
OPERATOR-THEORETIC FRAMEWORK FOR GRADIENT-FREE FEDERATED LEARNING

Mohit Kumar

University of Rostock, Germany

Software Competence Center Hagenberg GmbH, Austria

Hagenberg, Austria

mohit.kumar@uni-rostock.de

Mathias Brucker, Alexander Valentinitisch

Software Competence Center Hagenberg GmbH

Hagenberg, Austria

{mathias.brucker, alexander.valentinitisch}@scch.at

Adnan Husakovic, Ali Abbas

Primetals Technologies Austria GmbH

Linz, Austria

{adnan.husakovic, ali.abbas}@primetals.com

Manuela Geiß, Bernhard A. Moser

Software Competence Center Hagenberg GmbH

Hagenberg, Austria

{manuela.geiss, bernhard.moser}@scch.at

ABSTRACT

Background: Federated learning in practice must address client heterogeneity, strict communication and computation requirements, and data privacy, while optimizing performance. **Objectives:** Develop an operator-theoretic framework for federated learning that simultaneously addresses statistical heterogeneity, performance guarantees, and privacy under practical communication and computation constraints. **Methods:** We first map the L^2 -optimal solution into a reproducing kernel Hilbert space (RKHS) using a forward operator. Using the available data in that RKHS, we approximate the optimal solution. We then map this solution back to the original L^2 function space via the inverse operator. This construction yields a gradient-free learning scheme. We derive explicit finite-sample performance bounds for this scheme using concentration inequalities over operator norms. The framework analytically identifies a data-dependent hypothesis space and provides guarantees on risk, prediction error, robustness, and approximation error. Within this space, we design a communication- and computation-efficient model using kernel machines, leveraging the *space folding* property of Kernel Affine Hull Machines (KAHMs). Clients transfer knowledge to the server using a novel scalar metric, *space folding measure*, derived from KAHMs. Being a scalar, this measure greatly reduces communication overhead. It also supports a simple differentially private FL protocol in which scalar space folding summaries are computed from noise-perturbed data matrices obtained via a single application of a noise-adding mechanism, thereby avoiding per-round gradient clipping and privacy accounting. Finally, the induced global prediction rule can be implemented using a small number of integer minimum and equality-comparison operations per test point, making it structurally compatible with fully homomorphic encryption (FHE) during inference. **Results:** Across four benchmarks (20Newsgroup, XGLUE-NC, CIFAR-10-LT, CIFAR-100-LT), the resulting gradient-free FL method built on fixed encoder embeddings is competitive with, and in several cases outperforms, strong gradient-based federated fine-tuning, with gains of up to 23.7 percentage points on the considered benchmarks. In differentially private experiments, the proposed kernel-based smoothing mechanism partially offsets the accuracy loss caused by noise in high-privacy regimes. The induced global prediction rule admits an FHE realization based on $Q \times C$ encrypted minimum and C equality-comparison operations per test point (where $Q = \#\text{clients}$ and $C = \#\text{classes}$), and our operation-level benchmarks for these primitives indicate latencies

compatible with practical secure inference at the evaluated cryptographic parameter settings. **Conclusions:** The operator-theoretic, gradient-free federated learning framework provides provable performance guarantees with low communication overhead, supports differentially private knowledge transfer via scalar space folding summaries, and yields an FHE-compatible prediction rule for which we report operation-level runtimes, offering a mathematically grounded alternative to gradient-based federated learning under client heterogeneity.

1 Introduction

Developing a competitive machine learning model often necessitates a substantial amount of high-quality data for training. However, integrating different data sources to utilize all distributed data is challenging due to data privacy requirements and cumbersome exchange procedures. Data privacy is becoming increasingly important due to regulations such as the GDPR (General Data Protection Regulation) [58] and Artificial Intelligence Act (AI Act) [59]. Federated learning (FL) offers a popular solution for collaborative learning from distributed, privately owned data under the orchestration of a central server, without requiring participating clients to share raw training data. Despite its appeal, FL faces several practical challenges. First, data across clients are often statistically heterogeneous (non-IID), which can degrade the performance of a one-size-fits-all model. Second, devices may have limited computation and communication budgets, making frequent or large exchanges infeasible. Third, strict privacy and security guarantees (e.g., via differential privacy or encryption) must be enforced without significantly degrading the model’s learning performance and computation-efficiency.

1.1 Requirements

We identify the following requirements for the development of an effective FL algorithm. **R1 (Hypothesis Space for Learning from Heterogeneous Distributed Data):** Provide a mathematical framework that, without imposing parametric form or homogeneity assumptions on the client data distributions (beyond mild regularity conditions such as square-integrability), determines a suitable hypothesis space for task learning in a federated setting. **R2 (Theoretical Guarantees):** Calculate theoretically the error bounds and evaluate the task learning solution in terms of 1) robustness of the prediction error against the *disturbances* arising from uncertainties and data noise, and 2) accuracy of the solution and asymptotic upper bound on the approximation error. **R3 (Communication Efficiency):** Ensure that the analytically derived task learning solution can be implemented in a federated setting with communication and computational efficiency. Specifically, the optimization of the global model should not require multiple rounds of communication between the server and clients. **R4 (Efficient Differentially Private Knowledge Transfer Across Global and Local Models):** Define a novel differentially private metric, that allows for knowledge transfer from clients to server by solving the global model optimization problem without requiring exchange of gradients or model parameters (that may be computationally challenging and not optimal for privacy preservation), while optimizing the utility-privacy tradeoff. **R5 (Computationally Efficient FHE-Based Secure Federated Learning):** Rather than transmitting high-dimensional gradient or parameter update vectors, which entail substantial computational overhead when transmitted and operated on within encrypted domains, define novel low-dimensional attributes. These attributes must enable the inference of the global model with low communication overhead and latency, thereby supporting a computationally efficient realization of fully homomorphically encrypted inference of the global model. Researchers have approached these issues from multiple angles including communication-efficient protocols, robust aggregation methods, differential privacy techniques, and secure computation. We briefly review these developments next.

Problem Setting and Scope Throughout this work we focus on a practically motivated FL setting in which pre-trained encoders (e.g., deep neural networks trained on large public corpora) are available to the participating clients, but are not jointly updated during federated training. This reflects scenarios where

1. encoder updates may be constrained by regulatory or validation requirements,
2. communication and computation budgets preclude repeated updates of large models, or
3. the clients intend to reuse a common frozen representation while collaboratively learning a task-specific prediction head.

Within this setting, our goal is not to improve representation learning itself, but to design and analyse a gradient-free, communication-efficient, and privacy-/security-aware federated head on top of such fixed encoders.

1.2 State of the Art

To address the above challenges, a variety of federated learning strategies have been explored. In this section, we review key state-of-the-art approaches related to communication efficiency and handling heterogeneous data, differentially private learning, and secure aggregation with homomorphic encryption, among others.

1.2.1 Communication Efficiency and Addressing Statistical Heterogeneity

A cloud-edge architecture effectively mitigates communication and computation cost challenges by offloading computationally intensive learning tasks to the edge [76]. To accelerate the model convergence and thus to reduce the number of communication rounds during the learning process, device-to-device communication can be leveraged for mitigating the local over-fitting issue [24]. To reduce the communication cost, *low-rank Hadamard product parametrization* of the model parameters has been suggested [31]. Instead of training and transmitting full models, sparse models can be considered for computational and communication efficiency [5]. To enhance robustness against heterogeneity and improve communication efficiency, clients and the server exchange *abstract prototypes*, while local prototypes (rather than gradients) are aggregated [73]. To tailor the local model size and consequently the computation, memory, and data exchange requirements to the available client resources, an importance-based pruning mechanism has been proposed to extract lower-footprint nested submodels [29]. To facilitate FL on heterogeneous devices, a split-mix strategy [27] enables the learning of base sub-networks with varying sizes and robustness levels, which can be aggregated on-demand to meet specific inference requirements. The challenges of system heterogeneity and connection uncertainty in federated learning can be tackled by developing models that are readily prunable to arbitrary sizes and thus can be structurally decomposed for learning, inference, and transmission [85]. The training on the devices can be accelerated by introducing sparsity [62]. To achieve a faster convergence rate in theory and practice, adaptive gradient methods has been integrated into the FL [77]. The sparse and complementary subsets of the dense model are instead exchanged between server and clients to reduce communication and computational cost [33]. The similarities among clients can be assessed to enable personalized FL while reducing communication overhead, achieved through the optimization of aggregation weights [48].

A study [25] identifies *local learning bias* as the pitfall of FL with heterogeneous data, and introduces an algorithm that leverages label-distribution-agnostic pseudo-data to reduce the learning bias on local features and classifiers. An empirical study [81] reports that the large sparse convolution kernels can lead to enhanced robustness against distribution shifts in FL. A robust FL approach is to alleviate the *worst-case* effect of distribution shifts on the model performance. This approach has been followed [66] for the case of *affine distribution shifts* by minimizing the maximum possible loss induced by distribution shifts across clients. Adversarial learning approach [44] has been considered for addressing distribution shifts, where the server aims to train a discriminator to distinguish the representations of the clients while the clients aim to generate a common representation distribution. *Knowledge distillation* is another approach to address heterogeneity, where e.g. the knowledge about the global view of data distribution is extracted by the server and is distilled to guide local models' learning [86]. The clustered FL approach [69, 75] addresses data distribution heterogeneity by grouping clients with similar distributions. This enables clients within the same cluster to mutually benefit from federated learning while reducing harmful interference from clients with dissimilar distributions. A clustered FL algorithm [22] alternately estimates the cluster identities of the clients and optimizes model parameters for the client clusters. Assuming that each client's data follows a mixture of multiple distributions, a method [68] facilitates the simultaneous training of cluster models and personalized local models.

A personalized FL approach to tackle statistical heterogeneity involves clients and the server aiming at learning a global representation together, while each client learns its unique head locally [15]. Along this line, a study [2] suggests learning a common kernel function (parameterized by a neural network) across all clients, while each client employs a personalized Gaussian Process model. A simple personalization mechanism can be provided using local k-nearest neighbors model based on the shared representation provided by the global model [50]. The personalized FL problem can be studied under the *model agnostic meta-learning framework* [19], where the goal is to find an initial shared model that clients can easily adapt to their local datasets. Instead of using a *meta-model* as the initialization, both the personalized and global models can be pursued in parallel by formulating a bi-level optimization problem using the Moreau envelope as a regularized loss function [18]. To enable pairwise collaborations between clients with similar data, a mechanism has been proposed for personalized federated learning, which exchanges weighted model-aggregation messages between personalized models and personalized cloud models [30]. To address the computational limitations of heterogeneous devices in personalized FL, optimized masking vectors (derived by minimizing the bias term in the convergence bound) can be employed to train a sub-network of the learning model for each device, tailored to device's computational capacity [70]. A personalized FL method [84] adaptively aggregates the global model and the previous local model to initialize the local model.

1.2.2 Differentially Private Learning

Differential privacy is the gold-standard approach to provide anonymization guarantees for the data used in FL [6]. Differential privacy can be enforced within machine learning pipeline at any of three stages: on the input training data, during model training, or on model predictions [61]. Releasing a differentially private version of the training data would enable any training algorithm to be applied to it, thanks to differential privacy's post-processing property. However, a large amount of noise would typically be needed in this setting to be added into the data for achieving differential privacy guarantee, leading to the loss of model utility. To address the utility-loss issue, the notion of differential privacy can be relaxed to allow defining privacy in terms of the distinguishability level between inputs by means of a distance function [20, 63]. Generating differentially private synthetic data for the training of models is a promising approach to privacy-preserving machine learning. Differentially private synthetic data can be generated using random projections [23], Bayesian networks [82], Markov random fields [8], GANs [79], iterative methods [49], neural tangent kernels [78], and *Kernel Affine Hull Machines* [40]. Perturbing the objective function of the optimization problem with noise is another approach to enforce differential privacy in models with strong convexity [60, 32]. The study in [57] introduced Private Aggregation of Teacher Ensembles (PATE) method based on knowledge aggregation from “teacher” models and transfer to a “student” model in differentially private fashion. The authors suggested to train the student model on public unlabeled data using GAN-like approach for semi-supervised learning. This approach achieves differential privacy by injecting noise in the aggregation of teacher models (that have been trained on disjoint splits of a private dataset) without placing any restrictions on teacher models, thus allowing any models to be used in *model-agnostic* manner. Further, a rigorous theoretical analysis of the PATE approach is available [4, 47].

As it is typical that large-scale machine learning models are optimized using gradient-based algorithms, the gradient perturbation-based methods such as DP-SGD (differentially private stochastic gradient descent) are widely used for achieving rigorous differential privacy guarantees [1]. DP-SGD operates by running stochastic gradient descent on noisy mini-batch gradients. The DP-SGD method’s privacy analysis relies on the concept of *privacy amplification by sampling* requiring that each mini-batch is sampled with replacement on each iteration. This requirement may be infeasible to achieve in distributed settings like FL. Thus, the authors in [35] introduced a solution, referred to as DP-FTRL algorithm, that does not rely on random sampling for privacy amplification, instead leverages the differentially private streaming of the cumulative sum of gradients. The problem of continual-release of cumulative sums was connected to the matrix mechanism in [17], yielding improvements in federated learning with user-level differential privacy. The matrix mechanism was extended in [14] to the multi-epoch setting, allowing for differentially private gradient-based machine learning with multiple epochs over a dataset.

The integration of FL and differential privacy is potentially an effective approach to privacy-preserving collaborative learning from distributed datasets [55]. In practice, the edge device heterogeneity may cause straggler effect that can be mitigated by an *asynchronous* approach allowing clients to synchronize with the central server independently and at different times [45]. FL models have typically deep learning architecture, where estimating the sensitivity of gradients (which is required for differential privacy), is difficult, and thus gradients are clipped (whenever their norm exceed some threshold) to control the sensitivity [1, 53]. Gradient clipping has been theoretically proven and empirically observed to accelerate gradient descent optimization in training [83]. However, the clipping may lead to a bias on the convergence to a stationary point and the clipping bias can be quantified with a disparity measure between the gradient distribution and a geometrically symmetric distribution [10]. The choice of clipping threshold is crucial to the performance of differentially private models, necessitating the development of automatic [7] and adaptive [45] gradient clipping methods.

1.2.3 Secure Learning with Homomorphic Encryption

Fully Homomorphic Encryption (FHE) allows arbitrary computation in encrypted space, and thus FHE can be applied to FL for protecting the privacy of the data (shared by clients with the server during training) against a server eavesdropper. Traditional single-key FHE (where all clients share one key) poses a security risk in the event of a client colluding with the server, and thus multi-key FHE schemes [54] have been considered for FL [72], where the suggested scheme remains secured against collusion attacks involving up to all but one participant. Moreover, their scheme reduces the computational load by reducing the complexity of the NAND gate. In order to reduce computational and communication overhead during HE-based secure model aggregation in federated training, the authors in [34] suggest to selectively encrypt only the most privacy-sensitive parameters. The CKKS scheme [11] is a leveled homomorphic encryption scheme designed specifically for approximate arithmetic on real or complex numbers, and thus has been considered for privacy-preserving FL [56]. The SPDZ protocol [16] uses somewhat homomorphic encryption and provides low-latency secure multi party computation due to its fast online phase, and thus can be considered for FL [74]. Smart network interface cards can be leveraged as hardware accelerators to offload compute-intensive HE operations of FL [13]. To implement FHE-based secure FL with reduced communication overhead and latency, the authors

in [26] have experimented with different approaches in which data can be encrypted and transmitted. TFHE (Torus Fully Homomorphic Encryption) is a lattice-based FHE scheme [12] designed for fast gate-by-gate computation on encrypted data with very fast bootstrapping. TFHE leads to a computationally efficient FHE-based FL [43, 39] by reformulating the FL problem in such a way that 1) exchange of only low-dimensional attributes is required between clients and server, and 2) the inference of the global model in encrypted space is not computationally heavy.

1.3 Research Gap

In summary, most existing FL algorithms address only a subset of requirements R1–R5. For example, the classic FedAvg algorithm [52] (and similar gradient-based variants) requires many SGD epochs and communication rounds, yet still lacks formal guarantees for heterogeneous data. Approaches using differentially private SGD enhance privacy, but in federated settings they face other issues (e.g. performance drop and reliance on random sampling for privacy amplification). Methods using homomorphic encryption can secure model updates, however, they incur high computational cost and often still depend on exchanging gradients. Recently, kernel-based and prototype-based methods (e.g. distillation of local prototypes, or KAHM-based aggregation) have been explored to reduce communication and improve robustness, but a unified theoretical framework is lacking. In particular, the state of the art (as reviewed in Section 1.2.1, 1.2.2, 1.2.3) doesn’t jointly solve R1–R5. A recent kernel-based FL scheme [42] addressed aspects of R2–R3 and enabled privacy [40] and security [39] by exchanging compact task-sufficient information, but it did not address R1 by deriving the underlying hypothesis space for learning from heterogeneous distributed data. Existing FL algorithms either rely on iterative gradient descent, repeated communication, or do not fully protect privacy, and they do not come with end-to-end learning guarantees under heterogeneity. To the best of our knowledge, no existing framework concurrently addresses all requirements R1–R5 within a single, mathematically unified treatment.

1.4 Contributions

To address these limitations, we argue that a fundamentally different (gradient-free) approach is needed to satisfy requirements R1–R5. In this paper, we therefore propose a rigorous operator-theoretic framework that is designed to address R1–R5 within a single, unified treatment. Our framework generalizes the methodology of [42] and extends the methodology to address not only R1 but also simultaneously R2, R3, R4, and R5 in a unified manner. Our main contributions are:

Operator-Theoretic Formulation. We formulate the FL problem in the L^2 function space as solving for the minimizer of mean-squared error. We define an invertible forward operator that maps the L^2 -optimal solution into an RKHS. This RKHS is associated with a generalized kernel whose feature-map serves as an estimator of class posterior probability. We derive a sample-based estimator in the RKHS and prove non-asymptotic upper bound on its risk (Theorem 1) using kernel and operator theory and concentration inequalities. Notably, under mild regularity conditions on the data-generating distributions (i.e., square-integrability of the class-posterior function, as stated in Section 3.1), we achieve a risk bound of $\mathcal{O}(1/\sqrt{N})$, where N is the total number of training data samples distributed across clients. We map the RKHS learning solution back to L^2 function space using the inverse-operator to obtain a generalized learning solution.

Performance Guarantees. The generalized learning solution (provided in Section 3.5) is evaluated for its performance in terms of risk (in Theorem 2), prediction error (in Theorem 3), robustness (in Remark 3), and approximation error (in Theorem 4). Under the same regularity condition on the data-generating distributions as in Section 3.1, it is shown that risk, prediction error, and approximation error bounds are of $\mathcal{O}(1/\sqrt{N})$. *The robustness property of the generalized learning solution is established by showing that small disturbances cannot lead to large prediction error.*

Determination of Hypothesis Space and Theoretical Analysis. A data-dependent hypothesis space for task learning is determined (in Section 3.6) from the generalized learning solution by tuning the kernel to the scale of the data. We then analyze the Rademacher complexity of the hypothesis space (in Theorem 5) and derive upper bound on prediction error (in Theorem 6) and approximation error (in Theorem 7). It is shown (in Remark 6) that the achieved error bounds are tighter than the existing bounds [42].

KAHM-Based Space Folding Kernel and Classification: We introduce a novel *space folding measure* for KAHS (Definition 1). It quantifies how much a data point must be “folded” to fit it into the subspace spanned by the training data (i.e. the KAHS’s data subspace). This space folding measure induces a new kernel whose feature-map aligns with the class membership (Remark 7). In fact, the feature value can be interpreted as an estimate of class posterior probability (Proposition 2).

1.5 The Proposed Operator-Theoretic Framework

Below, we outline the key steps of our approach (illustrated in Fig. 1) before delving into the technical details.

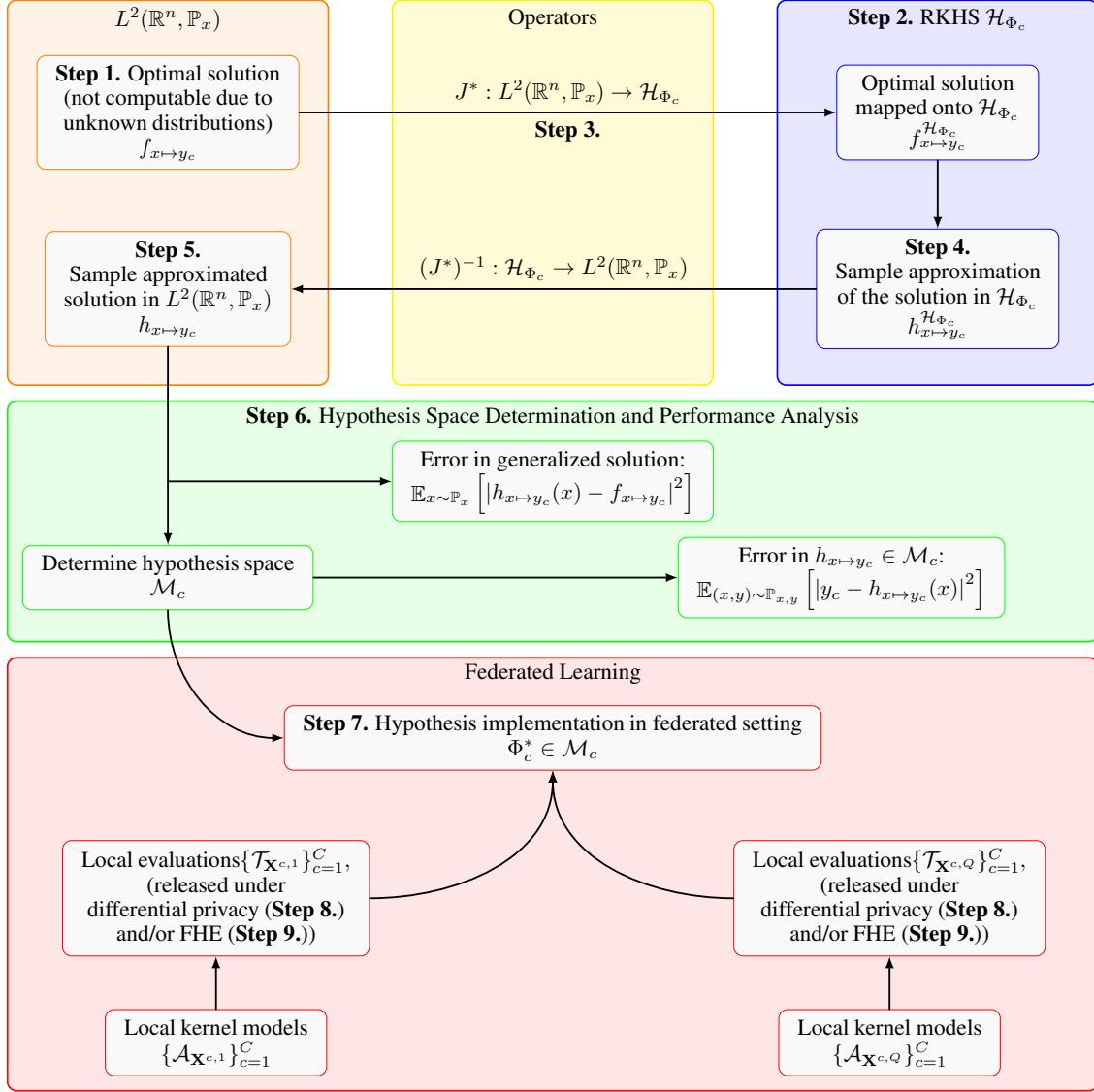


Figure 1: The operator-theoretic kernel FL framework is developed by 1) considering the optimal learning solution in $L^2(\mathbb{R}^n, \mathbb{P}_x)$, 2) mapping the optimal solution onto a RKHS (associated to a generalized kernel) using an operator, 3) approximating the optimal solution using available data samples in RKHS, 4) mapping the sample approximated solution onto $L^2(\mathbb{R}^n, \mathbb{P}_x)$ using the inverse-operator, 5) analyzing the sample approximated solution in $L^2(\mathbb{R}^n, \mathbb{P}_x)$ and identifying conditions on kernel choice to define hypothesis space, 6) implementing a suitable hypothesis with the minimum computational and communication cost in the federated setting using kernel models.

Step 1. Analytical Formulation of the Learning Problem in L^2 Function Space The task learning problem is mathematically formulated in the L^2 function space and the optimal solution is analytically derived. Since the probability distributions involved in the analytical solution are unknown, the analytically derived solution cannot be practically computed.

Step 2. A Reproducing Kernel Hilbert Space (RKHS) for the Learning Problem We consider the RKHS induced by a generalized kernel to approximate the optimal solution. Because the kernel choice is not obvious a priori, we adopt

a generalized kernel whose feature map is derived from theoretical analysis aimed at estimating the class-posterior probability.

Step 3. An Invertible Operator for Mapping Optimal Solution onto RKHS For an approximation and analysis of the optimal solution in RKHS using powerful kernel theory, an integral kernel operator with its inverse existing is defined from L^2 function space to RKHS.

Step 4. Sample Approximation and Analysis in RKHS: The mapped (onto RKHS) optimal solution is approximated by means of training data samples distributed across clients to obtain the *RKHS learning solution*. Further, kernel theory and concentration inequalities are applied to derive upper bound on the risk for the RKHS learning solution.

Step 5. Generalized Learning Solution and Analysis The learning solution, obtained in RKHS via sample approximation, is mapped onto the L^2 function space through the inverse-operator to obtain the *generalized learning solution*. The risk and error bounds for the generalized learning solution are easily obtained, thanks to the derived risk bound for the RKHS learning solution.

Step 6. Determination of the Hypothesis Space and Analysis The generalized learning solution is modulated by the kernel choice. We identify conditions on kernel under which the generalized learning solution captures the data’s scale. The set of learning solutions with the kernel satisfying the identified conditions defines the hypothesis space. Rademacher complexity of the hypothesis space is calculated to derive upper bounds on prediction error and approximation error for the hypothesis space.

Step 7. Choosing a Communication Efficient Hypothesis Having determined the hypothesis space and provided theoretical performance guarantees, the next step is to choose a suitable hypothesis that can be efficiently implemented in the federated setting. It is highlighted that it is possible to define such a hypothesis by means of Kernel Affine Hull Machines (KAHMs) [40, 42]. Specifically, the *space folding property* of the KAHMs (where the space folding property refers to the mapping of an arbitrary point by a KAHM onto the data subspace represented by the KAHM) is leveraged to implement a communication efficient hypothesis. This is done by specifying the feature-map of the kernel as an estimator of class posterior probability from the space folding measures, enabling gradient-free FL protocol where local KAHM-based models are aggregated by means of space folding measures without requiring rounds of communication between server and the clients.

Step 8. Differentially Private Release of Space Folding Measures Privacy-preserving knowledge transfer from clients to server is enabled by providing differentially private approximations to the space folding measures. Since estimating the sensitivity of space folding measure is challenging, we consider the differentially private release of data samples using an optimized noise adding mechanism [41]. The adverse effect of the added noise is mitigated by leveraging the post-processing property of differential privacy for a smoothing of noise added data samples. The study introduces a kernel-based smoothing function, with the degree of smoothing optimized to minimize the deviation of smoothed data points from original data points.

Step 9. Computationally Efficient Secure Inference of Global Model Using FHE Since the space folding measure (unlike high-dimensional gradient or parameter update vectors) is scalar-valued, inference for a C -class problem with Q participating clients reduces to evaluating $Q \times C$ space folding measures and performing $Q \times C$ minimum operations and C equality-comparison operations per test point. This fixed and low-dimensional operation pattern is well suited to secure implementation under fully homomorphic encryption. In our experiments, we instantiate this using the TFHE scheme and report runtimes for the encrypted minimum and equality-comparison primitives to characterize the computational profile of secure inference.

1.6 Novelty

Kernel methods, empowered by a strong mathematical theory on kernel machine learning, have been considered for FL [28, 21]. However, only a recent study [42] has introduced a kernel FL method that departs from gradient descent. That study proposed a KAHM-based federated scheme which considered a particular convex-hull hypothesis and derived Rademacher-complexity-based error bounds, and the study is complemented by separate works on KAHM-based differentially private [40] and FHE-secured [39] FL protocols. Taken together, these works partially address R2–R3 and demonstrate that KAHM-based scalar, task-sufficient summaries can support privacy and security. They do not, however, provide a unified operator-theoretic formulation that derives the hypothesis space from first principles, nor do they jointly address requirements R1–R5 within a single framework.

The present paper goes substantially beyond those earlier works in several ways:

Operator-Theoretic Formulation and Hypothesis-Space Derivation Instead of postulating a particular convex hypothesis set as in [42], we formulate the learning problem in the L^2 function space, derive the L^2 -optimal solution, map it into an RKHS via an invertible operator, and then map a sample-based RKHS approximation back to L^2 . This forward-inverse operator construction yields a generalized learning solution and, by identifying conditions on the generalized kernel, induces a data-dependent hypothesis space tailored to heterogeneous client distributions.

Integrated Operator- and Complexity-Based Analysis with Tighter Bounds We derive non-asymptotic risk, prediction-error, robustness, and approximation-error bounds for the generalized solution via operator-theoretic arguments, and then combine these with new Rademacher-complexity bounds. The resulting bounds are strictly tighter than those in [42].

Generalized space folding Kernel and Probabilistic Interpretation We extend the KAHM-induced distance used for aggregation in [42, 40, 39] to a new space folding measure and associated generalized kernel whose feature map admits an interpretation as an estimator of the class posterior probability. This construction is derived from the operator-theoretic framework rather than chosen ad hoc.

Unified Treatment of Communication Efficiency, DP, and FHE Within the same operator-theoretic framework, we show that scalar space folding summaries are sufficient statistics for the global prediction rule, that they admit an optimal univariate noise distribution for (ϵ, δ) -differential privacy together with a kernel-based post-processing smoother, and that the resulting decision rule has a fixed gate-level structure compatible with FHE-based secure inference. Earlier works [42, 40, 39] studied these aspects in isolation, without connecting them to a formally derived hypothesis space.

Expanded Empirical Study We provide a new empirical evaluation on four benchmarks (20Newsgroup, XGLUE-NC, CIFAR-10-LT, CIFAR-100-LT) under heterogeneous and long-tailed data partitions, including ablations on space folding, batch size, and embedding combinations. These experiments go beyond those reported in [42, 40, 39].

1.7 Organization

The remainder of the paper is organized as follows: Section 2 introduces the necessary notations and formal problem setup. Section 3 presents the development of our operator-theoretic framework, including theoretical analysis of its performance. Section 4 reports experimental results on benchmarks, and 5 concludes with a discussion of the findings and future work.

2 Mathematical Prerequisites

This section introduces the used notations, presents the considered distributed data setting, and reviews the necessary definitions.

2.1 Notations

We use the boldface font to denote the matrices. The following notations are introduced:

- Let $n, N, c, C, q, Q \in \mathbb{Z}_+$ be the positive integers.
- For a scalar $a \in \mathbb{R}$, $|a|$ denotes its absolute value. For a set A , $|A|$ denotes its cardinality. For a real matrix \mathbf{X} , \mathbf{X}^T is the transpose of \mathbf{X} .
- For a vector $y \in \mathbb{R}^C$, $\|y\|$ denotes the Euclidean norm and y_j (and also $(y)_j$) denotes the j^{th} element. For a matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$, $\|\mathbf{X}\|_2$ denotes the spectral norm, $\|\mathbf{X}\|_F$ denotes the Frobenius norm, $(\mathbf{X})_{i,:}$ denotes the i^{th} row, $(\mathbf{X})_{:,j}$ denotes the j^{th} column, and $(\mathbf{X})_{i,j}$ denotes the $(i, j)^{th}$ element.
- For a set $\{x^1, \dots, x^N\} \subset \mathbb{R}^n$, its affine hull is denoted as $\text{aff}(\{x^1, \dots, x^N\})$.
- The square brackets are used to represent the construction of a matrix from columns e.g. $[x^1 \dots x^N]$ is a matrix with vectors x^1, \dots, x^N as the columns.

2.2 Definitions

- Let $(\Omega_x, \mathcal{F}_x, \mu_x)$ be a probability space and $x : \Omega_x \rightarrow \mathbb{R}^n$ be a random vector on Ω_x . Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ -algebra on \mathbb{R}^n . Let $\mathbb{P}_x : \mathcal{B}(\mathbb{R}^n) \rightarrow [0, 1]$ be the distribution of x given as

$$\mathbb{P}_x := \mu_x \circ x^{-1}. \quad (1)$$

- Let $(\Omega_{x,y}, \mathcal{F}_{x,y}, \mu_{x,y})$ be a probability space and $(x, y) : \Omega_{x,y} \rightarrow \mathbb{R}^n \times \{0, 1\}^C$ be a random vector on $\Omega_{x,y}$. Let $\mathcal{B}(\mathbb{R}^n \times \{0, 1\}^C)$ denote the Borel σ -algebra on $\mathbb{R}^n \times \{0, 1\}^C$. Let $\mathbb{P}_{x,y} : \mathcal{B}(\mathbb{R}^n \times \{0, 1\}^C) \rightarrow [0, 1]$ be the distribution of (x, y) given as

$$\mathbb{P}_{x,y} := \mu_{x,y} \circ (x, y)^{-1}. \quad (2)$$

- Let $(\Omega_{x,y,q}, \mathcal{F}_{x,y,q}, \mu_{x,y,q})$ be a probability space and $(x, y, q) : \Omega_{x,y,q} \rightarrow \mathbb{R}^n \times \{0, 1\}^C \times \{1, 2, \dots, Q\}$ be a random vector on $\Omega_{x,y,q}$. Let $\mathcal{B}(\mathbb{R}^n \times \{0, 1\}^C \times \{1, 2, \dots, Q\})$ denote the Borel σ -algebra on $\mathbb{R}^n \times \{0, 1\}^C \times \{1, 2, \dots, Q\}$. Let $\mathbb{P}_{x,y,q} : \mathcal{B}(\mathbb{R}^n \times \{0, 1\}^C \times \{1, 2, \dots, Q\}) \rightarrow \mathbb{R}$ be the distribution of (x, y, q) given as

$$\mathbb{P}_{x,y,q} := \mu_{x,y,q} \circ (x, y, q)^{-1}. \quad (3)$$

- Let $L^2(\mathbb{R}^n, \mathbb{P}_x)$ be the space of all complex-valued measurable functions on \mathbb{R}^n that satisfy

$$\int_{\mathbb{R}^n} |f(x)|^2 d\mathbb{P}_x(x) < \infty. \quad (4)$$

The norm of a $f \in L^2(\mathbb{R}^n, \mathbb{P}_x)$ is given as

$$\|f\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)} := \left(\int_{\mathbb{R}^n} |f(x)|^2 d\mathbb{P}_x(x) \right)^{1/2}. \quad (5)$$

2.3 Statistically Heterogeneously Distributed Data Setting

Let \mathcal{D} be a set consisting of N number of samples drawn IID according to the distribution $\mathbb{P}_{x,y}$:

$$\mathcal{D} := \{(x^i, y^i) \in \mathbb{R}^n \times \{0, 1\}^C \mid i \in \{1, 2, \dots, N\}\} \sim (\mathbb{P}_{x,y})^N. \quad (6)$$

Let \mathcal{I}^c be the set of indices of those samples in the sequence $((x^i, y^i) \in \mathcal{D})_{i=1}^N$ which are c^{th} class labelled, i.e.,

$$\mathcal{I}^c := \{i \in \{1, 2, \dots, N\} \mid y_c^i = 1\}. \quad (7)$$

Let N_c be the number of c^{th} class labelled samples, i.e.,

$$N_c = |\mathcal{I}^c|. \quad (8)$$

Let $\mathbf{I}^c = (I_1^c, \dots, I_{N_c}^c)$ be the sequence of elements of \mathcal{I}^c in ascending order, i.e.,

$$I_1^c = \min(\mathcal{I}^c) \quad (9)$$

$$I_i^c = \min(\mathcal{I}^c \setminus \{I_1^c, \dots, I_{i-1}^c\}), \forall i \in \{2, \dots, N_c\}. \quad (10)$$

Let $\mathbf{X}^c \in \mathbb{R}^{N_c \times n}$ be the matrix storing c^{th} class labelled samples as its rows, i.e.,

$$\mathbf{X}^c = [x^{I_1^c} \dots x^{I_{N_c}^c}]^T. \quad (11)$$

We consider the distributed data setting where total data samples are distributed among Q ($Q > 1$) different clients. Let $q^i \in \{1, 2, \dots, Q\}$ be the client characterizing variable associated to the i^{th} sample pair (x^i, y^i) indicating which of the Q clients owns the i^{th} sample pair. Let $\mathcal{I}^{c,q}$ be the set of indices of those samples in the sequence $((x^i, y^i) \in \mathcal{D})_{i=1}^N$ which are c^{th} class labelled and owned by client q , i.e.,

$$\mathcal{I}^{c,q} := \{i \in \{1, 2, \dots, N\} \mid (y^i)_c = 1, q^i = q\}. \quad (12)$$

Let $(I_1^{c,q}, \dots, I_{|\mathcal{I}^{c,q}|}^{c,q})$ be the sequence of elements of $\mathcal{I}^{c,q}$ in ascending order, i.e.,

$$I_1^{c,q} = \min(\mathcal{I}^{c,q}), \quad (13)$$

$$I_i^{c,q} = \min(\mathcal{I}^{c,q} \setminus \{I_1^{c,q}, \dots, I_{i-1}^{c,q}\}), \quad (14)$$

for $i \in \{2, \dots, |\mathcal{I}^{c,q}|\}$. Let $\mathbf{X}^{c,q} \in \mathbb{R}^{|\mathcal{I}^{c,q}| \times n}$ be the matrix storing the c^{th} class labelled and q^{th} client owned samples, i.e.,

$$\mathbf{X}^{c,q} = [x^{I_1^{c,q}} \dots x^{I_{|\mathcal{I}^{c,q}|}^{c,q}}]^T. \quad (15)$$

Since the c^{th} class labelled samples are distributed among Q clients, we have

$$N_c = |\mathcal{I}^{c,1}| + |\mathcal{I}^{c,2}| + \dots + |\mathcal{I}^{c,Q}|. \quad (16)$$

Remark 1 (Data Heterogeneity across Clients). *We assume that data samples are statistically heterogeneously distributed, i.e., for arbitrary clients q^i and q^j with $i \neq j$, we assume that*

$$\mathbb{P}_{x,y|q}(\cdot, \cdot | q = q^i) \neq \mathbb{P}_{x,y|q}(\cdot, \cdot | q = q^j), \quad (17)$$

$$\mathbb{P}_{y|x,q}(\cdot | x, q = q^i) \neq \mathbb{P}_{y|x,q}(\cdot | x, q = q^j). \quad (18)$$

2.4 Kernel Affine Hull Machine (KAHM)

The KAHS, originally defined in [40], have been considered for automated machine learning in [42]. Given a finite number of samples: $\mathbf{X} = [x^1 \dots x^N]^T$ with $x^1, \dots, x^N \in \mathbb{R}^n$, a KAHM $\mathcal{A}_\mathbf{X} : \mathbb{R}^n \rightarrow \text{aff}(\{x^1, \dots, x^N\})$ is defined as

$$\mathcal{A}_\mathbf{X}(x) := \frac{h_\mathbf{X}^1(\mathbf{P}_\mathbf{X}x)}{\sum_{i=1}^N h_\mathbf{X}^i(\mathbf{P}_\mathbf{X}x)} x^1 + \dots + \frac{h_\mathbf{X}^N(\mathbf{P}_\mathbf{X}x)}{\sum_{i=1}^N h_\mathbf{X}^i(\mathbf{P}_\mathbf{X}x)} x^N. \quad (19)$$

Appendix A presents a comprehensive description of the variables and functions associated with (19).

3 Operator-Theoretic Kernel Federated Learning Framework

This section provides an operator-theoretic framework for kernel FL. As stated previously in Section 1.5, the framework development approach consists of 7 steps. Each step is described separately in a subsection.

3.1 Step 1: Learning Problem in $L^2(\mathbb{R}^n, \mathbb{P}_x)$

We consider the learning problem in $L^2(\mathbb{R}^n, \mathbb{P}_x)$. Our goal is to learn a function $f_{x \mapsto y_c} : \mathbb{R}^n \rightarrow \mathbb{R}$ that minimizes the mean squared error:

$$f_{x \mapsto y_c} := \underset{g \in L^2(\mathbb{R}^n, \mathbb{P}_x)}{\operatorname{argmin}} \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - g(x)|^2] \quad (20)$$

$$= \underset{g \in L^2(\mathbb{R}^n, \mathbb{P}_x)}{\operatorname{argmin}} \left(\int_{\mathbb{R}^n \times \{0,1\}^C} |y_c - g(x)|^2 d\mathbb{P}_{x,y}(x,y) \right). \quad (21)$$

It is well-known and also shown in Appendix B that the conditional expectation, also known as regression function, minimizes the mean squared error. That is,

$$f_{x \mapsto y_c}(x) = \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c | x], \quad (22)$$

where we have made the following regularity assumption:

Assumption 1 (Square-Integrability of the Regression Function).

$$\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c | x] \in L^2(\mathbb{R}^n, \mathbb{P}_x). \quad (23)$$

Due to $y_c \in \{0, 1\}$, we have

$$f_{x \mapsto y_c}(x) \in [0, 1]. \quad (24)$$

For an analysis, the *disturbance function*, $\xi_c : \mathbb{R}^n \times \{0, 1\}^C \rightarrow \mathbb{R}$, is defined as

$$\xi_c(x, y) := y_c - f_{x \mapsto y_c}(x) \quad (25)$$

$$= y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c | x]. \quad (26)$$

It is obvious that

$$\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [\xi_c(x, y)] = 0, \quad (27)$$

and

$$\xi_c(x, y) \in [-1, 1]. \quad (28)$$

3.2 Step 2: A RKHS Associated to a Generalized Kernel Function

We consider a generalized kernel function such that for each class $c \in \{1, 2, \dots, C\}$,

$$\mathcal{K}_{\Phi_c}(x, x') := \Phi_c(x)\Phi_c(x'), \quad (29)$$

where $\Phi_c : \mathbb{R}^n \rightarrow [0, 1]$ is the feature-map (which will be determined based on theoretical analysis to estimate class posterior probability).

Remark 2 (Rational for the Restrictive Feature-Map). *The rational for the restrictive nature of the feature-map, $\Phi_c : \mathbb{R}^n \rightarrow [0, 1]$, is the intent of setting it as an estimator of the class posterior probability, i.e., $\Phi_c(x) \approx \mathbb{P}_{y|x}(y_c = 1|x)$. The bound on the error in approximating class posterior probability through the feature-map will be derived (in Proposition 2).*

Since for any $x \in \mathbb{R}^n$, $\Phi_c(x) \in [0, 1]$, thus we have

$$\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \geq 0 \quad (30)$$

$$\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \leq 1. \quad (31)$$

It is shown in Appendix C that \mathcal{K}_{Φ_c} is a positive semi-definite kernel. Now the RKHS associated to \mathcal{K}_{Φ_c} is given as

$$\mathcal{H}_{\Phi_c} := \left\{ f = \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_{\Phi_c}(\cdot, x^i) \mid \alpha_i \in \mathbb{R}, x^i \in \mathbb{R}^n, \|f\|_{\mathcal{H}_{\Phi_c}}^2 := \sum_{i,j=1}^{\infty} \alpha_i \alpha_j \mathcal{K}_{\Phi_c}(x^i, x^j) < \infty \right\} \quad (32)$$

with inner product for any $f = \sum_{i=1}^N a_i \mathcal{K}_{\Phi_c}(\cdot, s^i) \in \mathcal{H}_{\Phi_c}$ and $g = \sum_{j=1}^M b_j \mathcal{K}_{\Phi_c}(\cdot, t^j) \in \mathcal{H}_{\Phi_c}$ defined as

$$\langle f, g \rangle_{\mathcal{H}_{\Phi_c}} := \sum_{i=1}^N \sum_{j=1}^M a_i b_j \mathcal{K}_{\Phi_c}(s^i, t^j). \quad (33)$$

3.3 Step 3: Operators between $L^2(\mathbb{R}^n, \mathbb{P}_x)$ and RKHS

To enable kernel-based approximation, we map the L^2 -optimal function into an RKHS using a bounded linear operator. We introduce an operator from $L^2(\mathbb{R}^n, \mathbb{P}_x)$ to \mathcal{H}_{Φ_c} such that it is invertible. For defining such an operator, we first consider the inclusion operator $J : \mathcal{H}_{\Phi_c} \hookrightarrow L^2(\mathbb{R}^n, \mathbb{P}_x)$, its adjoint operator $J^* : L^2(\mathbb{R}^n, \mathbb{P}_x) \rightarrow \mathcal{H}_{\Phi_c}$, and the inverse of the adjoint operator $(J^*)^{-1} : \mathcal{H}_{\Phi_c} \rightarrow L^2(\mathbb{R}^n, \mathbb{P}_x)$. It is shown in Appendix D that J is well defined. Consider for any $f \in \mathcal{H}_{\Phi_c}$,

$$\langle Jf, g \rangle_{L^2(\mathbb{R}^n, \mathbb{P}_x)} = \mathbb{E}_{x \sim \mathbb{P}_x} [f(x)g(x)] \quad (34)$$

$$= \mathbb{E}_{x \sim \mathbb{P}_x} [\langle f, \mathcal{K}_{\Phi_c}(x, \cdot) \rangle_{\mathcal{H}_{\Phi_c}} g(x)] \quad (35)$$

$$= \left\langle f, \mathbb{E}_{x \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x, \cdot)g(x)] \right\rangle_{\mathcal{H}_{\Phi_c}}. \quad (36)$$

It follows that the adjoint of J , $J^* : L^2(\mathbb{R}^n, \mathbb{P}_x) \rightarrow \mathcal{H}_{\Phi_c}$, is given as

$$(J^*g)(x) := \mathbb{E}_{x' \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x', x)g(x')] \quad (37)$$

$$= \int_{\mathbb{R}^n} \mathcal{K}_{\Phi_c}(x', x)g(x') d\mathbb{P}_x(x'). \quad (38)$$

Consider for any $f = \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_{\Phi_c}(\cdot, x^i) \in \mathcal{H}_{\Phi_c}$,

$$\left(J^* \frac{f}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \right) (x) = \frac{1}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \mathbb{E}_{x' \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x', x) f(x')] \quad (39)$$

$$= \frac{\Phi_c(x)}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \mathbb{E}_{x' \sim \mathbb{P}_x} [\Phi_c(x') f(x')] \quad (40)$$

$$= \frac{\Phi_c(x)}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \mathbb{E}_{x' \sim \mathbb{P}_x} \left[\Phi_c(x') \sum_{i=1}^{\infty} \alpha_i \mathcal{K}_{\Phi_c}(x', x^i) \right] \quad (41)$$

$$= \frac{\Phi_c(x)}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \mathbb{E}_{x' \sim \mathbb{P}_x} [|\Phi_c(x')|^2] \sum_{i=1}^{\infty} \alpha_i \Phi_c(x^i) \quad (42)$$

$$= \Phi_c(x) \sum_{i=1}^{\infty} \alpha_i \Phi_c(x^i) \quad (43)$$

$$= f(x). \quad (44)$$

It follows that the inverse of J^* , $(J^*)^{-1} : \mathcal{H}_{\Phi_c} \rightarrow L^2(\mathbb{R}^n, \mathbb{P}_x)$, is given as

$$((J^*)^{-1} f)(x) := \frac{f(x)}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2}, \quad (45)$$

where $\Phi_c : \mathbb{R}^n \rightarrow [0, 1]$ characterizes the kernel \mathcal{K}_{Φ_c} , as stated in (29). It is shown in Appendix E that $(J^*)^{-1}$ is well defined on the range of J^* .

3.3.1 A Few Propositions

It is shown in Appendix F that $J^* J$ is given as

$$J^* J = \mathbb{E}_{x' \sim \mathbb{P}_x} [(\mathcal{K}_{\Phi_c}(x', \cdot) \otimes \mathcal{K}_{\Phi_c}(x', \cdot))], \quad (46)$$

where for any $f, f' \in \mathcal{H}_{\Phi_c}$, $f \otimes f' : \mathcal{H}_{\Phi_c} \rightarrow \mathcal{H}_{\Phi_c}$ is defined as

$$(f \otimes f')(g) := \langle g, f' \rangle_{\mathcal{H}_{\Phi_c}} f. \quad (47)$$

It is shown in Appendix G that the norm of the operator $J^* J$ is upper bounded as

$$\|J^* J\|_{\text{op}} \leq 1. \quad (48)$$

It is clear that $J^* J$ is self-adjoint. $J^* J$ is a positive operator, since for all $f \in \mathcal{H}_{\Phi_c}$,

$$\langle f, (J^* J)f \rangle_{\mathcal{H}_{\Phi_c}} = \langle Jf, Jf \rangle_{L^2(\mathbb{R}^n, \mathbb{P}_x)} \quad (49)$$

$$= \|Jf\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \quad (50)$$

$$\geq 0. \quad (51)$$

Since $J^* J$ is a positive self-adjoint operator, there is a unique positive self-adjoint square root of $J^* J$ and is denoted by $(J^* J)^{1/2}$. It follows from (48) that

$$\|(J^* J)^{1/2}\|_{\text{op}} \leq 1. \quad (52)$$

3.3.2 Sample Approximation of Operators

For a given sample sequence $(x^i \in \mathbb{R}^n)_{i=1}^N$, let $\widehat{S}_{(x^i)_{i=1}^N} : \mathcal{H}_{\Phi_c} \rightarrow (\mathbb{R}^N, \langle \cdot, \cdot \rangle_{\mathbb{R}^N})$ be the *sample evaluation* operator defined as

$$\widehat{S}_{(x^i)_{i=1}^N} f := (f(x^1), \dots, f(x^N)), \quad (53)$$

with inner product for any $u, v \in \mathbb{R}^N$ defined as

$$\langle u, v \rangle_{\mathbb{R}^N} := \frac{1}{N} \sum_{i=1}^N u_i v_i. \quad (54)$$

$\widehat{S}_{(x^i)_{i=1}^N}$ is viewed as the sample approximation of J . For any $u \in \mathbb{R}^N$, we have

$$\langle \widehat{S}_{(x^i)_{i=1}^N} f, u \rangle_{\mathbb{R}^N} = \frac{1}{N} \sum_{i=1}^N f(x^i) \cdot u_i \quad (55)$$

$$= \frac{1}{N} \sum_{i=1}^N \langle f, \mathcal{K}_{\Phi_c}(x^i, \cdot) \rangle_{\mathcal{H}_{\Phi_c}} u_i \quad (56)$$

$$= \left\langle f, \frac{1}{N} \sum_{i=1}^N u_i \mathcal{K}_{\Phi_c}(x^i, \cdot) \right\rangle_{\mathcal{H}_{\Phi_c}}. \quad (57)$$

It follows that adjoint of $\widehat{S}_{(x^i)_{i=1}^N}$, $\widehat{S}_{(x^i)_{i=1}^N}^* : (\mathbb{R}^N, \langle \cdot, \cdot \rangle_{\mathbb{R}^N}) \rightarrow \mathcal{H}_{\Phi_c}$, is given as

$$\widehat{S}_{(x^i)_{i=1}^N}^*(u_1, \dots, u_N) := \frac{1}{N} \sum_{i=1}^N u_i \mathcal{K}_{\Phi_c}(x^i, \cdot). \quad (58)$$

$\widehat{S}_{(x^i)_{i=1}^N}^*$ is viewed as the sample approximation of J^* .

3.4 Step 4: RKHS Learning Solution

The optimal solution (22), $f_{x \mapsto y_c} \in L^2(\mathbb{R}^n, \mathbb{P}_x)$, is mapped onto \mathcal{H}_{Φ_c} through the operator J^* , i.e.,

$$f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} := J^* f_{x \mapsto y_c}. \quad (59)$$

To approximate $f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}$ in RKHS, a natural approach is of approximating J^* in (59) using available data samples \mathcal{D} (6). This leads to

$$h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} := \widehat{S}_{(x^i)_{i=1}^N}^*(y_c^1, \dots, y_c^N), \quad (60)$$

where $h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} \in \mathcal{H}_{\Phi_c}$ is viewed as the sample approximation of $f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}$. For a given sequence $(x^i \in \mathbb{R}^n)_{i=1}^N$, let $\text{Ev}_{(x^i)_{i=1}^N} : L^2(\mathbb{R}^n, \mathbb{P}_x) \rightarrow (\mathbb{R}^N, \langle \cdot, \cdot \rangle_{\mathbb{R}^N})$ be the function evaluation operator defined as

$$\text{Ev}_{(x^i)_{i=1}^N} g := (g(x^1), \dots, g(x^N)). \quad (61)$$

The given data samples can be represented using evaluation operator as

$$(y_c^1, \dots, y_c^N) = \text{Ev}_{(x^i)_{i=1}^N} f_{x \mapsto y_c} + (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)). \quad (62)$$

Combining (60) and (62), we get

$$h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} = (\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N}) f_{x \mapsto y_c} + \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)). \quad (63)$$

Theorem 1 (Risk for Sample Approximation of the Optimal Solution in RKHS). *The following holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$:*

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[\left| h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) \right|^2 \right] \leq \frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}}. \quad (64)$$

Proof. The proof is provided in five parts:

Part 1: It is shown in Appendix H that

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[\left| h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) \right|^2 \right] \leq \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}}. \quad (65)$$

Part 2: It is shown in Appendix I that

$$\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} \right] \leq \frac{2}{\sqrt{N}}. \quad (66)$$

Part 3: It is shown in Appendix J that

$$\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right] \leq \frac{1}{\sqrt{N}}. \quad (67)$$

Part 4: It is shown in Appendix K that we have with probability at least $1 - \delta$,

$$\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \geq \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \quad (68)$$

Part 5: Finally, we get (64) by using (68) in (65). \square

3.5 Step 5: Generalized Learning Solution in $L^2(\mathbb{R}^n, \mathbb{P}_x)$

The learning solution $h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} \in \mathcal{H}_{\Phi_c}$ is mapped onto $L^2(\mathbb{R}^n, \mathbb{P}_x)$ using inverse-operator $(J^*)^{-1}$, i.e.,

$$h_{x \mapsto y_c} := (J^*)^{-1} h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}. \quad (69)$$

The obtained solution $h_{x \mapsto y_c}$ is referred to as *generalized learning solution* reflecting upon the considered generalized kernel function. The learning solution is evaluated in Theorem 2 for its risk with respect to the optimal solution $f_{x \mapsto y_c}$.

Theorem 2 (Risk for Generalized Learning Solution). *The following holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$:*

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - f_{x \mapsto y_c}(x)|^2 \right] \leq \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \right)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right). \quad (70)$$

Proof. Consider

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - f_{x \mapsto y_c}(x)|^2 \right] = \mathbb{E}_{x \sim \mathbb{P}_x} \left[\left| \left((J^*)^{-1} h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} \right)(x) - \left((J^*)^{-1} f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} \right)(x) \right|^2 \right] \quad (71)$$

$$= \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \right)^2} \mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x)|^2 \right]. \quad (72)$$

Using Theorem (1) in (72), we get the result (70). \square

Theorem 2 allows to bound the error associated to $h_{x \mapsto y_c}$ in predicting the output y_c , as stated in Theorem 3.

Theorem 3 (Prediction Error Bound for Generalized Learning Solution). *The following holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$:*

$$\underbrace{\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right]}_{\text{mean-squared prediction error}} \leq \underbrace{\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[\left| y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right|^2 \right]}_{\text{mean-squared disturbance magnitude}} + \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \right)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right). \quad (73)$$

Proof. Consider the prediction error

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right] &= \mathbb{E}_{x \sim \mathbb{P}_x} \left[|f_{x \mapsto y_c}(x) - h_{x \mapsto y_c}(x)|^2 \right] + \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|\xi_c(x, y)|^2 \right] \\ &\quad + 2 \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [(f_{x \mapsto y_c}(x) - h_{x \mapsto y_c}(x)) \xi_c(x, y)]. \end{aligned} \quad (74)$$

Using (27), we get

$$\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right] = \mathbb{E}_{x \sim \mathbb{P}_x} \left[|f_{x \mapsto y_c}(x) - h_{x \mapsto y_c}(x)|^2 \right] + \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|\xi_c(x, y)|^2 \right]. \quad (75)$$

Using (26) and Theorem 2, the result is obtained. \square

Remark 3 (Robustness of Generalized Learning Solution). *Inequality (73) establishes the robustness property in the sense that if the disturbances are small (i.e. $\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] |^2]$ is small), then the prediction error cannot be large (i.e. $\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x) |^2]$ cannot be large). As the number of training samples N increases, the upper bound on prediction error decreases. In the limiting case, we have*

$$\lim_{N \rightarrow \infty} \frac{\underbrace{\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x) |^2]}_{\text{mean-squared prediction error}}}{\underbrace{\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[\left| y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right|^2 \right]}_{\text{mean-squared disturbance magnitude}}} \leq 1. \quad (76)$$

Theorem 4 provides an upper bound on the error in approximating the target function $\mathbb{P}_{y|x}(y_c = 1|x)$ through $h_{x \mapsto y_c}$.

Theorem 4 (Approximation Error Bound for Generalized Learning Solution). *The following holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$:*

$$\underbrace{\mathbb{E}_{x \sim \mathbb{P}_x} [|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x) |^2]}_{\text{mean-squared approximation error}} \leq \frac{1}{(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right). \quad (77)$$

Proof. Since $y_c \in \{0, 1\}$, we have

$$\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] = \mathbb{P}_{y|x}(y_c = 1|x). \quad (78)$$

Using (22), we get

$$\mathbb{P}_{y|x}(y_c = 1|x) = f_{x \mapsto y_c}(x). \quad (79)$$

Using (79) in Theorem 2 leads to the result. \square

Remark 4 (Asymptotic Convergence of Generalized Learning Solution). *Theorem 4 indicates that the approximation error bound decays with an increasing number of training samples. In the limiting case, the approximation error reduces to zero i.e.*

$$\lim_{N \rightarrow \infty} \mathbb{E}_{x \sim \mathbb{P}_x} [|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x) |^2] = 0. \quad (80)$$

3.6 Step 6: Identification of Conditions on Kernel and Determination of Hypothesis Space

It can be seen from (69) and (60) that the obtained solution $h_{x \mapsto y_c}$ is given as

$$h_{x \mapsto y_c}(x) = \frac{1}{\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \frac{1}{N} \sum_{i=1}^N y_c^i \mathcal{K}_{\Phi_c}(x^i, x) \quad (81)$$

$$= \frac{1}{N \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \sum_{i=1}^{N_c} \mathcal{K}_{\Phi_c}(x^{I_i^c}, x) \quad (82)$$

$$= \frac{\Phi_c(x)}{N \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \sum_{i=1}^{N_c} \Phi_c(x^{I_i^c}), \quad (83)$$

where $(x^{I_1^c}, \dots, x^{I_{N_c}^c})$ is the sequence of c^{th} class labelled samples, i.e., $y_c^{I_1^c} = \dots = y_c^{I_{N_c}^c} = 1$. Since $\Phi_c : \mathbb{R}^n \rightarrow [0, 1]$, we have for any $x \in \mathbb{R}^n$,

$$h_{x \mapsto y_c}(x) \in \left[0, \frac{N_c}{N \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \right]. \quad (84)$$

Since $h_{x \mapsto y_c}$ is viewed as the predictor of c^{th} class label $y_c \in \{0, 1\}$, we ensure that $h_{x \mapsto y_c}(x) \in [0, 1]$ by constraining the feature-map Φ_c as

$$\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{N_c}{N}, \quad (85)$$

resulting in

$$h_{x \mapsto y_c}(x) \in [0, 1]. \quad (86)$$

Remark 5 (Justification of the Normalization (85)). *Condition (85) is not an ad hoc constraint but a normalization that aligns the feature-map Φ_c with the class prior. The KAHM-based construction in Section 3.7.2 is designed such that $\Phi_c(x) \in [0, 1]$ behaves as an (almost) binary membership score for class c : in the idealized case encoded by (106) we have $\Phi_c(x) \in \{0, 1\}$, and Proposition 2 shows that $\Phi_c(x)$ provides a calibrated approximation of the posterior probability $\mathbb{P}_{y|x}(y_c = 1|x)$. In the binary idealization $\Phi_c(x) \in \{0, 1\}$ we obtain $|\Phi_c(x)|^2 = \Phi_c(x)$ and hence $\mathbb{E}_{x \sim \mathbb{P}_x} [|\Phi_c(x)|^2] = \mathbb{E}_{x \sim \mathbb{P}_x} [\Phi_c(x)]$. Under the posterior interpretation $\Phi_c(x) \approx \mathbb{P}_{y|x}(y_c = 1|x)$ we further have, by the law of total expectation, $\mathbb{E}_{x \sim \mathbb{P}_x} [\Phi_c(x)] \approx \mathbb{E}_{x \sim \mathbb{P}_x} [\mathbb{P}_{y|x}(y_c = 1|x)] = \mathbb{P}_y(y_c = 1)$. Combining these relations yields $\mathbb{E}_{x \sim \mathbb{P}_x} [|\Phi_c(x)|^2] \approx \mathbb{P}_y(y_c = 1)$. Imposing (85) for our design, i.e. $\mathbb{E}_{x \sim \mathbb{P}_x} [|\Phi_c(x)|^2] = N_c/N$, is therefore equivalent to requiring $\mathbb{P}_y(y_c = 1) \approx N_c/N$, which is the standard assumption that the empirical class frequency N_c/N approximates the true class prior. Thus, in our design of Φ_c , condition (85) encodes a natural and statistically consistent alignment with the empirical class prior, rather than an artificial restriction on the hypothesis space.*

With the kernel feature-map normalization condition (85), we define our hypothesis space as

$$\mathcal{M}_c := \left\{ h_{x \mapsto y_c} = \frac{\Phi_c(\cdot)}{N \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2} \sum_{i=1}^{N_c} \Phi_c(x^{I_i}) \mid \Phi_c : \mathbb{R}^n \rightarrow [0, 1], \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{N_c}{N} \right\} \quad (87)$$

$$= \left\{ h_{x \mapsto y_c} = \frac{\Phi_c(\cdot)}{N_c} \sum_{i=1}^{N_c} \Phi_c(x^{I_i}) \mid \Phi_c : \mathbb{R}^n \rightarrow [0, 1], \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{N_c}{N} \right\} \quad (88)$$

$$= \left\{ h_{x \mapsto y_c} = \sum_{i=1}^{N_c} \frac{1}{N_c} \mathcal{K}_{\Phi_c}(\cdot, x^{I_i}) \mid \Phi_c : \mathbb{R}^n \rightarrow [0, 1], \|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{N_c}{N} \right\} \quad (89)$$

$$\subset \mathcal{H}_{\Phi_c}. \quad (90)$$

For any $h_{x \mapsto y_c} \in \mathcal{M}_c$, we have

$$\|h_{x \mapsto y_c}\|_{\mathcal{H}_{\Phi_c}} = \frac{\sum_{i=1}^{N_c} \Phi_c(x^{I_i})}{N_c}, \forall h_{x \mapsto y_c} \in \mathcal{M}_c. \quad (91)$$

3.6.1 Rademacher Complexity of the Hypothesis Space

For a given data set \mathcal{D} (defined in (6)), the empirical Rademacher complexity of the hypothesis space \mathcal{M}_c is given as

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) = \frac{1}{N} \mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \sum_{i=1}^N \sigma_i h_{x \mapsto y_c}(x^i) \right], \quad (92)$$

where $\sigma = (\sigma_1, \dots, \sigma_N)$ with $\sigma_1, \dots, \sigma_N$ as the independent random variables drawn from the Rademacher distribution.

Theorem 5 (Bound on Rademacher Complexity of the Hypothesis Space). *Given a dataset \mathcal{D} , as defined in (6), we have*

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) \leq \frac{1}{\sqrt{N}}. \quad (93)$$

Thus,

$$\mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c)] \leq \frac{1}{\sqrt{N}}. \quad (94)$$

Proof. The proof is provided in Appendix L. \square

3.6.2 Error Bounds

Theorem 3 and Theorem 4 have provided error bounds for the generalized solution using operator-theoretic analysis. Further, the Rademacher complexity of the hypothesis space can be used to derive error bounds for the hypothesis space as in [42]. The results obtained by the two approaches can be combined to obtain the tighter bounds as in Theorem 6 and Theorem 7.

Theorem 6 (Prediction Error Bound for Hypothesis Space). *Given a data set $\mathcal{D} = \{(x^i, y^i) \mid i \in \{1, 2, \dots, N\}\} \sim (\mathbb{P}_{x,y})^N$, for any $h_{x \mapsto y_c} \in \mathcal{M}_c$, we have with probability at least $1 - \delta$ for any $\delta \in (0, 1)$,*

$$\begin{aligned} & \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right] \\ & \leq \min \left(\frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}, \right. \\ & \quad \left. \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[\left| y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right|^2 \right] + \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right) \right). \end{aligned} \quad (95)$$

Proof. The proof is provided in Appendix M. \square

Theorem 7 (Approximation Error Bound for Hypothesis Space). *Given a data set $\mathcal{D} = \{(x^i, y^i) \mid i \in \{1, 2, \dots, N\}\} \sim (\mathbb{P}_{x,y})^N$, for any $h_{x \mapsto y_c} \in \mathcal{M}_c$, we have with probability at least $1 - \delta$ for any $\delta \in (0, 1)$,*

$$\begin{aligned} \mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] & \leq \min \left(\frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}, \right. \\ & \quad \left. \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right) \right). \end{aligned} \quad (96)$$

Proof. The proof is provided in Appendix N. \square

Remark 6 (Comparison with the Existing Error Bounds). *Since*

$$\begin{aligned} \text{r.h.s. of inequality (95)} & \leq \underbrace{\frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}}, \\ & \quad = \text{prediction error bound of [42]} \end{aligned} \quad (97)$$

$$\begin{aligned} \text{r.h.s. of inequality (96)} & \leq \underbrace{\frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}}, \\ & \quad = \text{approximation error bound of [42]} \end{aligned} \quad (98)$$

we achieve tighter bounds on prediction and approximation errors.

3.7 Step 7: Hypothesis Implementation in Federated Setting

We now describe how the RKHS-based method is deployed across decentralized clients in a federated setup. Till-now, we have not fixed the kernel feature-map Φ_c and the corresponding hypothesis for an implementation in federated setting. Our idea is to leverage KAHMs for defining Φ_c in such a way that the corresponding hypothesis can be evaluated efficiently from the distributed training data.

3.7.1 Space Folding Property of a KAHM

KAHMs exhibit the space folding property in a sense that a KAHM associated to a given set of data samples, folds any arbitrary point in the data space around the data samples.

Theorem 8 (KAHM as a Bounded Function [40]). *The KAHM $\mathcal{A}_{\mathbf{X}}$, associated to $\mathbf{X} = [x^1 \dots x^N]^T$ with $x^1, \dots, x^N \in \mathbb{R}^n$, is a bounded function on \mathbb{R}^n such that for any $x \in \mathbb{R}^n$,*

$$\|\mathcal{A}_{\mathbf{X}}(x)\| < \|\mathbf{X}\|_2 \left(1 + \frac{nN^2}{2\|\mathbf{X}\|_F^2} \right). \quad (99)$$

Thus, the image of \mathcal{A}_X is bounded such that

$$\mathcal{A}_X[\mathbb{R}^n] \subset \left\{ x \in \mathbb{R}^n \mid \|x\| < \|\mathbf{X}\|_2 \left(1 + \frac{nN^2}{2\|\mathbf{X}\|_F^2} \right) \right\}. \quad (100)$$

Theorem 9 (KAHM Induced Distance Measure [40]). *The ratio of the distance of a point $x \in \mathbb{R}^n$ from its image under \mathcal{A}_X to the distance of x from $\{x^1, \dots, x^N\}$ evaluated as $\|[x - x^1 \dots x - x^N]\|_2$ remains upper bounded as*

$$\frac{\|x - \mathcal{A}_X(x)\|}{\|[x - x^1 \dots x - x^N]\|_2} < 1 + \frac{nN^2}{2\|\mathbf{X}\|_F^2}. \quad (101)$$

Theorem 8 establishes the boundedness property of the KAHM. Theorem 9 states that if a point x is close to the points $\{x^1, \dots, x^N\}$, then the value $\|x - \mathcal{A}_X(x)\|$ cannot be large. Theorem 8 and Theorem 9 indicate that any arbitrary point $x \in \mathbb{R}^n$ can be mapped to a point closer to data samples \mathbf{X} through a KAHM, and thus the space folding property is established. This is illustrated in Fig. 2(a), where the KAHM folds the data space around given data samples.

Definition 1 (Space Folding Measure). *To evaluate the amount of folding required for an arbitrary point $x \in \mathbb{R}^n$ to map it (by the KAHM \mathcal{A}_X) to a point closer to data samples \mathbf{X} , we define a space folding measure, $\mathcal{T}_X : \mathbb{R}^n \rightarrow [0, 1]$, associated to data samples \mathbf{X} , as*

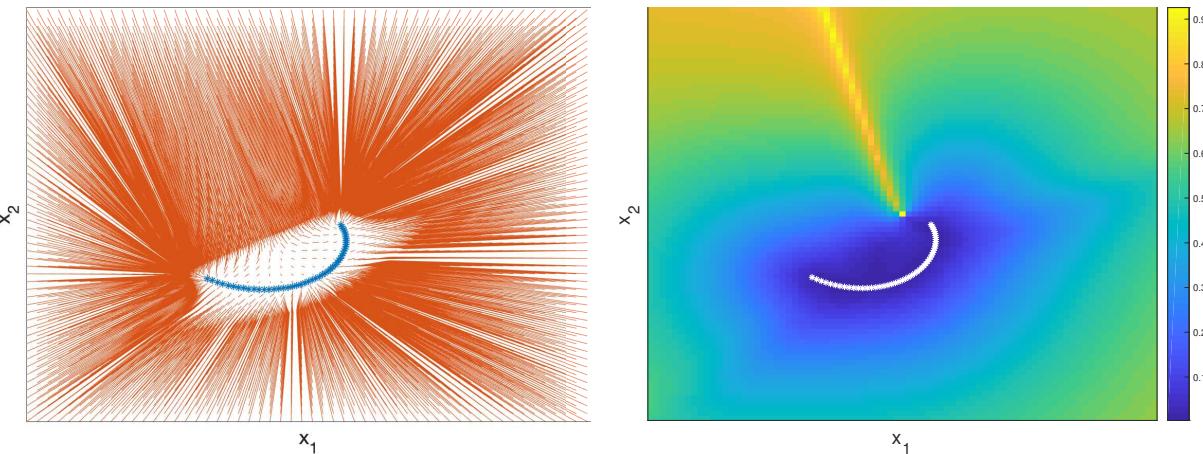
$$\mathcal{T}_X(x) := \begin{cases} \sqrt{\frac{1}{2}(|\mathcal{T}_X^{Euc}(x)|^2 + |\mathcal{T}_X^{Cos}(x)|^2)} & \text{option 1} \\ \mathcal{T}_X^{Euc}(x)\mathcal{T}_X^{Cos}(x) & \text{option 2} \\ \min(\mathcal{T}_X^{Euc}(x), \mathcal{T}_X^{Cos}(x)) & \text{option 3} \\ \max(\mathcal{T}_X^{Euc}(x), \mathcal{T}_X^{Cos}(x)) & \text{option 4} \end{cases} \quad (102)$$

where

$$\mathcal{T}_X^{Euc}(x) := 1 - \exp(-\|x - \mathcal{A}_X(x)\|) \quad (103)$$

$$\mathcal{T}_X^{Cos}(x) := \frac{1}{\pi} \arccos \left(\frac{(\mathcal{A}_X(x))^T x}{\|\mathcal{A}_X(x)\| \|x\|} \right). \quad (104)$$

The space folding measure \mathcal{T}_X combines both Euclidean distance and cosine distance to define a composite measure of the distance between x and $\mathcal{A}_X(x)$. Since there are different possibilities to combine the Euclidean and cosine distances resulting in different possible definitions of the space folding measure, (102) provides four different possibilities among others. Fig. 2(b) displays the color-plot of the space folding measure function associated to a set of 2-dimensional data samples.



(a) The given 2-dimensional data samples \mathbf{X} have been marked in blue color as “*”, and a red line connects a point to its image under KAHM (i.e. x is connected to $\mathcal{A}_X(x)$ through a line).

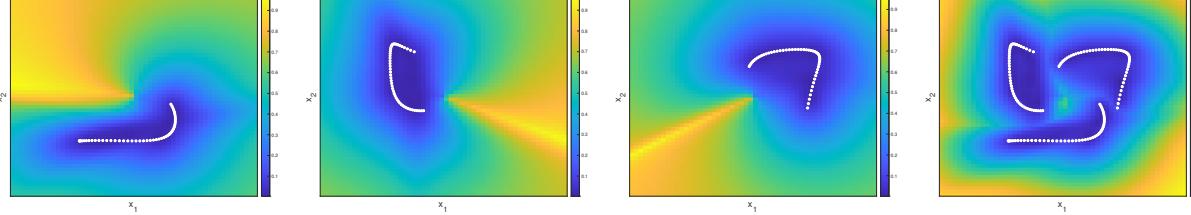
(b) The color plot of the space folding measure function \mathcal{T}_X (option 1). The given 2-dimensional data samples \mathbf{X} have been marked in white color as “*”.

Figure 2: An illustration of the space folding property possessed by a KAHM.

Definition 2 (Space Folding Measure Associated to Distributed Data). *Under the scenario that total data samples are distributed among Q number of parties such that matrix \mathbf{X}^q represents local data samples owned by q^{th} party, one possible way to define a global space folding measure, $\mathcal{T}_{\mathbf{X}^1, \dots, \mathbf{X}^Q} : \mathbb{R}^n \rightarrow [0, 1]$, associated to distributed data samples $\mathbf{X}^1, \dots, \mathbf{X}^Q$, is as follows*

$$\mathcal{T}_{\mathbf{X}^1, \dots, \mathbf{X}^Q}(x) := \min_{q \in \{1, \dots, Q\}} \mathcal{T}_{\mathbf{X}^q}(x). \quad (105)$$

Fig. 3 shows an example of the global space folding measure associated to distributed 2-dimensional data samples.



(a) The color plot of the space folding measure function $\mathcal{T}_{\mathbf{X}^1}$ (option 1) associated to data samples \mathbf{X}^1 (marked in white color as “*”).
(b) The color plot of the space folding measure function $\mathcal{T}_{\mathbf{X}^2}$ (option 1) associated to data samples \mathbf{X}^2 (marked in white color as “*”).
(c) The color plot of the space folding measure function $\mathcal{T}_{\mathbf{X}^3}$ (option 1) associated to data samples \mathbf{X}^3 (marked in white color as “*”).
(d) The color plot of the global space folding measure function $\mathcal{T}_{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3}$ associated to total data samples (marked in white color as “*”).

Figure 3: An example of the global space folding measure associated to the distributed data.

3.7.2 Defining Kernel Feature-Map Using Space Folding Measure and a Hypothesis

Our approach is to define the kernel feature-map based on the space folding measure as

$$\Phi_c^*(x) := \begin{cases} 1 & \text{if } \mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}(x) = \min_{c \in \{1, 2, \dots, C\}} \mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}(x), \\ 0 & \text{otherwise,} \end{cases} \quad (106)$$

where $\mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}$ is the global space folding measure associated to c^{th} class labelled samples that are distributed among Q different clients, and $\mathbf{X}^{c,q}$ is the matrix of c^{th} class labelled and q^{th} client owned samples. *Definition of the kernel feature-map (106) implies that $\Phi_c^*(x)$ is equal to 1, if, among all the classes, x can be mapped (by the KAHMs) to the point closer to the samples of c^{th} class with the least amount of folding.*

Remark 7 (The Interpretation of the Space Folding Kernel $\mathcal{K}_{\Phi_c^*}$). $\mathcal{K}_{\Phi_c^*}(x, x')$ will be equal to 1, if, among all the classes, both x and x' can be mapped (by the KAHMs) to the points closer to the samples of c^{th} class with the least amount of folding. Similarly, $\mathcal{K}_{\Phi_c^*}(x, x')$ will be equal to 0, if, either of x and x' cannot be mapped (by the KAHMs) to the points closer to the samples of c^{th} class with the least amount of folding. Thus, $\mathcal{K}_{\Phi_c^*}(x, x')$ estimates an association (of x and x' as together) to the c^{th} class.

Assumption 2. It is reasonable to assume for the training data samples that the global space folding measure associated to c^{th} class (i.e. $\mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}$) will take minimum values for c^{th} class labelled samples, i.e.,

$$\Phi_c^*(x^i) = \begin{cases} 1 & \text{if } i \in \mathcal{I}^c, \\ 0 & \text{if } i \notin \mathcal{I}^c. \end{cases}, \quad \forall i \in \{1, \dots, N\}. \quad (107)$$

Assumption 3. The number of training data samples is sufficiently large i.e. $N \gg 1$, so that $\|\Phi_c^*\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2$ can be approximated by sample-averaging:

$$\|\Phi_c^*\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{1}{N} \sum_{i=1}^N |\Phi_c^*(x^i)|^2. \quad (108)$$

Remark 8 (Consistency of Normalization Condition (85) with the Space Folding Kernel). *Using (107) in (108), we have*

$$\|\Phi_c^*\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{N_c}{N}. \quad (109)$$

That is, KAHM-induced feature-map Φ_c^* satisfies the normalization condition (85) under Assumptions 2-3.

Proposition 1. Under Assumption 2 and Assumption 3, we have

$$\Phi_c^* \in \mathcal{M}_c. \quad (110)$$

Proof. Define

$$h_{x \mapsto y_c}^* := \frac{\Phi_c^*(\cdot)}{N_c} \sum_{i=1}^{N_c} \Phi_c^*(x^{I_i^c}). \quad (111)$$

It follows from (109) and (111) that

$$h_{x \mapsto y_c}^* \in \mathcal{M}_c. \quad (112)$$

Using (107) in (111), we have

$$h_{x \mapsto y_c}^* = \Phi_c^*. \quad (113)$$

Combining (112) and (113) leads to (110). \square

Proposition 2 (Approximation Error Bound for Φ_c^*). Given a data set $\mathcal{D} = \{(x^i, y^i) \mid i \in \{1, 2, \dots, N\}\} \sim (\mathbb{P}_{x,y})^N$, under Assumption 2 and Assumption 3, we have with probability at least $1 - \delta$ for any $\delta \in (0, 1)$,

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|\Phi_c^*(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] \leq \min \left(\frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}, \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right) \right). \quad (114)$$

Proof. Since $\Phi_c^* \in \mathcal{M}_c$, it follows from Theorem 7 that we have with probability at least $1 - \delta$ for any $\delta \in (0, 1)$,

$$\begin{aligned} \mathbb{E}_{x \sim \mathbb{P}_x} \left[|\Phi_c^*(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] &\leq \min \left(\frac{1}{N} \sum_{i=1}^N |y_c^i - \Phi_c^*(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}, \right. \\ &\quad \left. \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right) \right). \end{aligned} \quad (115)$$

Due to (107),

$$\Phi_c^*(x^i) = y_c^i, \quad \forall i \in \{1, \dots, N\}. \quad (116)$$

Hence the result follows. \square

Remark 9 (Practical Significance of Proposition 2). Proposition 2 allows to make the following approximation:

$$\Phi_c^*(x) \approx \mathbb{P}_{y|x}(y_c = 1|x). \quad (117)$$

That is, $\Phi_c^*(x)$ estimates the probability that x is associated to the c^{th} class.

3.7.3 Federated Learning Applications

The fact, $\Phi_c^*(x)$ (which is the estimated probability of x being associated to the c^{th} class) can be evaluated from the distributed data, is leveraged for FL, as illustrated in Fig. 4. The hypothesis Φ_c^* is inferred for all $c \in \{1, 2, \dots, C\}$ from the locally computed space folding measures using (106). The global classifier, $\hat{\mathcal{C}} : \mathbb{R}^n \rightarrow \{1, 2, \dots, C\}$, is defined as

$$\hat{\mathcal{C}}(x) := \operatorname{argmax}_{c \in \{1, 2, \dots, C\}} \Phi_c^*(x). \quad (118)$$

Remark 10 (Batch KAHM Modeling for Enhanced Accuracy). To enhance the KAHM modeling accuracy of each class's data samples (which is crucial for datasets with long-tailed imbalance), the total samples are partitioned into subsets and each subset is modelled through a separate KAHM. If $|\mathcal{I}^{c,q}|$ (i.e. the number of c^{th} class labelled samples that are owned by client q) is more than a specified number N_b (representing the batch-size of samples to be modeled by a KAHM), then we have

$$S_{c,q} = \lceil |\mathcal{I}^{c,q}| / N_b \rceil \quad (119)$$

$$\mathbf{X}^{c,q} = \begin{bmatrix} \mathbf{X}_1^{c,q} \\ \vdots \\ \mathbf{X}_{S_{c,q}}^{c,q} \end{bmatrix} \quad (120)$$

$$\mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x) = \min \left(\mathcal{T}_{\mathbf{X}_1^{c,q}}(x), \dots, \mathcal{T}_{\mathbf{X}_{S_{c,q}}^{c,q}}(x) \right). \quad (121)$$

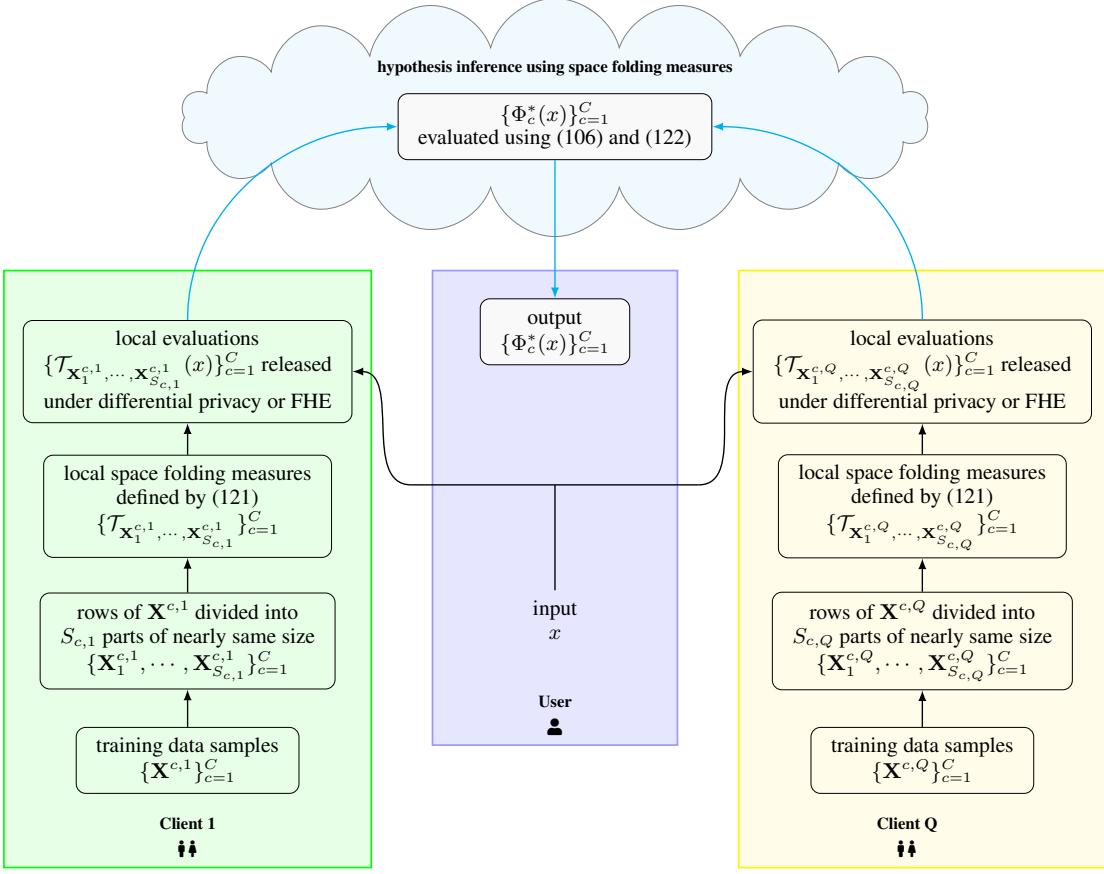


Figure 4: The proposed space folding measure-based methodology, referred to as SFM, estimates for a given input x the probability of c^{th} class, $\Phi_c^*(x)$, without imposing statistical assumptions on clients' data distributions, thereby ensuring robustness towards statistical heterogeneity.

That is, the total data samples (stored in the rows of matrix $\mathbf{X}^{c,q}$) are partitioned into $S_{c,q}$ sub-matrices of nearly same size (i.e. each sub-matrix has nearly N_b number of rows), where $S_{c,q}$ equals the rounding of $|\mathcal{I}^{c,q}|/N_b$ towards the nearest integer, and the space folding measure is defined by aggregating all of the $S_{c,q}$ individual measures associated to the $S_{c,q}$ sub-matrices. In this case, the global space folding measure associated to c^{th} class labelled samples $\mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}$ is defined as

$$\mathcal{T}_{\mathbf{X}^{c,1}, \mathbf{X}^{c,2}, \dots, \mathbf{X}^{c,Q}}(x) := \min_{q \in \{1, \dots, Q\}} \mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x). \quad (122)$$

Remark 11 (Data Partitioning via Clustering for Batch Processing of Big Data). *To address the computational challenge of processing big datasets, the previous studies [40, 42] have suggested clustering as a method for partitioning a large number of data samples into subsets for their batch processing. That is, $\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}$ are instead obtained by the clustering the rows of $\mathbf{X}^{c,q}$.*

Remark 12 (A Generalization of the Federated Learning Method of [42]). *It can be observed that the proposed federated learning methodology generalizes the method of [42] by considering a generalized space folding measure $\mathcal{T}_{\mathbf{X}}(x)$ instead of the distance measure $\|x - \mathcal{A}_{\mathbf{X}}(x)\|$ considered in [42]. That is, the solution of [42] is obtained with our approach by defining $\mathcal{T}_{\mathbf{X}}(x) := \|x - \mathcal{A}_{\mathbf{X}}(x)\|$.*

Remark 13 (Clients with Missing Classes). *If the q^{th} client has zero c^{th} class labelled samples, then $\mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x)$ equals the highest possible value of 1, i.e., the space folding measure is defined as*

$$\mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x) = \{1, \text{ if } |\mathcal{I}^{c,q}| = 0.\} \quad (123)$$

Remark 14 (Communication and Computation Efficiency). *The federated learning methodology, as suggested in Fig. 4, is communication and computation efficient since the space folding measures are computed using KAHMs and KAHMs are efficiently built [40] from the local data samples requiring neither any gradients nor any communication with the server.*

3.8 Step 8: Differentially Private Federated Setting

Existing differentially private FL approaches often require gradient clipping and complex per-round privacy budgeting. In contrast, our method relies only on scalar outputs (i.e., space folding measures), enabling a natural (ϵ, δ) -differential privacy mechanism that directly acts on these real-valued summaries and thus avoids the need for per-round gradient clipping and privacy accounting. The FL methodology (as illustrated in Fig.4) is made differentially private by ensuring that the evaluation of the space folding measure is differentially private with respect to the local dataset, and the subsequent aggregation steps then inherit privacy by post-processing property of differential privacy. It follows from (121) that $\mathcal{T}_{\mathbf{X}_s^{c,q}}(x)$ must be differentially private with respect to $\mathbf{X}_s^{c,q}$ for all $s \in \{1, \dots, S_{c,q}\}$. We need to basically address the privacy of data samples in matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$ that may be leaked during inference through the output of space folding measure function $\mathcal{T}_{\mathbf{X}} : \mathbb{R}^n \rightarrow [0, 1]$. However, estimating the sensitivity of space folding measure is challenging. Thus, we consider the more practical approach of approximating \mathbf{X} under differential privacy followed by a smoothing, while ensuring that differentially private smoothed data points are as close to original data points as possible. This approach leads to defining a private version of the space folding measure:

Definition 3 (Private Space Folding Measure). *The private version of space folding measure, $\mathcal{T}_{\mathbf{X}}^+ : \mathbb{R}^n \rightarrow [0, 1]$, is defined as*

$$\mathcal{T}_{\mathbf{X}}^+(\cdot) := \mathcal{T}_{\mathcal{F}(\mathbf{X} + \mathbf{V})}(\cdot), \quad (124)$$

where $\mathbf{V} \in \mathbb{R}^{N \times n}$ is a random real matrix added to \mathbf{X} for preserving the privacy of elements of \mathbf{X} , such that elements of \mathbf{V} are independently distributed from a distribution $\mathbb{P}_v : \mathcal{B}(\mathbb{R}) \rightarrow [0, 1]$:

$$\mathbf{V} \sim (\mathbb{P}_v)^{N \times n}, \quad (125)$$

where $v \in \mathbb{R}$ is the random noise with some distribution \mathbb{P}_v and $\mathcal{F} : \mathbb{R}^{N \times n} \rightarrow \mathbb{R}^{N \times n}$ is a matrix-valued function, referred to as smoothing function, meant for mitigating the effect of added noise.

The noise adding mechanism and smoothing function, involved in Definition 3, will be designed in subsection 3.8.1 and subsection 3.8.2.

3.8.1 Optimal Noise Adding Mechanism for Differential Privacy

The output of $\mathcal{T}_{\mathbf{X}}^+$ is a random variable defined as

$$t_{\mathbf{X}}^+ := \mathcal{T}_{\mathcal{F}(\mathbf{X} + \mathbf{V})}(x), \quad \mathbf{V} \sim (\mathbb{P}_v)^{N \times n} \quad (126)$$

Let the distribution of $t_{\mathbf{X}}^+$ be denoted by $\mathbb{P}_{t_{\mathbf{X}}^+} : \mathcal{B}([0, 1]) \rightarrow [0, 1]$.

Definition 4 (d -Adjacent Matrices). *Two matrices $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{N \times n}$ are d -adjacent if for a given $d \in \mathbf{R}_+$, there exist $i_0 \in \{1, 2, \dots, N\}$ and $j_0 \in \{1, 2, \dots, n\}$ such that for all $i \in \{1, 2, \dots, N\}$ and $j \in \{1, 2, \dots, n\}$,*

$$|(\mathbf{X})_{i,j} - (\mathbf{X}')_{i,j}| \leq \begin{cases} d & \text{if } i = i_0 \text{ and } j = j_0, \\ 0 & \text{otherwise} \end{cases} \quad (127)$$

Thus, two d -adjacent matrices differ by only one element and the difference is bounded by a scalar $d > 0$.

Definition 5 ((ϵ, δ) -Differential Privacy for $\mathcal{T}_{\mathbf{X}}^+$). *The space folding measure $\mathcal{T}_{\mathbf{X}}^+ : \mathbb{R}^n \rightarrow [0, 1]$ is (ϵ, δ) -differentially private if*

$$\mathbb{P}_{t_{\mathbf{X}}^+}(\mathcal{O}) \leq \exp(\epsilon) \mathbb{P}_{t_{\mathbf{X}'}^+}(\mathcal{O}) + \delta \quad (128)$$

for any $\mathcal{O} \in \mathcal{B}([0, 1])$ and d -adjacent matrices $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{N \times n}$.

Result 1 (Optimal Noise for (ϵ, δ) -Differential Privacy [41, 40]). *The distribution of noise v , that minimizes expected noise magnitude together with satisfying the sufficient condition for (ϵ, δ) -differential privacy of $\mathcal{T}_{\mathbf{X}}^+$, is given as*

$$\mathbb{P}_v((-\infty, v]) = \begin{cases} \frac{1-\delta}{2} \exp(-\frac{\epsilon}{d}v) & \text{if } v < 0, \\ \frac{1+\delta}{2} & \text{if } v = 0, \\ 1 - \frac{1-\delta}{2} \exp(-\frac{\epsilon}{d}v) & \text{if } v > 0. \end{cases} \quad (129)$$

The method of inverse transform sampling can be used to generate random samples from (129) and approximate \mathbf{X} under differential privacy as

$$\mathbf{X}^+ = \mathbf{X} + \mathbf{V}, \quad \mathbf{V} \sim (\mathbb{P}_v)^{N \times n}. \quad (130)$$

3.8.2 Kernel-Based Smoothing of Data Samples

Given a dataset $\{x^i \in \mathbb{R}^n\}_{i=1}^N$ (that can be equivalently represented as matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$), a kernel-based smoothing can be represented as

$$\widehat{\mathbf{X}} = \mathbf{H}_\mathbf{X}^T \mathbf{X}, \quad (131)$$

where $\mathbf{H}_\mathbf{X} \in \mathbb{R}^{N \times N}$ is a matrix defined as

$$(\mathbf{H}_\mathbf{X})_{i,j} = h_\mathbf{X}^i(\mathbf{P}_\mathbf{X} x^j), \quad (132)$$

where $\mathbf{P}_\mathbf{X}$ is an encoding matrix (computed by Algorithm 1 of Appendix A), and $h_\mathbf{X}^i(\mathbf{P}_\mathbf{X} x^j)$, given by (155), evaluates kernel-smoothed membership of x^j to x^i .

Definition 6 (A Kernel-Based Smoother). A *kernel-based smoother*, $\mathcal{S} : \mathbb{R}^{N \times n} \rightarrow \mathbb{R}^{N \times n}$, is defined as

$$\mathcal{S}(\mathbf{X}) := \mathbf{H}_\mathbf{X}^T \mathbf{X}. \quad (133)$$

Proposition 3 (Smoothing Property of \mathcal{S}). There exists a $\beta \in (0, 1)$ such that

$$\|\mathcal{S}(\mathbf{X})\|_2 \leq \beta \|\mathbf{X}\|_2 < \|\mathbf{X}\|_2. \quad (134)$$

Proof. It can be seen using (155) that

$$\mathbf{H}_\mathbf{X} = (\mathbf{K}_\mathbf{X} + \lambda_\mathbf{X}^* \mathbf{I}_N)^{-1} \mathbf{K}_\mathbf{X}, \quad (135)$$

where the kernel matrix $\mathbf{K}_\mathbf{X}$ and regularization parameter $\lambda_\mathbf{X}^*$ are defined by (152) and (149), respectively. The spectral decomposition of the real symmetric positive define matrix $\mathbf{K}_\mathbf{X}$ is given as $\mathbf{K}_\mathbf{X} = E_\mathbf{X} \Lambda_\mathbf{X} E_\mathbf{X}^T$, where $E_\mathbf{X}$ is an orthogonal matrix (i.e. $E_\mathbf{X}^T E_\mathbf{X} = E_\mathbf{X} E_\mathbf{X}^T = \mathbf{I}_N$), and $\Lambda_\mathbf{X}$ is the diagonal matrix of eigenvalues i.e. $\Lambda_\mathbf{X} = \text{diag}(\text{eig}_1(\mathbf{K}_\mathbf{X}), \dots, \text{eig}_N(\mathbf{K}_\mathbf{X}))$. Now, we can express $\mathbf{H}_\mathbf{X}$ as

$$\mathbf{H}_\mathbf{X} = E_\mathbf{X} (\Lambda_\mathbf{X} + \lambda_\mathbf{X}^* \mathbf{I}_N)^{-1} \Lambda_\mathbf{X} E_\mathbf{X}^T. \quad (136)$$

The i^{th} eigenvalue of $\mathbf{H}_\mathbf{X}$ is given as

$$\text{eig}_i(\mathbf{H}_\mathbf{X}) = \frac{\text{eig}_i(\mathbf{K}_\mathbf{X})}{\text{eig}_i(\mathbf{K}_\mathbf{X}) + \lambda_\mathbf{X}^*} > 0. \quad (137)$$

It follows from (135) that $\mathbf{H}_\mathbf{X}$, being a product of commuting symmetric matrices, is also symmetric. Therefore,

$$\|\mathbf{H}_\mathbf{X}\|_2 = \max_{i \in \{1, 2, \dots, N\}} |\text{eig}_i(\mathbf{H}_\mathbf{X})| \quad (138)$$

$$= \max_{i \in \{1, 2, \dots, N\}} \frac{\text{eig}_i(\mathbf{K}_\mathbf{X})}{\text{eig}_i(\mathbf{K}_\mathbf{X}) + \lambda_\mathbf{X}^*}. \quad (139)$$

If we define

$$\beta = \max_{i \in \{1, 2, \dots, N\}} \frac{\text{eig}_i(\mathbf{K}_\mathbf{X})}{\text{eig}_i(\mathbf{K}_\mathbf{X}) + \lambda_\mathbf{X}^*} < 1, \quad (140)$$

then we have

$$\|\mathbf{H}_\mathbf{X}\|_2 = \beta. \quad (141)$$

Now, (134) follows from (133) and (141). \square

Inequality (134) establishes the smoothness property by ensuring that the norm of smoothed data matrix remains smaller than that of input data matrix. The degree-of-smoothness can be enhanced by repeatedly applying the smoother, leading to m -fold composition of \mathcal{S} on data:

$$\mathcal{S}^m(\mathbf{X}) = \mathcal{S}(\mathcal{S}^{m-1}(\mathbf{X})), \quad m \in \{1, 2, \dots\}, \quad (142)$$

$$\mathcal{S}^0(\mathbf{X}) = \mathbf{X}. \quad (143)$$

Our idea is to apply m -fold composition of \mathcal{S} on noisy data samples \mathbf{X}^+ (130), with m chosen optimally to minimize the difference of smoothed-noisy data from noise-free data. That is, the smoothing function \mathcal{F} (used in Definition 3 of differentially private space folding measure) is defined as

$$\mathcal{F} = \mathcal{S}^{m^*}, \quad (144)$$

where $m^* \in \{1, 2, \dots\}$ is such that

$$\|\mathcal{S}^1(\mathbf{X}^+) - \mathbf{X}\|_F > \|\mathcal{S}^2(\mathbf{X}^+) - \mathbf{X}\|_F > \dots > \|\mathcal{S}^{m^*}(\mathbf{X}^+) - \mathbf{X}\|_F \leq \|\mathcal{S}^{m^*+1}(\mathbf{X}^+) - \mathbf{X}\|_2. \quad (145)$$

The inequalities (145) imply that *an iteration of smoothing function \mathcal{S} is applied only if it reduces the mismatch between smoothed-noisy data and noise-free data*.

3.8.3 Scope of Privacy Guarantees

In the proposed FL protocol, each client locally applies the noise-adding mechanism exactly once to its data matrix $\mathbf{X}_s^{c,q}$, yielding a noise-perturbed matrix $\mathbf{X}_s^{c,q} + \mathbf{V}_s^{c,q}$. This noise-perturbed data matrix is then smoothed and subsequently used to construct a KAHM, and induce the associated space folding measure $\mathcal{T}_{\mathcal{F}(\mathbf{X}_s^{c,q} + \mathbf{V}_s^{c,q})}(\cdot)$ for an aggregation and inference of the global model. Our method does not require any iterative sampling of clients' raw data for federated training and does not involve multiple communication rounds. The smoothing, KAHM construction, and the model inference steps operate solely on $\mathbf{X}_s^{c,q} + \mathbf{V}_s^{c,q}$, and by the post-processing property of differential privacy, these steps cannot weaken the (ϵ, δ) -DP guarantee already established for the release of $\mathbf{X}_s^{c,q} + \mathbf{V}_s^{c,q}$. *The (ϵ, δ) -DP guarantee established for the single application of the noise-adding mechanism thus fully characterizes the privacy of the entire training and inference pipeline, and therefore no per-round privacy accounting or multi-round composition analysis is needed.*

3.9 Step 9: Secure Federated Learning with FHE

In privacy-critical domains, inference must often be performed on encrypted data (e.g., via FHE), which prohibits complex operations. Since our aggregation involves only scalar space folding measures, the inference is implementable using basic arithmetic gates supported by FHE. The FL methodology (as illustrated in Fig. 4) can be secured against untrustworthy server by sharing fully homomorphically encrypted local evaluations (of the space folding measure) with the server for an inference of the global model on encrypted data. The computational efficiency stems from the fact that the space folding measure, unlike high-dimensional gradients or model parameters, is a scalar. Moreover, inference in the encrypted space is likewise not computationally demanding. Since the inference of hypothesis from the locally computed space folding measures (see (106)) is not arithmetic-heavy and can be expressed as a boolean circuit, TFHE [12] is selected as the FHE scheme, owing to its ability to evaluate binary gates with exceptionally low latency. However, the encryption of locally evaluated space folding measure (i.e. $\mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x) \in [0, 1]$) requires encoding $\mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x)$ as unsigned p -bits (e.g. $p = 16$) integer, i.e., $\lceil (2^p - 1) \mathcal{T}_{\mathbf{X}_1^{c,q}, \dots, \mathbf{X}_{S_{c,q}}^{c,q}}(x) \rceil$.

3.9.1 Computational Efficient FHE Secured Inference of Global Model

The inference of the global model using (106) and (122) involves performing the `minimum` operation $Q \times C$ times and `equality-comparison` operation C times on unsigned integers that encode the space folding measure evaluations. According to the TFHE-rs library benchmarks [80], operations on fully homomorphically encrypted 16-bit unsigned integers demonstrate practical performance on modern CPU hardware. Specifically, the `minimum` operations require approximately 96.4ms, while the `equality-comparison` requires around 31.3ms, when executed on an AMD EPYC 9R14 @ 2.60 GHz (AWS hpc7a.96xlarge) CPU. These measurements correspond to the default high-level parameter set in TFHE-rs, which provides at least 128-bit security under the IND-CPA-D model, with a bootstrapping failure probability not exceeding 2^{-128} . This configuration therefore balances strong cryptographic guarantees with practical computational efficiency for fully homomorphic operations on 16-bit encrypted data.

3.10 Summary

Together, these nine steps demonstrate that our operator-theoretic approach offers a unified kernel framework for gradient-free FL that is communication-efficient, supports rigorous differential privacy mechanisms on scalar space folding summaries, and is compatible with secure inference via FHE, while being grounded in non-asymptotic finite-sample analysis and hypothesis-space complexity bounds.

4 Experiments

We design our experimental study to address the following questions:

- **Q1 (Robustness to heterogeneity and imbalance).** How robust is the proposed operator-theoretic, gradient-free federated learning method to long-tailed class imbalance and non-IID label distributions across clients?
- **Q2 (Privacy-utility trade-off).** How well does the method perform under differential privacy constraints, and what is the impact of the proposed smoothing mechanism?
- **Q3 (Secure inference efficiency).** Is secure inference of the global model via fully homomorphic encryption (FHE) computationally feasible on commodity hardware?
- **Q4 (Sensitivity to design choices).** How sensitive is performance to the choice of space folding variant, batch size, and feature embeddings?

4.1 Datasets

We evaluate the proposed method on four benchmark datasets.

20Newsgroup [67] This text classification dataset is a collection of newsgroup documents split across 20 distinct topics. The “bydate” version of the dataset contains 11314 training documents and 7532 test documents.

XGLUE-NC [46] This is a multilingual news classification benchmark dataset containing English, German, Spanish, French, and Russian language text documents belonging to 10 distinct categories. Each language is represented by 10000 training examples and 10000 test examples.

CIFAR-10-LT The original CIFAR-10 dataset [38] contains 50000 training images and 10000 test images divided across 10 classes. Following prior work [9], the original CIFAR-10 dataset is turned into a long-tailed imbalance with imbalance ratio (which is the ratio between sample sizes of the most frequent and least frequent class) $\rho \in \{10, 50, 100\}$.

CIFAR-100-LT Following [71], the original CIFAR-100 dataset [37], containing 100 classes with 500 training images and 100 test images in each class, is turned into a long-tailed imbalance with imbalance ratio $\rho \in \{10, 50, 100\}$.

4.2 Preprocessing and Feature Extraction

In all experiments, the proposed FL method operates on fixed feature vectors extracted from existing encoders such that encoders are not updated during training. Our method therefore plays the role of a gradient-free “head” on top of pretrained feature extractors.

Image datasets. For CIFAR-10-LT and CIFAR-100-LT, a 2048-dimensional feature vector is obtained for each image from the activations of the “avg_pool” layer (the final average pooling layer preceding the fully connected layer) of a pretrained ResNet-50 neural network [51]. The image feature vectors are processed through the hyperbolic tangent function to limit values within the range $[-1, 1]$.

20Newsgroup For each document, “mxbai-embed-large” English sentence embedding model [3] is used to extract 1024-dimensional feature vector. Since the raw embeddings exhibit relatively small variance across dimensions, we rescale them along all dimensions by a factor of 10.

XGLUE-NC For the multilingual setting, we first extract 512-dimensional feature vectors using “distiluse-base-multilingual-cased-v2” multilingual sentence embedding model [64] and rescale them by a factor of 10 to increase variance. We additionally compute 768-dimensional embeddings using the “paraphrase-multilingual” sentence embedding model [65], again rescaled by a factor of 10. Concatenating both embeddings yields a 1280-dimensional feature vector for each document, which is used in the FL experiments.

4.3 Client Partition

We follow established client-partitioning protocols to match prior FL studies on these benchmarks.

20Newsgroup Following the experimental setting as in [36], the training documents are distributed across 100 clients in a non-IID manner using Dirichlet distribution with concentration parameter $\alpha \in \{0.1, 1, 5\}$.

XGLUE-NC Again following [36], 100 clients are divided into five distinct groups, with a specific language assigned to a group such that all training examples of that language are distributed among the clients of the group in a non-IID manner using Dirichlet distribution with concentration parameter $\alpha \in \{0.5, 2, 5\}$.

CIFAR-10-LT and CIFAR-100-LT For both long-tailed image benchmarks, like previous study [71], a non-IID scenario of training images distribution across 20 clients is simulated using Dirichlet distribution with concentration parameter $\alpha = 0.5$.

4.4 FL Protocol

Our FL protocol, illustrated in Fig. 4, is used for all experiments. A key feature of the proposed framework is that it requires only a small number of method-specific choices. Beyond selecting a variant of space folding measure in (102)

Method	$\alpha = 5$	$\alpha = 1$	$\alpha = 0.1$
SFM (proposed)	85.3	84.7	84.7
Adapter [36]	69.1	65.5	56.1
LoRA [36]	69.5	67.7	56.6
Compacter [36]	65.9	62.8	50.1
Prompt-tuning [36]	51.6	46.4	28.2
BitFit [36]	67.1	66.5	55.1
AdaMix [36]	68.7	65.3	54.5
C2A [36]	<u>71.6</u>	<u>70.4</u>	<u>61.0</u>

Table 1: Comparison of the test data accuracy (%) obtained by proposed method against previously available results [36] of federated learning experiments on 20Newsgroup dataset under the non-IID label distribution scenarios.

Method	$\rho = 100$	$\rho = 50$	$\rho = 10$
SFM (proposed)	80.99	83.86	87.95
FedAvg [71]	56.17	59.36	77.45
FedAvgM [71]	52.03	57.11	70.81
FedProx [71]	56.92	60.89	76.53
FedDF [71]	55.15	58.74	76.51
FedBE [71]	55.79	59.55	77.78
CCVR [71]	69.53	71.89	78.48
FedNova [71]	57.79	63.91	77.79
Fed-Focal Loss [71]	53.83	57.42	73.74
Ratio Loss [71]	59.75	64.77	78.14
FedAvg+ τ -norm [71]	49.95	51.41	72.08
CREFF [71]	<u>70.55</u>	<u>73.08</u>	<u>80.71</u>

Table 2: Comparison of the test data accuracy (%) obtained by proposed method against previously available results [71] of federated learning experiments on CIFAR-10-LT dataset under the long-tailed imbalance and non-IID label distribution scenarios.

and the batch-size for local processing (Remark 10), no additional hyperparameters specific to our method are tuned. In all main experiments, we adopt option 1 in (102) to define the space folding measure \mathcal{T}_X . For the 20Newsgroup and XGLUE-NC datasets, we use a batch size of $N_b = 100$. For CIFAR-10-LT and CIFAR-100-LT, which exhibit pronounced long-tailed class imbalance, we set $N_b = 20$ and model each batch of 20 samples via a separate KAHM to better preserve minority classes. The effect of varying N_b is further examined in an ablation study (Table 9).

The proposed FL method performs a single aggregation of scalar space folding summaries at the server, rather than iterative gradient exchanges, and thus operates in a communication-efficient, gradient-free regime.

4.5 Software and Reproducibility

All experiments were conducted in MATLAB R2024a. The reported numbers correspond to reference runs on an Apple iMac (M1, 2021) with 8 GB RAM. We release the implementation so that the experimental results can be reproduced from the source code, which is publicly available at:

<https://drive.mathworks.com/sharing/4cef6387-1a62-46c8-a7e7-a7439bbbd9ef>.

We considered GitHub for hosting, but several precomputed embedding matrices and auxiliary .mat files used in our pipelines exceed the 100 MB per-file size limit imposed on standard GitHub repositories. Using MATLAB Drive avoids splitting the material across multiple services or requiring reviewers to configure Git Large File Storage, and therefore offers a more practical way to distribute the full reproducibility package, including large precomputed artifacts.

4.6 Results

Unlike most prior FL studies on these benchmarks, our method operates exclusively on feature vectors derived from pretrained encoders and does not update the encoder parameters. Consequently, the comparisons should be interpreted

Method	$\alpha = 5$	$\alpha = 2$	$\alpha = 0.5$
SFM (proposed)	82.5	82.2	82.2
Adapter [36]	78.6	75.0	74.3
LoRA [36]	80.4	78.4	74.6
Compacter [36]	75.9	73.4	71.0
Prompt-tuning [36]	61.2	60.6	58.0
BitFit [36]	78.4	76.8	72.1
AdaMix [36]	79.6	79.1	76.6
C2A [36]	82.8	82.2	80.2

Table 3: Comparison of the test data accuracy (%) obtained by proposed method against previously available results [36] of federated learning experiments on XGLUE-NC dataset under the non-IID label distribution scenarios.

Method	$\rho = 100$	$\rho = 50$	$\rho = 10$
SFM (proposed)	46.16	50.92	64.98
FedAvg [71]	30.34	36.35	45.87
FedAvgM [71]	30.80	35.33	44.66
FedProx [71]	31.67	36.30	46.10
FedDF [71]	31.43	36.22	46.19
FedBE [71]	31.97	36.39	46.25
CCVR [71]	33.43	36.98	46.88
FedNova [71]	32.64	36.62	46.75
Fed-Focal Loss [71]	30.67	35.25	45.52
Ratio Loss [71]	32.95	36.88	46.79
FedAvg+ τ -norm [71]	26.22	33.71	43.65
CREFF [71]	34.67	37.64	47.08

Table 4: Comparison of the test data accuracy (%) obtained by proposed method against previously available results [71] of federated learning experiments on CIFAR-100-LT dataset under the long-tailed imbalance and non-IID label distribution scenarios.

ϵ	20Newsgroup		XGLUE-NC	
	$\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathbf{X}+\mathbf{v}}$	$\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathcal{F}(\mathbf{X}+\mathbf{v})}g$	$\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathbf{X}+\mathbf{v}}$	$\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathcal{F}(\mathbf{X}+\mathbf{v})}$
1	69.46	70.03	76.76	77.53
1.5	74.03	74.55	78.31	79.47
2	78.52	78.89	79.89	80.34
3	78.89	79.22	80.89	81.10
5	80.88	81.35	81.62	81.60
8	83.05	82.97	82.03	81.89

Table 5: Test data accuracy (%) obtained by proposed method during differential privacy federated learning experiments on 20Newsgroup and XGLUE-NC datasets under the non-IID label distribution scenarios with $\alpha = 0.1$ for 20Newsgroup and $\alpha = 0.1$ for XGLUE-NC. For each value of privacy-loss bound ϵ (and fixed $\delta = 10^{-5}$), the performance was evaluated under two scenarios: 1) when the noise-perturbed samples were not smoothed (i.e. $\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathbf{X}+\mathbf{v}}$), 2) when the noise-perturbed samples were smoothed (i.e. $\mathcal{T}_{\mathbf{X}}^+ = \mathcal{T}_{\mathcal{F}(\mathbf{X}+\mathbf{v})}$).

Precision	Computational Time (ms)		Test Data Accuracy (%)	
	minimum	equality-comparison	20Newsgroup	XGLUE-NC
8-bits	101	44	84.75	81.76
16-bits	195	82	84.75	81.76

Table 6: Results of the FHE secured federated learning experiments on 20Newsgroup and XGLUE-NC datasets under the non-IID label distribution scenarios with $\alpha = 0.1$ for 20Newsgroup and $\alpha = 0.1$ for XGLUE-NC. The FHE secured inference of the global model involves performing `minimum` operation $Q \times C$ times and `equality-comparison` operations C times, where Q is the number of clients and C is the number of classes. The reported computational time is required by an iMac (M1, 8 GB RAM) for performing `minimum` and `equality-comparison` operations on fully homomorphically encrypted integers using TFHE-rs Rust library. These values correspond to the default high-level parameter set in TFHE-rs, which provides at least 128-bit security under the IND-CPA-D model, with a bootstrapping failure probability not exceeding 2^{-128} .

Method	$\rho = 100$	$\rho = 50$	$\rho = 10$
SFM-1	46.16	50.92	64.98
SFM-2	46.17	50.92	65.01
SFM-3	46.17	50.92	65.01
SFM-4	45.58	50.14	64.71

Table 7: Comparison of the test data accuracy (%) obtained by proposed SFM-i method (where i denotes the selected option in the definition of \mathcal{T}_X in Equation (102)) in federated learning experiments on CIFAR-100-LT dataset under the long-tailed imbalance and non-IID label distribution scenarios.

Method	$\alpha = 5$	$\alpha = 2$	$\alpha = 0.5$
SFM-1	82.54	82.21	82.20
SFM-2	82.53	82.19	82.20
SFM-3	82.53	82.19	82.20
SFM-4	82.68	82.27	82.31

Table 8: Comparison of the test data accuracy (%) obtained by proposed SFM-i method (where i denotes the selected option in the definition of \mathcal{T}_X in Equation (102)) in federated learning experiments on XGLUE-NC dataset under the non-IID label distribution scenarios.

N_b	$\rho = 100$	$\rho = 50$	$\rho = 10$
20	46.16	50.92	64.98
50	41.34	44.77	58.17
100	38.93	42.07	54.32

Table 9: Comparison of the test data accuracy (%) obtained by proposed method across varying batch-size N_b in federated learning experiments on CIFAR-100-LT dataset under the long-tailed imbalance and non-IID label distribution scenarios.

Embedding Model	$\alpha = 5$	$\alpha = 2$	$\alpha = 0.5$
mdl ₁ (distiluse-base-multilingual-cased-v2)	81.79	81.64	82.02
mdl ₂ (paraphrase-multilingual)	81.33	80.78	81.00
mdl ₁ + mdl ₂	82.54	82.21	82.20

Table 10: Comparison of the test data accuracy (%) obtained by proposed method across different embedding models in federated learning experiments on XGLUE-NC dataset under the non-IID label distribution scenarios.

in a head-only setting: we ask whether our gradient-free, operator-theoretic FL aggregation of fixed embeddings can be competitive with, and in several cases superior to, strong gradient-based FL baselines that optimize full or partially trainable models under matched data-distribution scenarios. This setup reflects applications where the encoder has already been trained and validated (or is provided by a third party), and only the task-specific prediction head is subject to federated training. A comparison of full end-to-end pipelines, including joint representation learning from raw text and image data, would require extending the operator-theoretic gradient-free kernel framework to encoder training, which we leave as an important direction for future work.

For 20Newsgroup and XGLUE-NC, we compare against the parameter-efficient fine-tuning baselines of [36], and for CIFAR-10-LT and CIFAR-100-LT against the gradient-based FL methods of [71]. Tables 1, 2, 3, and 4 report the experimental results on 20Newsgroup, CIFAR-10-LT, XGLUE-NC, and CIFAR-100-LT, respectively. The top two performances have been highlighted. The results of differentially private federated learning experiments on 20Newsgroup and XGLUE-NC datasets are provided in Table 5. The goal of differentially private federated learning experiments was to study the effect of data smoothing mechanism on the performance. Table 6 presents the results of FHE secured federated learning experiments on 20Newsgroup and XGLUE-NC datasets. The results of the experiments studying different variants of the space folding measure \mathcal{T}_X are provided in Table 7 and Table 8 for CIFAR-100-LT and XGLUE-NC, respectively. The effect of the batch-size N_b is experimentally studied in Table 9. Finally, the performance of different embedding models is evaluated in Table 10.

4.6.1 Benchmark performance under heterogeneity and imbalance (Q1)

On 20Newsgroup, the proposed space folding method (SFM) consistently yields the highest accuracy across all three non-IID settings ($\alpha \in \{5, 1, 0.1\}$). In the most heterogeneous case ($\alpha = 0.1$), SFM achieves 84.7% test accuracy, improving upon the strongest gradient-based baseline of [36] by up to 23.7 percentage points (Table 1). For CIFAR-10-LT, SFM substantially improves upon the baselines of [71] under long-tailed imbalance and non-IID label distributions (Table 2). At imbalance ratio $\rho = 100$, SFM attains 80.99% test accuracy, outperforming the best competing method (CReFF) by 10.44 percentage points. On the multilingual XGLUE-NC benchmark, SFM again attains the best results (Table 3). For $\alpha = 0.5$, SFM reaches 82.2% accuracy, exceeding the strongest baseline (C2A) by 2.0 percentage points. Finally, on CIFAR-100-LT (Table 4), SFM provides notable gains in the most imbalanced setting. At $\rho = 100$, SFM achieves 46.16% test accuracy, 11.49 percentage points higher than the baseline (CReFF).

Overall, these results indicate that, when coupled with fixed pretrained encoders, the proposed gradient-free FL method is robust to severe label skew and long-tailed imbalance, and can match or exceed the performance of state-of-the-art gradient-based FL methods on the considered benchmarks.

4.6.2 Differentially private federated learning (Q2)

To examine the privacy-utility trade-off (Q2), we perform differentially private FL experiments on 20Newsgroup and XGLUE-NC with non-IID label distributions ($\alpha = 0.1$ for both datasets). For varying values of privacy-loss bound ϵ (with fixed $\delta = 10^{-5}$), Table 5 reports test accuracy under two scenarios:

1. noise-perturbed samples without additional smoothing, i.e., $\mathcal{T}_X^+ = \mathcal{T}_{X+v}$,
2. noise-perturbed samples with the proposed kernel-based smoothing, i.e., $\mathcal{T}_X^+ = \mathcal{T}_{F(X+v)}$.

For high-privacy regimes ($\epsilon \leq 3$), smoothing yields modest but consistent improvements in accuracy on both datasets, indicating that the smoothing mechanism can partially counteract the distortion introduced by noise. The gains remain relatively small, which is consistent with the fact that the underlying KAHM-based autoencoder already enforces a smooth representation. In low-privacy regimes (e.g., $\epsilon = 8$), additional smoothing offers no benefit and may slightly degrade performance, suggesting that smoothing is most useful when stringent privacy guarantees are required.

4.6.3 FHE-secured inference (Q3)

We next examine the suitability of the global prediction rule for secure inference using fully homomorphic encryption. In the proposed framework, inference reduces to computing, for each class, a scalar score based on the aggregated space folding measures and then selecting the class with minimum score. When realized over encrypted integers, this decision rule requires $Q \times C$ homomorphic minimum operations and C equality-comparison, where Q is the number of clients and C is the number of classes. Using the TFHE-rs Rust library with its default high-level parameter set (providing at least 128-bit security under the IND-CPA-D model and a bootstrapping failure probability not exceeding 2^{-128}), we benchmark the primitive operations that dominate the cost of the encrypted prediction rule. On an iMac (M1, 8 GB RAM), minimum and equality-comparison operations on 8-bit and 16-bit encrypted integers are computationally practical, with the corresponding latencies reported in Table 6. Since the total cost of the FHE

realization grows linearly in $Q \times C$ for minimum and C for equality-comparison, these measurements provide an operation-level indication that the induced prediction rule is structurally amenable to FHE-secured inference on standard hardware. These measurements should therefore be interpreted as lower-level building blocks, quantifying the dominant cryptographic operations induced by our decision rule under a particular TFHE implementation, while system-level optimizations (e.g., batching, specialized hardware, or multi-key schemes) are not the focus of our experiments.

4.6.4 Ablation studies (Q4)

Finally, we investigate the sensitivity of the method to the design choices.

Space folding variants Table 7 (CIFAR-100-LT) and Table 8 (XGLUE-NC) compare the four variants of the space folding measure \mathcal{T}_X defined in (102). Across all settings, the performance differences between SFM-1, SFM-2, SFM-3, and SFM-4 are small, indicating that the method is robust to the particular choice of space folding variant.

Batch size Table 9 reports the effect of varying the batch size N_b on CIFAR-100-LT. Smaller batches (e.g., $N_b = 20$) lead to noticeably better performance under strong imbalance, whereas larger batches can hurt accuracy, particularly for the most imbalanced settings. This supports the intuition that, on long-tailed image datasets, modelling smaller batches via separate KAHMs helps capture minority classes more faithfully.

Embedding combinations Table 10 examines the effect of different embedding models on performance. Using either `distiluse-base-multilingual-cased-v2` (mdl_1) or `paraphrase-multilingual` (mdl_2) alone yields strong performance, but their concatenation ($\text{mdl}_1 + \text{mdl}_2$) systematically improves accuracy for all values of α . This suggests that the operator-theoretic gradient-free FL model can effectively exploit complementary information from multiple embedding spaces.

4.7 Inferences Drawn from Experimental Results

We summarize the main empirical findings in terms of the research questions posed at the beginning of this section.

Q1 (Robustness to heterogeneity and imbalance) Across all four datasets, the proposed method achieves accuracy that matches or exceeds strong gradient-based FL baselines under non-IID label distributions and long-tailed class imbalance. The sizeable gains on 20Newsgroup, CIFAR-10-LT, and CIFAR-100-LT, together with the improvements on the multilingual XGLUE-NC benchmark, indicate that the proposed FL method is robust to challenging data heterogeneity when built on top of pretrained encoders.

Q2 (Privacy-utility trade-off) In the differentially private FL experiments, the method maintains competitive accuracy even under tight privacy budgets. The kernel-based smoothing improves performance in high-privacy regimes ($\epsilon \leq 3$), while having limited or no benefit when privacy constraints are relaxed. This suggests that smoothing should be applied primarily when strong privacy guarantees are required.

Q3 (Secure inference efficiency) The global gradient-free FL model admits a simple inference procedure suitable for FHE: for a C -class problem with Q clients, encrypted inference requires $Q \times C$ minimum and C equality-comparison operations per test point. The measured latencies for these encrypted primitive operations show that, under the evaluated cryptographic parameter settings and for the dataset and client scales studied here, the resulting FHE-secured inference appears computationally feasible on standard hardware.

Q4 (Sensitivity to design choices) The ablation studies demonstrate that the method is robust to the choice of space folding variant, benefits from smaller batch sizes in the presence of long-tailed imbalance, and can exploit complementary embeddings to further improve accuracy.

Overall summary Taken together, the experiments suggest that the proposed operator-theoretic framework offers a favourable combination of robustness to heterogeneity, privacy preservation, and practical efficiency for federated learning, within the scope of the benchmarks and settings considered in this study.

5 Conclusion

The primary contribution of this study is the development of an operator-theoretic kernel framework for the design and analysis of gradient-free federated learning algorithms. The framework addresses the key requirements identified in Section 1 by reformulating the FL problem in the L^2 function space, mapping the L^2 -optimal solution into a reproducing kernel Hilbert space (RKHS) via an invertible operator, and deriving finite-sample performance guarantees using concentration inequalities over operator norms. This yields a gradient-free learning scheme together with non-asymptotic bounds on risk, prediction error, robustness, and approximation error. Within this formulation, we determine a data-dependent hypothesis space by tuning the kernel to the scale of the data and analyse its complexity via Rademacher complexity. The analysis shows that scalar space folding summaries derived from Kernel Affine Hull Machines (KAHMs) are sufficient for the global task learning solution, characterizing when high-dimensional gradient exchanges and multiple communication rounds are not required. In this way, the framework offers a mathematically grounded alternative to traditional gradient-based FL in heterogeneous settings.

The framework further integrates privacy-enhancing and security mechanisms into FL. Differentially private FL is achieved by applying a single optimized noise-adding mechanism to each client’s data matrices, followed by kernel-based smoothing and the computation of scalar space folding summaries. By the post-processing property of differential privacy, the resulting global decision rule inherits the (ϵ, δ) -DP guarantee. Secure FL is enabled by fully homomorphic encryption (FHE) of space folding measures. Because the global decision rule for a C -class problem with Q participating clients can be implemented using $Q \times C$ minimum and C equality-comparison operations per test point, the induced FHE-secured inference has a simple and low-dimensional computational structure. Operation-level benchmarks of these encrypted primitives indicate that, for the problem sizes and cryptographic parameter settings studied here, such FHE-secured inference is practically feasible on standard hardware.

Empirically, when combined with embeddings from existing encoders, the resulting gradient-free FL method is competitive with, and in several settings outperforms, strong gradient-based FL methods on non-IID and long-tailed benchmarks. The experiments also indicate that the proposed smoothing mechanism can mitigate the accuracy loss induced by differential privacy in high-privacy regimes, and that the structural simplicity of the FHE-secured decision rule, together with the measured primitive latencies, supports its practical feasibility at the evaluated scales.

A limitation of the present framework is that it treats feature extractors as fixed and focuses exclusively on the design and analysis of the federated prediction head. While this matches the settings where pretrained encoders are frozen for regulatory, engineering, or cost reasons, it does not directly address end-to-end representation learning under federated constraints. Extending the operator-theoretic construction to encompass joint encoder and head learning, for example via operator-theoretic formulations of representation learning objectives or hybrid gradient-free/gradient-based schemes, is an interesting avenue for future work.

Overall, the main value of this work lies in the unifying operator-theoretic perspective and the associated guarantees, which are largely architectural and model-agnostic and do not depend on a particular dataset, feature encoder, or hardware platform. While our experiments focus on standard non-IID partitions, long-tailed settings, and a reference implementation for concreteness, the theoretical results are intended to remain informative as future work broadens tasks, systems, and deployment conditions.

Acknowledgments

The research reported in this paper has been supported by the state of Upper Austria as part of #upperVISION2030 under the Secure Prescriptive Analytics (SPA) project; project ARIKI [(FFG grant no. 897904)] funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), which is linked to a two-stage German funding call “Development of Digital Technologies” of the Federal Ministry for Economic Affairs and Climate Action (BMWK); Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry for Digital and Economic Affairs, and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [(FFG grant no. 892418)] part of the FFG COMET Competence Centers for Excellent Technologies Programme.

Appendix A. Description of the KAHM Expression (19)

With reference to the KAHM expression (19), the following definitions are provided:

- $\mathbf{P}_X \in \mathbb{R}^{n \times n}$ ($n \in \{1, 2, \dots, n\}$) is an encoding matrix such that product $\mathbf{P}_X x$ is a lower-dimensional (i.e. n -dimensional) encoding for x . The encoding matrix is computed from the data samples \mathbf{X} using the following algorithm:

Algorithm 1 Determination of Encoding Matrix \mathbf{P}_X

Require: Matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$, equivalently represented as dataset $\{x^i \in \mathbb{R}^n\}_{i=1}^N$.

- 1: $\underline{n} \leftarrow \min(20, n, N - 1)$.
 - 2: Define $\mathbf{P}_X \in \mathbb{R}^{n \times n}$ such that the i -th row of \mathbf{P}_X is equal to transpose of eigenvector corresponding to i -th largest eigenvalue of sample covariance matrix of samples $\{x^1, \dots, x^N\}$.
 - 3: **while** $\min_{1 \leq j \leq \underline{n}} (\max_{1 \leq i \leq N} (\mathbf{P}_X x^i)_j - \min_{1 \leq i \leq N} (\mathbf{P}_X x^i)_j) < 1e-3$ **do**
 - 4: $\underline{n} \leftarrow \underline{n} - 1$.
 - 5: Define $\mathbf{P}_X \in \mathbb{R}^{n \times n}$ such that the i -th row of \mathbf{P}_X is equal to transpose of eigenvector corresponding to i -th largest eigenvalue of sample covariance matrix of dataset $\{x^1, \dots, x^N\}$.
 - 6: **end while**
 - 7: **return** \mathbf{P}_X .
-

- We have

$$\underline{\mathcal{X}} := \{\mathbf{P}_X x \mid x \in \mathbb{R}^n\}, \quad (146)$$

and a positive-definite real-valued kernel, $k_X : \underline{\mathcal{X}} \times \underline{\mathcal{X}} \rightarrow \mathbb{R}$ on $\underline{\mathcal{X}}$ with a corresponding reproducing kernel Hilbert space $\mathcal{H}_{k_X}(\underline{\mathcal{X}})$, as

$$k_X(\underline{x}^i, \underline{x}^j) := \exp\left(-\frac{1}{2\underline{n}}(\underline{x}^i - \underline{x}^j)^T \theta_X^{-1}(\underline{x}^i - \underline{x}^j)\right), \quad (147)$$

where $\underline{x}^i, \underline{x}^j \in \underline{\mathcal{X}}$ and $\theta_X \succ 0$ is sample covariance matrix of dataset $\{\mathbf{P}_X x^1, \dots, \mathbf{P}_X x^N\}$.

- The function $h_X^i : \underline{\mathcal{X}} \rightarrow \mathbb{R}$, such that $h_X^i \in \mathcal{H}_{k_X}(\underline{\mathcal{X}})$, approximates the indicator function $\mathbb{1}_{\{\mathbf{P}_X x^i\}} : \underline{\mathcal{X}} \rightarrow \{0, 1\}$ as the solution of following kernel regularized least squares problem:

$$h_X^i = \arg \min_{f \in \mathcal{H}_{k_X}(\underline{\mathcal{X}})} \left(\sum_{j=1}^N |\mathbb{1}_{\{\mathbf{P}_X x^i\}}(\mathbf{P}_X x^j) - f(\mathbf{P}_X x^j)|^2 + \lambda_X^* \|f\|_{\mathcal{H}_{k_X}(\underline{\mathcal{X}})}^2 \right), \quad (148)$$

where the regularization parameter $\lambda_X^* \in \mathbb{R}_+$ is given as

$$\lambda_X^* = \hat{e} + \frac{2}{nN} \|\mathbf{X}\|_F^2, \quad (149)$$

where \hat{e} is the unique fixed point of the function r such that

$$\hat{e} = r(\hat{e}, \frac{2}{nN} \|\mathbf{X}\|_F^2), \quad (150)$$

with $r : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined as

$$r(e, \tau) := \frac{1}{nN} \sum_{j=1}^n \|(\mathbf{X})_{:,j} - \mathbf{K}_X (\mathbf{K}_X + (e + \tau) \mathbf{I}_N)^{-1} (\mathbf{X})_{:,j}\|^2, \quad (151)$$

where $(\mathbf{I}_N)_{i,:}$ denotes the i -th row of identity matrix of size N and \mathbf{K}_X is $N \times N$ kernel matrix with its (i, j) -th element defined as

$$(\mathbf{K}_X)_{ij} := k_X(\mathbf{P}_X x^i, \mathbf{P}_X x^j). \quad (152)$$

The following iterations

$$e|_{it+1} = r(e|_{it}, \frac{2}{nN} \|\mathbf{X}\|_F^2), \quad it \in \{0, 1, \dots\} \quad (153)$$

$$e|_0 \in (0, \frac{1}{nN} \|\mathbf{X}\|_F^2) \quad (154)$$

converge to \hat{e} . The solution of the kernel regularized least squares problem follows as

$$h_X^i(\cdot) = (\mathbf{I}_N)_{i,:} (\mathbf{K}_X + \lambda_X^* \mathbf{I}_N)^{-1} [k_X(\cdot, \mathbf{P}_X x^1) \ \dots \ k_X(\cdot, \mathbf{P}_X x^N)]^T \quad (155)$$

The value $h_X^i(\mathbf{P}_X x)$ represents the kernel-smoothed membership of point $\mathbf{P}_X x$ to the set $\{\mathbf{P}_X x^i\}$.

- The image of \mathcal{A}_X defines a region in the affine hull of $\{x^1, \dots, x^N\}$. That is,

$$\mathcal{A}_X[\mathbb{R}^n] := \{\mathcal{A}_X(x) \mid x \in \mathbb{R}^n\} \subset \text{aff}(\{x^1, \dots, x^N\}). \quad (156)$$

Appendix B. Proof of Equation (22)

Consider

$$\operatorname{argmin}_{g \in L^2(\mathbb{R}^n, \mathbb{P}_x)} \left(\int_{\mathbb{R}^n \times \{0,1\}^C} |y_c - g(x)|^2 d\mathbb{P}_{x,y}(x,y) \right) = \operatorname{argmin}_{g \in L^2(\mathbb{R}^n, \mathbb{P}_x)} \left(\mathbb{E}_{x \sim \mathbb{P}_x} \left[\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [|y_c - g(x)|^2] \right] \right) \quad (157)$$

$$= \operatorname{argmin}_{g \in L^2(\mathbb{R}^n, \mathbb{P}_x)} \left(\mathbb{E}_{x \sim \mathbb{P}_x} \left[|g(x)|^2 - 2g(x) \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right] \right) \quad (158)$$

$$= \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x], \quad (159)$$

where we have considered that $\mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \in L^2(\mathbb{R}^n, \mathbb{P}_x)$. Thus, (22) follows.

Appendix C. Proof of \mathcal{K}_{Φ_c} Being a Positive Semi-Definite Kernel

\mathcal{K}_{Φ_c} is a positive semi-definite kernel, since

- $\mathcal{K}_{\Phi_c}(x^1, x^2) = \mathcal{K}_{\Phi_c}(x^2, x^1)$, and
- for every $x^1, \dots, x^N \in \mathbb{R}^n$ and $\alpha_1, \dots, \alpha_N \in \mathbb{R}$,

$$\sum_{i,j=1}^N \alpha_i \alpha_j \mathcal{K}_{\Phi_c}(x^i, x^j) \geq 0. \quad (160)$$

Inequality (160) can be proved by considering that

$$\sum_{i,j=1}^N \alpha_i \alpha_j \mathcal{K}_{\Phi_c}(x^i, x^j) = \sum_{i,j=1}^N \alpha_i \Phi_c(x^i) \alpha_j \Phi_c(x^j) \quad (161)$$

$$= \left| \sum_{i=1}^N \alpha_i \Phi_c(x^i) \right|^2 \quad (162)$$

$$\geq 0. \quad (163)$$

Appendix D. Proof of $J : \mathcal{H}_{\Phi_c} \hookrightarrow L^2(\mathbb{R}^n, \mathbb{P}_x)$ Being Well Defined

Consider for any $f \in \mathcal{H}_{\Phi_c}$,

$$\|Jf\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \int_{\mathbb{R}^n} |f(x)|^2 d\mathbb{P}_x(x) \quad (164)$$

$$= \int_{\mathbb{R}^n} |\langle f, \mathcal{K}_{\Phi_c}(x, \cdot) \rangle_{\mathcal{H}_{\Phi_c}}|^2 d\mathbb{P}_x(x) \quad (165)$$

$$\leq \int_{\mathbb{R}^n} \|f\|_{\mathcal{H}_{\Phi_c}}^2 \|\mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 d\mathbb{P}_x(x) \quad (166)$$

$$= \|f\|_{\mathcal{H}_{\Phi_c}}^2 \int_{\mathbb{R}^n} \|\mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 d\mathbb{P}_x(x). \quad (167)$$

Since

$$\sup_x \mathcal{K}_{\Phi_c}(x, x) \leq 1, \quad (168)$$

we have

$$\|\mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 \leq 1, \quad (169)$$

and thus

$$\|Jf\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 \leq \|f\|_{\mathcal{H}_{\Phi_c}}^2 < \infty. \quad (170)$$

That is $Jf \in L^2(\mathbb{R}^n, \mathbb{P}_x)$. Hence, J is well defined.

Appendix E. Proof of $(J^*)^{-1}$ Being Well Defined on the Range of J^*

Consider for any $f \in \mathcal{H}_{\Phi_c}$,

$$\|(J^*)^{-1}f\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 = \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2\right)^2} \int_{\mathbb{R}^n} |f(x)|^2 d\mathbb{P}_x(x) \quad (171)$$

$$= \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2\right)^2} \int_{\mathbb{R}^n} |\langle f, \mathcal{K}_{\Phi_c}(x, \cdot) \rangle_{\mathcal{H}_{\Phi_c}}|^2 d\mathbb{P}_x(x) \quad (172)$$

$$\leq \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2\right)^2} \int_{\mathbb{R}^n} \|f\|_{\mathcal{H}_{\Phi_c}}^2 \|\mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 d\mathbb{P}_x(x) \quad (173)$$

$$= \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2\right)^2} \|f\|_{\mathcal{H}_{\Phi_c}}^2 \int_{\mathbb{R}^n} \|\mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 d\mathbb{P}_x(x) \quad (174)$$

$$\leq \frac{1}{\left(\|\Phi_c\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2\right)^2} \|f\|_{\mathcal{H}_{\Phi_c}}^2, \quad (175)$$

where we have used (169). Due to (30), we have

$$\|(J^*)^{-1}f\|_{L^2(\mathbb{R}^n, \mathbb{P}_x)}^2 < \infty. \quad (176)$$

That is, $(J^*)^{-1}f \in L^2(\mathbb{R}^n, \mathbb{P}_x)$. Hence, $(J^*)^{-1}$ is well defined on the range of J^* .

Appendix F. Proof of Equation (46)

Consider

$$(J^* J)f = \mathbb{E}_{x' \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x', \cdot)(Jf)(x')] \quad (177)$$

$$= \mathbb{E}_{x' \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x', \cdot)f(x')] \quad (178)$$

$$= \mathbb{E}_{x' \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x', \cdot)\langle f, \mathcal{K}_{\Phi_c}(x', \cdot) \rangle_{\mathcal{H}_{\Phi_c}}] \quad (179)$$

$$= \mathbb{E}_{x' \sim \mathbb{P}_x} [(\mathcal{K}_{\Phi_c}(x', \cdot) \otimes \mathcal{K}_{\Phi_c}(x', \cdot))(f)]. \quad (180)$$

Thus, (46) follows.

Appendix G. Proof of Inequality (48)

Consider for any $f \in \mathcal{H}_{\Phi_c}$,

$$\|(J^* J)f\|_{\mathcal{H}_{\Phi_c}} = \left\| \mathbb{E}_{x' \sim \mathbb{P}_x} [(\mathcal{K}_{\Phi_c}(x', \cdot) \otimes \mathcal{K}_{\Phi_c}(x', \cdot))(f)] \right\|_{\mathcal{H}_{\Phi_c}} \quad (181)$$

$$\leq \mathbb{E}_{x' \sim \mathbb{P}_x} \left[\|(\mathcal{K}_{\Phi_c}(x', \cdot) \otimes \mathcal{K}_{\Phi_c}(x', \cdot))(f)\|_{\mathcal{H}_{\Phi_c}} \right] \quad (182)$$

$$= \mathbb{E}_{x' \sim \mathbb{P}_x} [\langle f, \mathcal{K}_{\Phi_c}(x', \cdot) \rangle_{\mathcal{H}_{\Phi_c}} \|\mathcal{K}_{\Phi_c}(x', \cdot)\|_{\mathcal{H}_{\Phi_c}}] \quad (183)$$

$$\leq \mathbb{E}_{x' \sim \mathbb{P}_x} \left[\|f\|_{\mathcal{H}_{\Phi_c}} \|\mathcal{K}_{\Phi_c}(x', \cdot)\|_{\mathcal{H}_{\Phi_c}}^2 \right] \quad (184)$$

$$\leq \|f\|_{\mathcal{H}_{\Phi_c}}. \quad (185)$$

Thus, (48) follows.

Appendix H. Proof of Inequality (65)

Consider

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[\left| h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) \right|^2 \right] = \| J(h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}) \|_{L^2(\mathbb{R}^n, \mathbb{P}_x)} \quad (186)$$

$$= \left(\left\langle J(h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}), J(h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}) \right\rangle_{L^2(\mathbb{R}^n, \mathbb{P}_x)} \right)^{1/2} \quad (187)$$

$$= \left(\left\langle (h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}), J^* J(h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}) \right\rangle_{\mathcal{H}_{\Phi_c}} \right)^{1/2} \quad (188)$$

$$= \left\| (J^* J)^{1/2} (h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}} - J^* f_{x \mapsto y_c}) \right\|_{\mathcal{H}_{\Phi_c}}. \quad (189)$$

Using (63), we have

$$\begin{aligned} & \mathbb{E}_{x \sim \mathbb{P}_x} \left[\left| h_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) - f_{x \mapsto y_c}^{\mathcal{H}_{\Phi_c}}(x) \right|^2 \right] \\ &= \left\| (J^* J)^{1/2} \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} + (J^* J)^{1/2} \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \\ &\leq \|(J^* J)^{1/2}\|_{\text{op}} \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \|(J^* J)^{1/2}\|_{\text{op}} \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}}. \end{aligned} \quad (190)$$

Using (52), we get (65).

Appendix I. Proof of Inequality (66)

Define, for a given $x \in \mathbb{R}^n$, $F_{x,c} \in \mathcal{H}_{\Phi_c}$ as

$$F_{x,c} := \mathcal{K}_{\Phi_c}(x, \cdot) f_{x \mapsto y_c}(x) - J^* f_{x \mapsto y_c}. \quad (191)$$

Considering x as a random variable, $F_{x,c}$ is a random variable taking values in \mathcal{H}_{Φ_c} with mean equal to the zero function, i.e.,

$$\mathbb{E}_{x \sim \mathbb{P}_x} [F_{x,c}] = \mathbf{0}, \quad (192)$$

where $\mathbf{0} : \mathbb{R}^n \rightarrow 0$. Consider

$$\|F_{x,c}\|_{\mathcal{H}_{\Phi_c}} = \left\| f_{x \mapsto y_c}(x) \mathcal{K}_{\Phi_c}(x, \cdot) - \mathbb{E}_{x \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x, \cdot) f_{x \mapsto y_c}(x)] \right\|_{\mathcal{H}_{\Phi_c}} \quad (193)$$

$$\leq \|f_{x \mapsto y_c}(x) \mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}} + \left\| \mathbb{E}_{x \sim \mathbb{P}_x} [\mathcal{K}_{\Phi_c}(x, \cdot) f_{x \mapsto y_c}(x)] \right\|_{\mathcal{H}_{\Phi_c}} \quad (194)$$

$$\leq \|f_{x \mapsto y_c}(x) \mathcal{K}_{\Phi_c}(x, \cdot)\|_{\mathcal{H}_{\Phi_c}} + \mathbb{E}_{x \sim \mathbb{P}_x} \left[\|\mathcal{K}_{\Phi_c}(x, \cdot) f_{x \mapsto y_c}(x)\|_{\mathcal{H}_{\Phi_c}} \right] \quad (195)$$

$$\leq 2, \quad (196)$$

where (196) follows from (195) using (24) and (169). Now, Consider

$$\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} \right] \leq \left(\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}}^2 \right] \right)^{1/2} \quad (197)$$

$$= \left(\frac{1}{N^2} \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \sum_{i=1}^N F_{x^i, c} \right\|_{\mathcal{H}_{\Phi_c}}^2 \right] \right)^{1/2} \quad (198)$$

$$= \left(\frac{1}{N^2} \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\sum_{i=1}^N \sum_{j=1}^N \langle F_{x^i, c}, F_{x^j, c} \rangle_{\mathcal{H}_{\Phi_c}} \right] \right)^{1/2} \quad (199)$$

$$= \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}_{(x^i, y^i) \sim \mathbb{P}_{x,y}, (x^j, y^j) \sim \mathbb{P}_{x,y}} [\langle F_{x^i, c}, F_{x^j, c} \rangle_{\mathcal{H}_{\Phi_c}}] \right)^{1/2}. \quad (200)$$

Using the independence of the samples $((x^i, y^i))_{i=1}^N$ and (192), we have

$$\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y}, (x^j, y^j) \sim \mathbb{P}_{x,y})} \left[\langle F_{x^i, c}, F_{x^j, c} \rangle_{\mathcal{H}_{\Phi_c}} \right] = 0, \text{ if } i \neq j. \quad (201)$$

Therefore,

$$\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)}_{i=1}^N - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} \right] \leq \left(\frac{1}{N^2} \sum_{i=1}^N \mathbb{E}_{(x^i, y^i) \sim \mathbb{P}_{x,y}} \left[\|F_{x^i, c}\|_{\mathcal{H}_{\Phi_c}}^2 \right] \right)^{1/2} \quad (202)$$

$$\leq \frac{2}{\sqrt{N}}. \quad (203)$$

where we have used (196).

Appendix J. Proof of Inequality (67)

Consider

$$\begin{aligned} & \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \widehat{S}_{(x^i)}_{i=1}^N (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right] \\ & \leq \left(\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \widehat{S}_{(x^i)}_{i=1}^N (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}}^2 \right] \right)^{1/2} \end{aligned} \quad (204)$$

$$= \left(\mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \xi_c(x^i, y^i) \xi_c(x^j, y^j) \mathcal{K}_{\Phi_c}(x^i, x^j) \right] \right)^{1/2} \quad (205)$$

$$= \left(\frac{1}{N^2} \sum_{i=1}^N \mathbb{E}_{(x^i, y^i) \sim \mathbb{P}_{x,y}} [|\xi_c(x^i, y^i)|^2 \mathcal{K}_{\Phi_c}(x^i, x^i)] \right)^{1/2}, \quad (206)$$

where we have used the independence of the samples $((x^i, y^i))_{i=1}^N$ and (27). Using (168) and (28), we get (67).

Appendix K: Proof of Inequality (68)

Define a function $\psi_c : (\mathbb{R}^n \times \{0, 1\}^C)^N \rightarrow \mathbb{R}_{\geq 0}$ as

$$\psi_c := \psi_{1,c}((x^1, y^1), \dots, (x^N, y^N)) + \psi_{2,c}((x^1, y^1), \dots, (x^N, y^N)), \quad (207)$$

where $\psi_{1,c} : (\mathbb{R}^n \times \{0, 1\}^C)^N \rightarrow \mathbb{R}_{\geq 0}$ and $\psi_{2,c} : (\mathbb{R}^n \times \{0, 1\}^C)^N \rightarrow \mathbb{R}_{\geq 0}$ are defined as

$$\psi_{1,c}((x^1, y^1), \dots, (x^N, y^N)) := \left\| \left(\widehat{S}_{(x^i)}_{i=1}^N - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} \quad (208)$$

$$\psi_{2,c}((x^1, y^1), \dots, (x^N, y^N)) := \left\| \widehat{S}_{(x^i)}_{i=1}^N (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}}. \quad (209)$$

It can be seen that

$$\psi_{1,c}((x^1, y^1), \dots, (x^N, y^N)) := \frac{1}{N} \left\| \sum_{i=1}^N F_{x^i, c} \right\|_{\mathcal{H}_{\Phi_c}}, \quad (210)$$

where $F_{x,c}$ is defined as in (191). Consider

$$\left| \frac{\|F_{x^1, c} + \dots + F_{x^i, c} + \dots + F_{x^N, c}\|_{\mathcal{H}_{\Phi_c}}}{N} - \frac{\|F_{x^1, c} + \dots + F_{x'^i, c} + \dots + F_{x^N, c}\|_{\mathcal{H}_{\Phi_c}}}{N} \right|$$

$$\leq \frac{1}{N} \|F_{x^i, c} - F_{x'^i, c}\|_{\mathcal{H}_{\Phi_c}} \quad (211)$$

$$= \frac{1}{N} \|f_{x \mapsto y_c}(x^i) \mathcal{K}_{\Phi_c}(x^i, \cdot) - f_{x \mapsto y_c}(x'^i) \mathcal{K}_{\Phi_c}(x'^i, \cdot)\|_{\mathcal{H}_{\Phi_c}} \quad (212)$$

$$\leq \frac{1}{N} (|f_{x \mapsto y_c}(x^i)| \|\mathcal{K}_{\Phi_c}(x^i, \cdot)\|_{\mathcal{H}_{\Phi_c}} + |f_{x \mapsto y_c}(x'^i)| \|\mathcal{K}_{\Phi_c}(x'^i, \cdot)\|_{\mathcal{H}_{\Phi_c}}) \quad (213)$$

$$\leq \frac{2}{N}. \quad (214)$$

Thus,

$$\frac{2}{N} \geq \sup_{(x'^i, y'^i) \in \mathbb{R}^n \times \{0,1\}^C} (|\psi_{1,c}((x^1, y^1), \dots, (x^i, y^i), \dots, (x^N, y^N)) \\ - \psi_{1,c}((x^1, y^1), \dots, (x'^i, y'^i), \dots, (x^N, y^N))|). \quad (215)$$

Now consider

$$\begin{aligned} & \left| \left\| \widehat{S}_{(x^1, \dots, x^i, \dots, x^N)}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^i, y^i), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right. \\ & \quad \left. - \left\| \widehat{S}_{(x^1, \dots, x'^i, \dots, x^N)}^* (\xi_c(x^1, y^1), \dots, \xi_c(x'^i, y'^i), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right| \\ & \leq \left\| \widehat{S}_{(x^1, \dots, x^i, \dots, x^N)}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^i, y^i), \dots, \xi_c(x^N, y^N)) \right. \\ & \quad \left. - \widehat{S}_{(x^1, \dots, x'^i, \dots, x^N)}^* (\xi_c(x^1, y^1), \dots, \xi_c(x'^i, y'^i), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \end{aligned} \quad (216)$$

$$= \frac{1}{N} \left\| \xi_c(x^i, y^i) \mathcal{K}_{\Phi_c}(x^i, \cdot) - \xi_c(x'^i, y'^i) \mathcal{K}_{\Phi_c}(x'^i, \cdot) \right\|_{\mathcal{H}_{\Phi_c}} \quad (217)$$

$$\leq \frac{1}{N} |\xi_c(x^i, y^i)| \|\mathcal{K}_{\Phi_c}(x^i, \cdot)\|_{\mathcal{H}_{\Phi_c}} + \frac{1}{N} |\xi_c(x'^i, y'^i)| \|\mathcal{K}_{\Phi_c}(x'^i, \cdot)\|_{\mathcal{H}_{\Phi_c}} \quad (218)$$

$$\leq \frac{2}{N}, \quad (219)$$

where we have used (169) and (28). Thus,

$$\frac{2}{N} \geq \sup_{(x'^i, y'^i) \in \mathbb{R}^n \times \{0,1\}^C} (|\psi_{2,c}((x^1, y^1), \dots, (x^i, y^i), \dots, (x^N, y^N)) \\ - \psi_{2,c}((x^1, y^1), \dots, (x'^i, y'^i), \dots, (x^N, y^N))|). \quad (220)$$

It follows from (215) and (220) that

$$\frac{4}{N} \geq \sup_{(x'^i, y'^i) \in \mathbb{R}^n \times \{0,1\}^C} (|\psi_c((x^1, y^1), \dots, (x^i, y^i), \dots, (x^N, y^N)) \\ - \psi_c((x^1, y^1), \dots, (x'^i, y'^i), \dots, (x^N, y^N))|). \quad (221)$$

Thus, ψ_c satisfies the bounded differences property with bound $4/N$, and therefore by McDiarmid's inequality, for any $\epsilon > 0$, with probability at most $\exp(-0.125N\epsilon^2)$, the following holds:

$$\epsilon \leq \psi_c((x^1, y^1), \dots, (x^i, y^i), \dots, (x^N, y^N)) - \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} [\psi_c((x^1, y^1), \dots, (x^i, y^i), \dots, (x^N, y^N))]. \quad (222)$$

That is, with probability at most $\exp(-0.125N\epsilon^2)$, the following holds:

$$\begin{aligned} \epsilon & \leq \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \\ & \quad - \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right] \end{aligned} \quad (223)$$

That is, with probability at most $\delta > 0$, the following holds:

$$\begin{aligned} & \sqrt{\frac{8 \log(1/\delta)}{N}} \\ & \leq \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \\ & \quad - \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right] \end{aligned} \quad (224)$$

In other words, with probability at least $1 - \delta$, the following holds:

$$\begin{aligned} & \sqrt{\frac{8 \log(1/\delta)}{N}} \\ & \geq \left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \\ & \quad - \mathbb{E}_{((x^i, y^i) \sim \mathbb{P}_{x,y})_{i=1}^N} \left[\left\| \left(\widehat{S}_{(x^i)_{i=1}^N}^* \text{Ev}_{(x^i)_{i=1}^N} - J^* \right) f_{x \mapsto y_c} \right\|_{\mathcal{H}_{\Phi_c}} + \left\| \widehat{S}_{(x^i)_{i=1}^N}^* (\xi_c(x^1, y^1), \dots, \xi_c(x^N, y^N)) \right\|_{\mathcal{H}_{\Phi_c}} \right] \end{aligned} \quad (225)$$

Using (66) and (67) in (225), we get (68).

Appendix L: Proof of Theorem 5

Consider

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) = \frac{1}{N} \mathbb{E} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \sum_{i=1}^N \sigma_i h_{x \mapsto y_c}(x^i) \right] \quad (226)$$

$$= \frac{1}{N} \mathbb{E} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \sum_{i=1}^N \sigma_i \langle h_{x \mapsto y_c}, \mathcal{K}_{\Phi_c}(\cdot, x^i) \rangle_{\mathcal{H}_{\Phi_c}} \right], \quad (227)$$

where we have used the reproducing property of the kernel, since $h_{x \mapsto y_c} \in \mathcal{H}_{\Phi_c}$. That is,

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) = \frac{1}{N} \mathbb{E} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left\langle h_{x \mapsto y_c}, \sum_{i=1}^N \sigma_i \mathcal{K}_{\Phi_c}(\cdot, x^i) \right\rangle_{\mathcal{H}_{\Phi_c}} \right] \quad (228)$$

$$\leq \frac{1}{N} \mathbb{E} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\|h_{x \mapsto y_c}\|_{\mathcal{H}_{\Phi_c}} \left\| \sum_{i=1}^N \sigma_i \mathcal{K}_{\Phi_c}(\cdot, x^i) \right\|_{\mathcal{H}_{\Phi_c}} \right) \right], \quad (229)$$

where (229) follows from (228) due to Cauchy–Schwarz inequality. Also,

$$\left\| \sum_{i=1}^N \sigma_i \mathcal{K}_{\Phi_c}(\cdot, x^i) \right\|_{\mathcal{H}_{\Phi_c}} = \left| \sum_{i=1}^N \sigma_i \Phi_c(x^i) \right|. \quad (230)$$

Using (91) and (230),

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) \leq \frac{1}{N} \mathbb{E} \left[\sup_{\Phi_c: \mathbb{R}^n \rightarrow [0,1]} \left(\frac{\sum_{i=1}^{N_c} \Phi_c(x^{I_i})}{N_c} \left| \sum_{i=1}^N \sigma_i \Phi_c(x^i) \right| \right) \right] \quad (231)$$

$$\leq \frac{1}{N} \mathbb{E} \left[\sup_{\Phi_c: \mathbb{R}^n \rightarrow [0,1]} \left(\left| \sum_{i=1}^N \sigma_i \Phi_c(x^i) \right| \right) \right]. \quad (232)$$

For any $\epsilon > 0$, let $\overline{\Phi}_{c,\sigma}: \mathbb{R}^n \rightarrow [0,1]$ be such that

$$\sup_{\Phi_c: \mathbb{R}^n \rightarrow [0,1]} \left(\left| \sum_{i=1}^N \sigma_i \Phi_c(x^i) \right| \right) = \left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right| + \epsilon. \quad (233)$$

Using (233) in (232),

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) \leq \frac{1}{N} \mathbb{E} \left[\left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right| \right] + \frac{\epsilon}{N}. \quad (234)$$

As per Jensen's inequality,

$$\left(\mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right| \right] \right)^2 \leq \mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right|^2 \right], \quad (235)$$

and thus

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) \leq \frac{1}{N} \sqrt{\mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right|^2 \right]} + \frac{\epsilon}{N}. \quad (236)$$

Consider

$$\mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \overline{\Phi}_{c,\sigma}(x^i) \right|^2 \right] = \mathbb{E}_{\sigma} \left[\sum_{i,j=1}^N \sigma_i \sigma_j \overline{\Phi}_{c,\sigma}(x^i) \overline{\Phi}_{c,\sigma}(x^j) \right] \quad (237)$$

$$= \sum_{i,j=1}^N \mathbb{E}_{\sigma} [\sigma_i \sigma_j] \overline{\Phi}_{c,\sigma}(x^i) \overline{\Phi}_{c,\sigma}(x^j). \quad (238)$$

Since $\sigma_1, \dots, \sigma_N$ are independent random variables drawn from the Rademacher distribution, we have

$$\mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \bar{\Phi}_{c,\sigma}(x^i) \right|^2 \right] = \sum_{i=1}^N |\bar{\Phi}_{c,\sigma}(x^i)|^2. \quad (239)$$

As $\bar{\Phi}_{c,\sigma} : \mathbb{R}^n \rightarrow [0, 1]$, we have

$$\mathbb{E}_{\sigma} \left[\left| \sum_{i=1}^N \sigma_i \bar{\Phi}_{c,\sigma}(x^i) \right|^2 \right] \leq N. \quad (240)$$

Using (240) in (236), we get

$$\widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c) \leq \frac{1}{\sqrt{N}} + \frac{\epsilon}{N}. \quad (241)$$

Since the inequality (241) holds for all $\epsilon > 0$, we have (93).

Appendix M: Proof of Theorem 6

Define, for a given dataset \mathcal{D} (as defined in (6)),

$$\widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) = \frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2. \quad (242)$$

Consider a function assessing the supremum of difference of expected loss value from empirically averaged loss value:

$$g_c(\mathcal{D}) := \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x)|^2] - \widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) \right). \quad (243)$$

Let $\mathcal{D}' = \{(x^1, y^1), \dots, (x^{i-1}, y^{i-1}), (x'^i, y'^i), (x^{i+1}, y^{i+1}), \dots, (x^N, y^N)\}$ be the *neighboring* set of \mathcal{D} such that \mathcal{D}' and \mathcal{D} differ by only a single entry, i.e. the entry $(x'^i, y'^i) \notin \mathcal{D}$. As the difference of suprema can't exceed the supremum of the difference, we have

$$g_c(\mathcal{D}') - g_c(\mathcal{D}) \leq \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{|y_c^i - h_{x \mapsto y_c}(x^i)|^2 - |y'^i - h_{x \mapsto y_c}(x'^i)|^2}{N} \right) \quad (244)$$

$$\leq \frac{1}{N}, \quad (245)$$

where we have used the facts that $y_c^i, y'^i \in \{0, 1\}$ and (86). Similarly, we can obtain

$$g_c(\mathcal{D}) - g_c(\mathcal{D}') \leq \frac{1}{N}. \quad (246)$$

Thus

$$|g_c(\mathcal{D}) - g_c(\mathcal{D}')| \leq \frac{1}{N}. \quad (247)$$

Thus, g_c satisfies the bounded differences property with bound $1/N$, and therefore by Mc-Diarmid's inequality, for any $\epsilon > 0$, with probability at most $\exp(-2N\epsilon^2)$, the following holds:

$$g_c(\mathcal{D}) - \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \geq \epsilon. \quad (248)$$

That is, with probability at most $\delta \in (0, 1)$, the following holds:

$$g_c(\mathcal{D}) - \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \geq \sqrt{\frac{\log(1/\delta)}{2N}}. \quad (249)$$

In other words, with probability at least $1 - \delta$, the following holds:

$$g_c(\mathcal{D}) \leq \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] + \sqrt{\frac{\log(1/\delta)}{2N}}. \quad (250)$$

Let

$$\tilde{\mathcal{D}} = \{(\tilde{x}^i, \tilde{y}^i) \mid i \in \{1, 2, \dots, N\}\} \sim (\mathbb{P}_{x,y})^N \quad (251)$$

be another set of IID samples and consider

$$\begin{aligned} & \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \\ &= \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x)|^2] - \widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) \right) \right] \end{aligned} \quad (252)$$

$$= \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\widehat{\mathbb{E}}_{\tilde{\mathcal{D}}} (h_{x \mapsto y_c}) - \widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) \right) \right] \quad (253)$$

$$= \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\widehat{\mathbb{E}}_{\tilde{\mathcal{D}}} (h_{x \mapsto y_c}) - \widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) \right) \right] \quad (254)$$

$$\leq \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\widehat{\mathbb{E}}_{\tilde{\mathcal{D}}} (h_{x \mapsto y_c}) - \widehat{\mathbb{E}}_{\mathcal{D}}(h_{x \mapsto y_c}) \right) \right] \quad (255)$$

$$= \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right]. \quad (256)$$

Consider

$$\begin{aligned} & \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N \sigma_i (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right] \\ &= \frac{1}{2} \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right] \\ &+ \frac{1}{2} \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|y_c^i - h_{x \mapsto y_c}(x^i)|^2 - |\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2) \right) \right] \end{aligned} \quad (257)$$

where we have used the fact that $\sigma_1, \dots, \sigma_N$ are Rademacher variables (i.e. taking value in $\{-1, 1\}$ with probability equal to 1/2). Further, due to the fact that

$$\mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|y_c^i - h_{x \mapsto y_c}(x^i)|^2 - |\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2) \right) \right] \quad (258)$$

$$= \mathbb{E}_{\tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N, \mathcal{D} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right], \quad (259)$$

(257) becomes

$$\begin{aligned} & \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N \sigma_i (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right] \\ &= \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right]. \end{aligned} \quad (260)$$

Using (260) in (256), we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \\ &\leq \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N \sigma_i (|\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 - |y_c^i - h_{x \mapsto y_c}(x^i)|^2) \right) \right] \end{aligned} \quad (261)$$

$$\leq \mathbb{E}_{\tilde{\mathcal{D}} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N \sigma_i |\tilde{y}_c^i - h_{x \mapsto y_c}(\tilde{x}^i)|^2 \right) \right] \quad (262)$$

$$+ \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N -\sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right]. \quad (263)$$

Since σ_i and $-\sigma_i$ are identically distributed, we have

$$\mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \leq 2 \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N, \sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\frac{1}{N} \sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \quad (264)$$

$$= 2 \mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} \left[\frac{1}{N} \mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \right]. \quad (265)$$

Define

$$u_j(h_{x \mapsto y_c}) := \sum_{i=1}^j \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \quad (266)$$

to express

$$\mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \quad (267)$$

$$= \mathbb{E}_{\sigma_1, \dots, \sigma_{N-1}} \left[\mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N |y_c^N - h_{x \mapsto y_c}(x^N)|^2) \right] \right] \quad (268)$$

For any $\epsilon > 0$, let $h^1, h^2 \in \mathcal{M}_c$ be such that

$$\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + |y_c^N - h_{x \mapsto y_c}(x^N)|^2) = u_{N-1}(h^1) + |y_c^N - h^1(x^N)|^2 + \epsilon \quad (269)$$

$$\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) - |y_c^N - h_{x \mapsto y_c}(x^N)|^2) = u_{N-1}(h^2) - |y_c^N - h^2(x^N)|^2 + \epsilon. \quad (270)$$

Now, consider

$$\mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N |y_c^N - h_{x \mapsto y_c}(x^N)|^2) \right] \quad (271)$$

$$= \frac{1}{2} \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + |y_c^N - h_{x \mapsto y_c}(x^N)|^2) + \frac{1}{2} \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) - |y_c^N - h_{x \mapsto y_c}(x^N)|^2) \quad (272)$$

$$= \frac{1}{2} (u_{N-1}(h^1) + u_{N-1}(h^2)) + \frac{1}{2} ((h^1(x^N) + h^2(x^N) - 2y_c^N)(h^1(x^N) - h^2(x^N))) + \epsilon. \quad (273)$$

Define

$$\eta = \text{sign}(h^1(x^N) - h^2(x^N)), \quad (274)$$

and consider (86) and $y_c^N \in \{0, 1\}$, leading to

$$h^1(x^N) + h^2(x^N) - 2y_c^N \leq 2, \quad (275)$$

so that

$$\mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N |y_c^N - h_{x \mapsto y_c}(x^N)|^2) \right] \leq \frac{1}{2} (u_{N-1}(h^1) + u_{N-1}(h^2)) + \frac{1}{2} (2\eta(h^1(x^N) - h^2(x^N))) + \epsilon \quad (276)$$

$$= \frac{1}{2} (u_{N-1}(h^1) + 2\eta h^1(x^N)) + \frac{1}{2} (u_{N-1}(h^2) - 2\eta h^2(x^N)) + \epsilon \quad (277)$$

$$\leq \frac{1}{2} \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + 2\eta h_{x \mapsto y_c}(x^N)) + \frac{1}{2} \sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) - 2\eta h_{x \mapsto y_c}(x^N)) + \epsilon \quad (278)$$

$$= \mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right] + \epsilon. \quad (279)$$

Since the inequality holds for all $\epsilon > 0$, we have

$$\begin{aligned} & \mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N |y_c^N - h_{x \mapsto y_c}(x^N)|^2) \right] \\ & \leq \mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right]. \end{aligned} \quad (280)$$

Equivalently,

$$\mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \leq \mathbb{E}_{\sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-1}(h_{x \mapsto y_c}) + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right]. \quad (281)$$

That is,

$$\begin{aligned} & \mathbb{E}_{\sigma_{N-1}, \sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \\ & \leq \mathbb{E}_{\sigma_{N-1}} \left[\mathbb{E}_{\sigma_{N-1}} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-2}(h_{x \mapsto y_c}) + \sigma_{N-1} |y_c^{N-1} - h_{x \mapsto y_c}(x^{N-1})|^2 + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right] \right] \end{aligned} \quad (282)$$

Following the same procedure for σ_{N-1} , as that for σ_N to derive the inequality (280), we will get

$$\begin{aligned} & \mathbb{E}_{\sigma_{N-1}} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-2}(h_{x \mapsto y_c}) + \sigma_{N-1} |y_c^{N-1} - h_{x \mapsto y_c}(x^{N-1})|^2 + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right] \\ & \leq \mathbb{E}_{\sigma_{N-1}} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-2}(h_{x \mapsto y_c}) + \sigma_{N-1} 2h_{x \mapsto y_c}(x^{N-1}) + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right], \end{aligned} \quad (283)$$

resulting in

$$\begin{aligned} & \mathbb{E}_{\sigma_{N-1}, \sigma_N} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \\ & \leq \mathbb{E}_{\sigma_N} \left[\mathbb{E}_{\sigma_{N-1}} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} (u_{N-2}(h_{x \mapsto y_c}) + \sigma_{N-1} 2h_{x \mapsto y_c}(x^{N-1}) + \sigma_N 2h_{x \mapsto y_c}(x^N)) \right] \right] \end{aligned} \quad (284)$$

Following the same procedure for $\sigma_{N-2}, \dots, \sigma_1$, we will get

$$\mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \leq 2 \mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i h_{x \mapsto y_c}(x^i) \right) \right] \quad (285)$$

$$= 2N \widehat{\mathcal{R}}_{\mathcal{D}}(\mathcal{M}_c). \quad (286)$$

Using (93), we get

$$\mathbb{E}_{\sigma} \left[\sup_{h_{x \mapsto y_c} \in \mathcal{M}_c} \left(\sum_{i=1}^N \sigma_i |y_c^i - h_{x \mapsto y_c}(x^i)|^2 \right) \right] \leq 2\sqrt{N}. \quad (287)$$

Combining (287) and (265), we have

$$\mathbb{E}_{\mathcal{D} \sim (\mathbb{P}_{x,y})^N} [g_c(\mathcal{D})] \leq \frac{4}{\sqrt{N}}. \quad (288)$$

Using (288) in (250), we have with probability at least $1 - \delta$,

$$g_c(\mathcal{D}) \leq \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}. \quad (289)$$

Considering the definition of g_c (as given in (243)) and (242), we have with probability at least $1 - \delta$,

$$\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x)|^2] \leq \frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}. \quad (290)$$

Using (85) together with Theorem 3, we have with probability at least $1 - \delta$:

$$\mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} [|y_c - h_{x \mapsto y_c}(x)|^2] \leq \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[\left| y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right|^2 \right] + \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right) \quad (291)$$

Combining (290) and (291) leads to the result.

Appendix N: Proof of Theorem 7

Consider

$$\begin{aligned} & \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right] - \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[\left| y_c - \mathbb{E}_{y \sim \mathbb{P}_{y|x}} [y_c|x] \right|^2 \right] \\ &= \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right] - \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] \end{aligned} \quad (292)$$

$$= \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|h_{x \mapsto y_c}(x)|^2 - |\mathbb{P}_{y|x}(y_c = 1|x)|^2 - 2y_c(h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)) \right] \quad (293)$$

$$= \mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x)|^2 \right] - \mathbb{E}_{x \sim \mathbb{P}_x} \left[|\mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] - 2 \mathbb{E}_{x \sim \mathbb{P}_x} \left[\mathbb{P}_{y|x}(y_c = 1|x)(h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)) \right] \quad (294)$$

$$= \mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right]. \quad (295)$$

It follows from (295) that

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] \leq \mathbb{E}_{(x,y) \sim \mathbb{P}_{x,y}} \left[|y_c - h_{x \mapsto y_c}(x)|^2 \right]. \quad (296)$$

Since (290) holds with probability at least $1 - \delta$, using (296), we have with probability at least $1 - \delta$:

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] \leq \frac{1}{N} \sum_{i=1}^N |y_c^i - h_{x \mapsto y_c}(x^i)|^2 + \frac{4}{\sqrt{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}. \quad (297)$$

Using (85) together with Theorem 4, we have with probability at least $1 - \delta$:

$$\mathbb{E}_{x \sim \mathbb{P}_x} \left[|h_{x \mapsto y_c}(x) - \mathbb{P}_{y|x}(y_c = 1|x)|^2 \right] \leq \frac{1}{(N_c/N)^2} \left(\frac{3}{\sqrt{N}} + \sqrt{\frac{8 \log(1/\delta)}{N}} \right). \quad (298)$$

Combining (297) and (298) leads to the result.

References

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) (CCS ’16). Association for Computing Machinery, New York, NY, USA, 308–318. <https://doi.org/10.1145/2976749.2978318>
- [2] Idan Aachitube, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. 2021. Personalized federated learning with Gaussian processes. In *Proceedings of the 35th International Conference on Neural Information Processing Systems* (NIPS ’21). Curran Associates Inc., Red Hook, NY, USA, Article 642, 15 pages.
- [3] Mixedbread AI. 2024. *mxbai-embed-large-v1: English sentence embedding model*. Retrieved September 16, 2025 from <https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1> Apache 2.0 License.
- [4] Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. 2018. Model-Agnostic Private Learning. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/aa97d584861474f4097cf13ccb5325da-Paper.pdf
- [5] Sameer Bibikar, Haris Vikalo, Zhangyang Wang, and Xiaohan Chen. 2022. Federated Dynamic Sparse Training: Computing Less, Communicating Less, Yet Learning Better. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 6 (Jun. 2022), 6080–6088. <https://doi.org/10.1609/aaai.v36i6.20555>
- [6] Kallista Bonawitz, Peter Kairouz, Brendan McMahan, and Daniel Ramage. 2021. Federated Learning and Privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. *Queue* 19, 5 (Nov. 2021), 87–114. <https://doi.org/10.1145/3494834.3500240>
- [7] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. 2023. Automatic Clipping: Differentially Private Deep Learning Made Easier and Stronger. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=e8i70aPj0q>

- [8] Kuntai Cai, Xiaoyu Lei, Jianxin Wei, and Xiaokui Xiao. 2021. Data synthesis via differentially private markov random fields. *Proc. VLDB Endow.* 14, 11 (July 2021), 2190–2202. <https://doi.org/10.14778/3476249.3476272>
- [9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf
- [10] Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. 2020. Understanding gradient clipping in private SGD: a geometric perspective. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (*NIPS* '20). Curran Associates Inc., Red Hook, NY, USA, Article 1155, 10 pages.
- [11] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yong Soo Song. 2017. Homomorphic Encryption for Arithmetic of Approximate Numbers. In *Advances in Cryptology – ASIACRYPT 2017 (Lecture Notes in Computer Science, Vol. 10624)*. Springer, 409–437. https://doi.org/10.1007/978-3-319-70694-8_15
- [12] Ilaria Chillotti, Nicolas Gama, Mariya Georgieva, and Malika Izabachène. 2020. TFHE: Fast Fully Homomorphic Encryption over the Torus. *Journal of Cryptology* 33 (2020), 34–91. <https://doi.org/10.1007/s00145-019-09319-x>
- [13] Sean Choi, Disha Patel, Diman Zad Tootaghaj, Lianjie Cao, Faraz Ahmed, and Puneet Sharma. 2024. FedNIC: Enhancing Privacy-Preserving Federated Learning via Homomorphic Encryption Offload on SmartNIC. *Frontiers in Computer Science* 6 (Nov. 2024). <https://doi.org/10.3389/fcomp.2024.1465352>
- [14] Christopher A. Choquette-Choo, H. Brendan McMahan, Keith Rush, and Abhradeep Thakurta. 2023. Multi-epoch matrix factorization mechanisms for private machine learning. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) (*ICML*'23). JMLR.org, Article 236, 40 pages.
- [15] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. 2021. Exploiting Shared Representations for Personalized Federated Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 2089–2099. <https://proceedings.mlr.press/v139/collins21a.html>
- [16] Ivan Damgård, Valerio Pastro, Nigel P. Smart, and Sarah Zakarias. 2012. Multiparty Computation from Some-what Homomorphic Encryption. In *Advances in Cryptology – CRYPTO 2012 (Lecture Notes in Computer Science, Vol. 7417)*. Springer, 643–662. https://doi.org/10.1007/978-3-642-32009-5_38
- [17] Serguei Denissov, Hugh Brendan McMahan, J Keith Rush, Adam Smith, and Abhradeep Guha Thakurta. 2022. Improved Differential Privacy for SGD via Optimal Private Linear Operators on Adaptive Streams. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=i9XrHJoyLqJ>
- [18] Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. 2020. Personalized federated learning with moreau envelopes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (*NIPS* '20). Curran Associates Inc., Red Hook, NY, USA, Article 1796, 12 pages.
- [19] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (*NIPS* '20). Curran Associates Inc., Red Hook, NY, USA, Article 300, 12 pages.
- [20] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) (*WSDM* '20). Association for Computing Machinery, New York, NY, USA, 178–186. <https://doi.org/10.1145/3336191.3371856>
- [21] Pouya M. Ghari and Yanning Shen. 2022. Personalized Online Federated Learning with Multiple Kernels. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

- [22] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. 2022. An Efficient Framework for Clustered Federated Learning. *IEEE Transactions on Information Theory* 68, 12 (2022), 8076–8091. <https://doi.org/10.1109/TIT.2022.3192506>
- [23] Lovedeep Gondara and Ke Wang. 2020. Differentially Private Small Dataset Release Using Random Projections. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) (Proceedings of Machine Learning Research, Vol. 124)*, Jonas Peters and David Sontag (Eds.). PMLR, 639–648. <https://proceedings.mlr.press/v124/gondara20a.html>
- [24] Yuanxiong Guo, Ying Sun, Rui Hu, and Yanmin Gong. 2022. Hybrid Local SGD for Federated Learning with Heterogeneous Communications. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=H0oaWl6THa>
- [25] Yongxin Guo, Xiaoying Tang, and Tao Lin. 2023. FedBR: improving federated learning on heterogeneous data via local learning bias reduction. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) (ICML'23). JMLR.org, Article 483, 21 pages.
- [26] Neveen Mohammad Hijazi, Moayad Aloqaily, Mohsen Guizani, Bassem Ouni, and Fakhri Karray. 2024. Secure Federated Learning With Fully Homomorphic Encryption for IoT Communications. *IEEE Internet of Things Journal* 11, 3 (2024), 4289–4300. <https://doi.org/10.1109/JIOT.2023.3302065>
- [27] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. 2022. Efficient Split-Mix Federated Learning for On-Demand and In-Situ Customization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=_QLmakITKg
- [28] Songnam Hong and Jeongmin Chae. 2022. Communication-Efficient Randomized Algorithm for Multi-Kernel Online Federated Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2022), 9872–9886.
- [29] Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Donald Lane. 2021. FjORD: Fair and Accurate Federated Learning under heterogeneous targets with Ordered Dropout. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=4fLr7H5D_eT
- [30] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. 2021. Personalized Cross-Silo Federated Learning on Non-IID Data. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 9 (May 2021), 7865–7873. <https://doi.org/10.1609/aaai.v35i9.16960>
- [31] Nam Hyeon-Woo, Moon Ye-Bin, and Tae-Hyun Oh. 2022. FedPara: Low-rank Hadamard Product for Communication-Efficient Federated Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=d71n4ftoCBy>
- [32] Roger Iyengar, Joseph P. Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. 2019. Towards Practical Differentially Private Convex Optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*. 299–316. <https://doi.org/10.1109/SP.2019.00001>
- [33] Xiaopeng Jiang and Cristian Borcea. 2023. Complement sparsification: low-overhead model pruning for federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 908, 9 pages. <https://doi.org/10.1609/aaai.v37i7.25977>
- [34] Weizhao Jin, Yuhang Yao, Shanshan Han, Carlee Joe-Wong, Srivatsan Ravi, Salman Avestimehr, and Chaoyang He. 2023. FedML-HE: An Efficient Homomorphic-Encryption-Based Privacy-Preserving Federated Learning System. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*. <https://openreview.net/forum?id=PuYD0fh5aq>
- [35] Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. 2021. Practical and Private (Deep) Learning Without Sampling or Shuffling. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5213–5225. <https://proceedings.mlr.press/v139/kairouz21b.html>

- [36] Yeachan Kim, Junho Kim, Wing-Lam Mok, Jun-Hyung Park, and SangKeun Lee. 2023. Client-Customized Adaptation for Parameter-Efficient Federated Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1159–1172. <https://doi.org/10.18653/v1/2023.findings-acl.75>
- [37] Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto. <https://www.cs.toronto.edu/~kriz/cifar.html> Technical Report.
- [38] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2009. *CIFAR-10 (Canadian Institute for Advanced Research)*. Retrieved September 16, 2025 from <https://www.cs.toronto.edu/~kriz/cifar.html>
- [39] Mohit Kumar, Bernhard Moser, and Lukas Fischer. 2023. Secure Federated Learning with Kernel Affine Hull Machines. In *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. Bruges, Belgium, 121–126. <https://doi.org/10.14428/esann/2023.ES2023-56>
- [40] Mohit Kumar, Bernhard A. Moser, and Lukas Fischer. 2024. On Mitigating the Utility-Loss in Differentially Private Learning: A New Perspective by a Geometrically Inspired Kernel Approach. *Journal of Artificial Intelligence Research* 79 (2024), 515–567.
- [41] Mohit Kumar, Michael Rossbory, Bernhard A. Moser, and Bernhard Freudenthaler. 2019. Deriving an Optimal Noise Adding Mechanism for Privacy-Preserving Machine Learning. In *Database and Expert Systems Applications (DEXA 2019)*, CCIS, vol. 1062, G. Anderst-Kotsis and et al. (Eds.). Springer, Cham, 108–118. https://doi.org/10.1007/978-3-030-27684-3_15
- [42] Mohit Kumar, Alexander Valentinitisch, Magdalena Fuchs, Mathias Brucker, Juliana Bowles, Adnan Husakovic, Ali Abbas, and Bernhard A. Moser. 2025. Geometrically Inspired Kernel Machines for Collaborative Learning Beyond Gradient Descent. *Journal of Artificial Intelligence Research* 83 (July 2025), 35 pages. <https://doi.org/10.1613/jair.1.16821>
- [43] Mohit Kumar, Weiping Zhang, Lukas Fischer, and Bernhard Freudenthaler. 2023. Membership Mappings for Practical Secure Distributed Deep Learning. *IEEE Transactions on Fuzzy Systems* 31, 8 (2023), 2617–2631. <https://doi.org/10.1109/TFUZZ.2023.3235440>
- [44] Qinbin Li, Bingsheng He, and Dawn Song. 2023. Adversarial Collaborative Learning on Non-IID Features. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19504–19526. <https://proceedings.mlr.press/v202/li23j.html>
- [45] Yanan Li, Shusen Yang, Xuebin Ren, Liang Shi, and Cong Zhao. 2024. Multi-Stage Asynchronous Federated Learning With Adaptive Differential Privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 2 (2024), 1243–1256. <https://doi.org/10.1109/TPAMI.2023.3332428>
- [46] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. *arXiv preprint arXiv:2004.01401* (2020). <https://arxiv.org/abs/2004.01401>
- [47] Chong Liu, Yuqing Zhu, Kamalika Chaudhuri, and Yu-Xiang Wang. 2021. Revisiting Model-Agnostic Private Learning: Faster Rates and Active Learning. *Journal of Machine Learning Research* 22, 262 (2021), 1–44. <http://jmlr.org/papers/v22/20-1251.html>
- [48] Jiahao Liu, Jiang Wu, Jinyu Chen, Miao Hu, Yipeng Zhou, and Di Wu. 2023. FedDWA: personalized federated learning with dynamic weight adjustment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (Macao, P.R.China) (IJCAI '23)*. Article 444, 9 pages. <https://doi.org/10.24963/ijcai.2023/444>
- [49] Terrance Liu, Giuseppe Vietri, and Zhiwei Steven Wu. 2021. Iterative methods for private synthetic data: unifying framework and new methods. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 53, 13 pages.

- [50] Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. 2022. Personalized Federated Learning through Local Memorization. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 15070–15092. <https://proceedings.mlr.press/v162/marfoq22a.html>
- [51] MathWorks. 2025. *Deep Learning Toolbox Model for ResNet-50 Network*. Retrieved September 16, 2025 from <https://www.mathworks.com/matlabcentral/fileexchange/64626-deep-learning-toolbox-model-for-resnet-50-network>
- [52] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 1273–1282.
- [53] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=BJ0hF1Z0b>
- [54] Pratyay Mukherjee and Daniel Wichs. 2016. Two Round Multiparty Computation via Multi-key FHE. In *Proceedings, Part II, of the 35th Annual International Conference on Advances in Cryptology—EUROCRYPT 2016 - Volume 9666*. Springer-Verlag, Berlin, Heidelberg, 735–763.
- [55] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. 2022. Local and Central Differential Privacy for Robustness and Privacy in Federated Learning. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. <https://doi.org/10.14722/ndss.2022.23054>
- [56] Yao Pan, Zheng Chao, Wang He, Yang Jing, Li Hongjia, and Wang Liming. 2024. FedSHE: privacy preserving and efficient federated learning with adaptive segmented CKKS homomorphic encryption. *Cybersecurity* 7 (2024), 40. <https://doi.org/10.1186/s42400-024-00232-w>
- [57] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. 2016. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. *CoRR* abs/1610.05755 (2016). <http://arxiv.org/abs/1610.05755>
- [58] European Parliament and European Council. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Retrieved April 22, 2025 from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>
- [59] European Parliament and European Council. 2024. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. Retrieved April 22, 2025 from <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [60] NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. 2016. Differential Privacy Preservation for Deep Auto-Encoders: an Application of Human Behavior Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (Feb. 2016). <https://doi.org/10.1609/aaai.v30i1.10165>
- [61] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. McMahan, Sergei Vasilitskii, Steve Chien, and Abhradeep Thakurta. 2023. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy. *Journal of Artificial Intelligence Research* 77 (07 2023), 1113–1201. <https://doi.org/10.1613/jair.1.14649>
- [62] Xinchi Qiu, Javier Fernandez-Marques, Pedro PB Gusmao, Yan Gao, Titouan Parcollet, and Nicholas Donald Lane. 2022. ZeroFL: Efficient On-Device Training for Federated Learning with Local Sparsity. In *International Conference on Learning Representations*. https://openreview.net/forum?id=2sDQwC_hmnM
- [63] Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural Language Understanding with Privacy-Preserving BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 1488–1497. <https://doi.org/10.1145/3459637.3482281>

- [64] Nils Reimers and Iryna Gurevych. 2020. *distiluse-base-multilingual-cased-v2*. Retrieved September 16, 2025 from <https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2>
- [65] Nils Reimers and Iryna Gurevych. 2020. *paraphrase-multilingual-MiniLM-L12-v2*. Retrieved September 16, 2025 from <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>
- [66] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. 2020. Robust Federated Learning: The Case of Affine Distribution Shifts. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21554–21565. https://proceedings.neurips.cc/paper_files/paper/2020/file/f5e536083a438cec5b64a4954abc17f1-Paper.pdf
- [67] Jason Rennie. 2004. *20 Newsgroups Dataset (bydate version)*. Retrieved September 16, 2025 from <http://qwone.com/~jason/20Newsgroups/>
- [68] Yichen Ruan and Carlee Joe-Wong. 2021. FedSoft: Soft Clustered Federated Learning with Proximal Local Updating. In *AAAI Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:245123703>
- [69] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. 2021. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems* 32, 8 (2021), 3710–3722. <https://doi.org/10.1109/TNNLS.2020.3015958>
- [70] Mehdi Setayesh, Xiaoxiao Li, and Vincent W.S. Wong. 2023. PerFedMask: Personalized Federated Learning with Optimized Masking Vectors. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=hxEIgUXLFF>
- [71] Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. 2022. Federated Learning on Heterogeneous and Long-Tailed Data via Classifier Re-Training with Federated Features. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2218–2224. <https://doi.org/10.24963/ijcai.2022/308> Main Track.
- [72] Jiachen Shen, Yekang Zhao, Shitao Huang, and Yongjun Ren. 2024. Secure and Flexible Privacy-Preserving Federated Learning Based on Multi-Key Fully Homomorphic Encryption. *Electronics* 13, 22 (2024). <https://doi.org/10.3390/electronics13224478>
- [73] Yue Tan, Guodong Long, LU LIU, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. 2022. FedProto: Federated Prototype Learning across Heterogeneous Clients. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (Jun. 2022), 8432–8440. <https://doi.org/10.1609/aaai.v36i8.20819>
- [74] Daniel Truhn, Soroosh Tayebi Arasteh, Oliver Lester Saldanha, Gustav Müller-Franzes, Firas Khader, Philip Quirke, Nicholas P. West, Richard Gray, G. Hutchins, Jacqueline M. James, Maurice B. Loughrey, Manuel Salto-Tellez, Hermann Brenner, Alexander Brobeil, Tanwei Yuan, Jenny H. Chang, Jenny Chang-Claude, Michael Hoffmeister, Sebastian Foersch, Tianyu Han, Sebastian Keil, Maximilian Schulze-Hagen, Peter Isfort, Philipp Bruners, Georgios Kaassis, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, et al. 2024. Encrypted federated learning for secure decentralized collaboration in cancer image analysis. *Medical Image Analysis* 92 (Feb. 2024), 103059. <https://doi.org/10.1016/j.media.2023.103059>
- [75] Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. 2023. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1128, 10 pages. <https://doi.org/10.1609/aaai.v37i8.26197>
- [76] Qiong Wu, Kaiwen He, and Xu Chen. 2020. Personalized Federated Learning for Intelligent IoT Applications: A Cloud-Edge Based Framework. *IEEE Open Journal of the Computer Society* 1 (2020), 35–44. <https://doi.org/10.1109/OJCS.2020.2993259>
- [77] Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. 2023. Faster adaptive federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1166, 9 pages. <https://doi.org/10.1609/aaai.v37i9.26235>

- [78] Yilin Yang, Kamil Adamczewski, Xiaoxiao Li, Danica J. Sutherland, and Mijung Park. 2025. Differentially Private Neural Tangent Kernels (DP-NTK) for Privacy-Preserving Data Generation. *Journal of Artificial Intelligence Research* 81 (Jan. 2025), 18 pages. <https://doi.org/10.1613/jair.1.15985>
- [79] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2019. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=S1zk9iRqF7>
- [80] Zama. 2025. *TFHE-rs: A pure Rust implementation of the TFHE scheme*. Retrieved September 16, 2025 from <https://docs.zama.ai/tfhe-rs/0.6-3/get-started/benchmarks>
- [81] Feilong Zhang, Yinchuan Li, Shiyi Lin, yunfeng shao, Junjun Jiang, and Xianming Liu. 2023. Large Sparse Kernels for Federated Learning. <https://openreview.net/forum?id=ZCv4E1unfJP>
- [82] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42, 4, Article 25 (Oct. 2017), 41 pages. <https://doi.org/10.1145/3134428>
- [83] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2020. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=BJgnXpVYwS>
- [84] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. FedALA: adaptive local aggregation for personalized federated learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1261, 8 pages. <https://doi.org/10.1609/aaai.v37i9.26330>
- [85] Zhuangdi Zhu, Junyuan Hong, Steve Drew, and Jiayu Zhou. 2022. Resilient and Communication Efficient Learning for Heterogeneous Federated Systems. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 27504–27526. <https://proceedings.mlr.press/v162/zhu22e.html>
- [86] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. 2021. Data-Free Knowledge Distillation for Heterogeneous Federated Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12878–12889. <https://proceedings.mlr.press/v139/zhu21b.html>