AFRL-RI-RS-TR-2023-199

INTEGRAL: A FOUNDATIONAL APPROACH TO LABEL COMPLEXITY VIA INFORMATION THEORY AND GRAPH SIGNAL PROCESSING

UNIVERSITY OF SOUTHERN CALIFORNIA

NOVEMBER 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-199 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
PETER A. JEDRYSIK
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Divsion
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2023	FINAL TECHNICAL REPORT	SEPTEMBER 2019	MAY 2023

4. TITLE AND SUBTITLE

INTEGRAL: A FOUNDATIONAL APPROACH TO LABEL COMPLEXITY VIA INFORMATION THEORY AND GRAPH SIGNAL PROCESSING

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-19-2-1005	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2VW

6. AUTHOR(S)

Salman Avestimehr, Antonio Ortega, Mahdi Soltanolkotabi, and Ilias Diakonikolas

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

PRIME
University of Southern California
3740 McClintock Avenue
Los Angeles CA 90089

SUB
University of Wisconsin-Madison
1210 W. Dayton St.
Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RISB 525 Brooks Road Rome NY 13441-4505 DARPA/I20 675 N. Randolph St. Arlington VA 22203-2114	AFRL/RI & DARPA	AFRL-RI-RS-TR-2023-199

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The project's primary accomplishments can be summarized as follows: Firstly, the project made significant contributions to the field of transfer learning by establishing statistical minimax bounds. These bounds offer a precise understanding of the limits of knowledge transfer between related domains in classification and regression tasks. The study also extends to scenarios involving multiple source domains. Secondly, the research introduced a novel approach to understanding and optimizing complex deep learning systems through non-negative kernel regression (NNK) graphs, facilitating improved generalization estimation, clustering, and geometric metrics for network invariance assessment. Thirdly, the project rigorously assessed the performance of popular heuristics for data reduction, feature learning, and transfer learning. Lastly, the team proposed Federated Alternate Training (FAT) as a framework for global semi-supervised federated learning, providing a solution for collaboration in machine learning when labeled data is limited. Additionally, the project made significant contributions to statistical query lower bounds, showcasing their relevance in the presence of noisy data and cryptographic hardness, and also proposing gradient-descent type algorithms matching some of the lower bounds in specific cases.

15. SUBJECT TERMS

Non-Negative Kernel (NNK), Federated Alternate Training (FAT), Learning with Less Labels (LwLL), Technical Area 2 (TA2)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
PETER A. JEDRYSIK	N/A

PREVIOUS EDITION IS OBSOLETE.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY

This report summarizes work performed within the DARPA Learning with Less Labels (LwLL) program, specifically as part of technical area 2 (TA2), which focused on the development of theoretical bounds on the performance of machine learning systems. In particular, the goal of this project was to understand the fundamental limits of transfer learning, where a classifier trained for a first (source) task is modified for a second (target) task. Ultimately the performance of transfer learning depends on factors such as (i) the accuracy of the source classifier, (ii) the similarity between source and target tasks, (iii) the amount of labeled data available for the target task, and (iv) the transfer learning method. In this context, a theoretical bound on performance would allow us to determine what is the best possible performance on transfer learning given the amount data (iii), as a function of factors such as (i) and (ii) and under conditions on the nature of the transfer learning method. In this project, we developed empirical and theoretical methods leading to bounds on transfer learning performance.

A summary of the primary accomplishments of the project is as follows:

Statistical minimax bounds for transfer learning: We have developed a variety of mini-max statistical lower bounds for transfer learning. While there have been many recent algorithmic advances in this domain, a fundamental understanding of when and how much one can transfer knowledge from a related domain to reduce the amount of labeled training data is far from understood. We provide a precise answer to this question for classification and regression problems by deriving a novel lower bound on the generalization error that can be achieved by *any* transfer learning algorithm (regardless of its computational complexity) as a function of the amount of source and target samples. Our lower bound depends on a natural notion of distance that can be easily computed on real-world data sets. Other key features of our lower bound are that it does not depend on the source/target data distributions and requires minimal assumptions that enables its application to a broad range of problems. We also consider a more general setting where there are more than one source domains for knowledge transfer to the target task and develop new bounds on generalization error in this setting. We also corroborate our theoretical findings on real image classification from the LwLL benchmark datasets.

Geometry-based understanding and optimization of practical deep learning systems: One of the main challenges in achieving a theoretical understanding of practical machine learning systems is that analysis tools are often limited to simpler systems (e.g., fewer layers, linearity assumptions, etc.) than those used in practice. In our research, we have sought to understand complex systems by developing methods to characterize datasets via graph constructions. Our proposed non-negative kernel

regression (NNK) graphs connect data examples based on their similarity in any embedded space (e.g., any layer of a deep network) while eliminating geometric redundancy. We have used the NNK graph construction to address several problems: i) we have shown that local graph-based label interpolation is better at estimating how well a trained network will generalize, ii) clustering based on NNK leads to cluster centers belong to the data space and can be used for outlier detection, and iii) geometric metrics derive from NNK can help quantify network invariance and intrinsic dimension, leading to improved understanding of self-supervised learning.

Analyzing the Performance of Popular Heuristics: We rigorously investigated the performance of popular heuristics for data reduction, feature learning, and transfer learning.

Federated Alternate Training for Global Semi-Supervised Federated Learning (FL): Collaboration between data owners who have labeled data and those who do not is an important consideration for machine learning tasks. However, most of the current collaboration (or federated learning) based medical imaging research assumes that each data silo has ground truth labels for training, which is often not feasible especially in the medical field due to the high cost and time required to obtain accurate annotations. To make use of unannotated data silos for model improvement, we propose an alternative training-based framework called Federated Alternate Training (FAT). This approach involves training on annotated data silos to learn a reasonable global segmentation model, while unannotated data silos use the global segmentation model to generate pseudo labels for self-supervised learning. We show via experimental results on two naturally partitioned datasets, KiTS19 and FETS2021, that our proposed approach can outperform the state-of-the-art method.

Statistical Query Lower Bounds and Beyond: We developed the first super-polynomial Statistical Query (SQ) Lower Bounds for a wide range of fundamental learning problems in the presence of noisy data. These include the basic problem of learning a single neuron with adversarial or semi-random noise. Our SQ lower bounds complement the aforementioned minimax lower bounds established. Additionally, we established connections between SQ lower bounds and cryptographic hardness, obtaining reduction-based computational hardness for several of our studied problems. Finally, in some cases we obtained simple gradient-descent type algorithms nearly matching our lower bounds.

2 INTRODUCTION

In this section, we give an overview of the problems we tackle in this research project.

While there have been many recent algorithmic advances in the transfer learning domain, a fundamental understanding of when and how much one can transfer knowledge from a related domain to reduce the amount of labeled training data is far from understood. In our research, we provide a precise answer to this question for classification and regression problems by deriving a novel lower bound on the generalization error that can be achieved by *any* transfer learning algorithm (regardless of its computational complexity) as a function of the amount of source and target samples.

One of the main challenges in achieving a theoretical understanding of practical machine learning systems is that analysis tools are often limited to simpler systems (e.g., fewer layers, linearity assumptions, etc.) than those used in practice. In our research, we have sought to understand complex systems by developing methods to characterize datasets via graph constructions. Our proposed non-negative kernel regression (NNK) graphs connect data examples based on their similarity in any embedded space (e.g., any layer of a deep network) while eliminating geometric redundancy.

In this project, we also rigorously analyzed the performance of popular heuristics for data reduction, feature learning, and transfer learning.

Collaboration between data owners who have labeled data and those who do not is an important consideration for machine learning tasks. However, most of the current collaboration (or federated learning) based medical imaging research assumes that each data silo has ground truth labels for training, which is often not feasible especially in the medical field due to the high cost and time required to obtain accurate annotations. To make use of unannotated data silos for model improvement, we propose an alternative training-based framework called Federated Alternate Training (FAT).

We developed the first super-polynomial Statistical Query (SQ) Lower Bounds for a wide range of fundamental learning problems in the presence of noisy data. These include the basic problem of learning a single neuron with adversarial or semi-random noise. Our SQ lower bounds complement the aforementioned minimax lower bounds established.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Approach I: Statistical minimax bound analysis

3.1.1 Contribution I: Minimax lower bound for regression with one-hidden layer neural networks

In [1], we develop statistical minimax lower bounds for transfer learning in regression with one-hidden layer neural network models. In this setting, we use least squares loss for the target error and derive minimax lower bounds for target generalization error over the class of Gaussian distributions as a function of the number of source and target samples as well as an appropriate notion of transfer distance.

We evaluate our lower bounds on LWLL challenge data, specifically DomainNet-Clipart and DomainNet-Sketch as source and target tasks. The derived lower bounds consist of some data-dependent parameters which need to be estimated.

The problem at hand is classification but the lower bounds are derived for regression with least squares loss. In order to address the problem, we use one-hot encoded labels and consider least squares loss.

3.1.2 Contribution II: Minimax lower bound for classification with linear models and Gaussian distributions

In [2], we focus on transfer learning in binary classification with linear models and Gaussian features and develop statistical minimax lower bounds in terms of the number of source and target samples and an appropriate notion of similarity between source and target tasks. We first define a transfer distance based on the geodesic distance of the true parameters in the source and target, and then derive minimax lower bound over Gaussian distributions within a distance.

Next, we evaluate the derived lower bounds on the DomainNet dataset. By plotting the theoretical lower bounds as well as upper bounds obtained by weighted empirical risk minimization we investigate the sharpness of the lower bounds. Furthermore, we investigate that the defined semantic transfer distance conforms with the dataset.

3.1.3 Contribution III: Minimax lower bound for classification with a general hypothesis class and general class of distributions

In this work, we study binary classification where the goal is to learn a classifier from an arbitrary binary hypothesis class with finite VC dimension and also develop extensions to multiclass classification. This covers most contemporary classification models including training deep neural networks for classification tasks. In this setting, we first define a natural notion of similarity between source and target tasks via the performance of the best source hypothesis on the target task. Then equipped with this notion of similarity, we derive a statistical minimax lower bound on the target generalization error in terms of the number of labeled data from source and target tasks as well as the VC dimension of the hypothesis class for binary classification and Natarajan dimension for multiclass classification and the similarity between source and target tasks. Furthermore, we extend this result to the case where there are multiple sources with different similarity to the target. Our results demonstrate that sources with high similarity to the target are more effective at reducing the target generalization error. Indeed, a key feature of our result is that our lower bounds can be easily and efficiently computed on real data sets and applied to a broad class of practical settings.

3.2 Approach II: Geometry-based understanding and optimization of practical deep learning systems

3.2.1 Contribution I: NNK Graph Construction

NNK graph constructions are initialized with conventional k nearest neighbor (KNN) graphs based on a suitable similarity metric between nodes (data points). Thus, an initial KNN graph can be built for any available data representation. The NNK procedure essentially recomputes the edge weights obtained from KNN to eliminate “geometrically redundant” connections. Our idea is based on framing the graph construction as a signal representation problem, where enforcing orthogonality among the atoms used in the approximation (the set of neighbors) is equivalent to removing connections.

3.2.2 Contribution II: Local graph-based label interpolation and generalization

When applied to the penultimate layer of a neural network (before the fully connected layer), NNK provides for each point a list of its closest non-redundant neighbors along with their respective weights. We have proposed a local label interpolation along with its theoretical analysis. The basic idea is to use the labels of the NNK neighbors of one point to estimate the label of that point. Our theoretical model relates the performance

of this technique to properties of the NNK neighborhood, e.g., the diameter of the NNK polytope.

3.2.3 Contribution III: NNK-Means clustering and outlier detection

A major challenge in understanding the performance of modern ML systems is the lack of structure in the training of datasets used for training. These datasets are increasingly large (millions of objects) and diverse (high-level semantic classes with high variation in their content). Thus, while increasing the amount of training data continues to be a sound design strategy, there may be underlying biases in the data that are hard to detect. For example, in a pre-trained system exposed to real data, there can be drifts in the real input that are hard to detect, other than by observing degradation in classification performance. To address this problem, we have developed a clustering method, NNK-Means [15], based on the NNK graph construction.

3.2.4 Contribution IV: Manifold Graph Metrics (MGM), intrinsic dimension and invariance

We have developed a set of manifold graph metrics (MGMs) based on the NNK graph construction. These are local measurements, such as number of neighbors or polytope diameter, that vary depending on the intrinsic dimension of the embedded space [16]. By intrinsic dimension we mean the local dimension of the data manifold, which is generally (much) lower than the dimension of the ambient space (the dimension of the feature vectors).

We have applied these ideas to analyze self-supervised learning (SSL) systems [17]. In an SSL system, a network is trained without labels, using instead data augmentations.

3.3 Approach III: Analyzing the performance of popular heuristics

In this project, we also rigorously analyzed the performance of popular heuristics for data reduction, feature learning, and transfer learning.

3.3.1 Contribution I: Feature learning via gradient descent with applications to transfer learning

Significant theoretical work has established that in specific regimes, neural networks trained by gradient descent behave like kernel methods. However, in practice, it is known that neural networks strongly outperform their associated kernels by learning data-dependent features. In [18] we explain this gap by demonstrating that there is a

large class of functions that cannot be efficiently learned by kernel methods but can be easily learned with gradient descent on a two-layer neural network outside the kernel regime by learning representations that are relevant to the target task. We also demonstrate that these representations allow for efficient transfer learning, which is impossible in the kernel regime.

3.3.2 Contribution II: Feature learning in transformers via prompt-tuning

Prompt-tuning is an emerging strategy to adapt large language models (LLM) to downstream tasks by learning a (soft-)prompt parameter from data. Despite its success in LLMs, there is limited theoretical understanding of the power of prompt-tuning and the role of the attention mechanism in prompting. In [19], we explore prompt-tuning for one-layer attention architectures and study contextual mixture models where each input token belongs to a context-relevant or -irrelevant set. We isolate the role of prompt-tuning through a self-contained prompt-attention model.

3.4 Approach IV: Collaboration between Annotated and Unannotated Data Silos

In this project, we also studied the performance of transfer learning, and semi-supervised learning in a data heterogeneous setting of Federated Learning (FL).

3.4.1 Contribution I: Transfer Learning in FL for non-IID data distribution

Pre-training is a well-explored technique for training machine learning models in Centralized Learning (CL) settings [3]. Model initialization with a pre-trained model has been shown to enhance the generalizability and accuracy of models in CL. In our recent work [4], we consider it for the challenging setting of FL where data can be scarce and non-IID across silos such as medical datasets. We evaluate the benefit of pre-trained model initialization on two naturally partitioned medical datasets, KiTS19 [5] and FETS2021 [6], and show that pre-training closes the accuracy gap between federated Learning and its counterpart centralized learning by a significant margin.

3.4.2 Contribution II: Federated Alternate training to leverage Unannotated Data Silos in FL

In recent years, Federated Learning (FL) has been widely explored for medical applications [7]. However, most current works focus on supervised federated learning where all silos have pixel-wise annotations available. In practical scenarios, pixel-level label acquisition for massive medical imaging datasets requires a radiologist expert and therefore, can be time-consuming and expensive, so not all silos can afford it. Examples

are silos from rural regions with limited expert resources. It has motivated us to study the research question: How can a server leverage unannotated data silos, that have no labeled data, along with a few labeled data silos in a realistic non-independent and identical (non-IID) data distribution-based FL regime to improve the global model performance. Further, we focus on a more realistic scenario where the number of unannotated data silos can be larger than the annotated data silos.

3.5 Approach V: Statistical Query Complexity and Beyond

In this project, we analyzed the Statistical Query complexity of a range of fundamental learning problems, most notably in the presence of various types of noise. For some of these problems, we also developed nearly matching efficient algorithms based on gradient descent.

The family of Statistical Query (SQ) algorithms is a natural and well-studied class of algorithms encompassing essentially all known techniques in machine learning. Lower Bounds in the SQ model provide strong evidence of hardness, or more precisely the existence of statistical-computational tradeoffs for the underlying learning problem. In a sequence of works, we developed a unified set of techniques that led to the first super-polynomial SQ lower bounds for several fundamental learning problems – most notably in the presence of noise. Interestingly, some of these problems turn out to be easy in the presence of clean data, but become significantly more challenging computationally when the data is noisy. Different types of noise have been explored ranging from random to adversarial.

While the main focus of this approach was to establish lower bounds in the SQ and related models, in some cases we were able to obtain simple efficient algorithms that nearly match our lower bounds.

4 RESULTS AND DISCUSSION

4.1 Approach I: Statistical minimax bound analysis

Our specific contributions/results are described below.

4.1.1 Contribution I: Minimax lower bound for regression with one-hidden layer neural networks:

In [1], we develop statistical minimax lower bounds for transfer learning in regression with one-hidden layer neural network models. Based on the transfer distance between source and target, the derived lower bounds consist of three different regimes, namely high similarity, moderate similarity, and low similarity. In the low similarity regime, the lower bound only depends on the number of samples available from the target which means that the source samples are useful only up to a point, and beyond that point increasing the number of source samples does not decrease the target generalization error. On the other hand, in the high similarity regime, both source and target samples play an important role in decreasing the target generalization error. The moderate regime interpolates the two extreme regimes.

We evaluate our lower bounds on LWLL challenge data, specifically DomainNet-Clipart and DomainNet-Sketch as source and target tasks. The derived lower bounds consist of some data-dependent parameters which need to be estimated. The estimated parameters on the data reveal that the source and target lie in the low similarity regime and hence the lower bound is only a function of the number of target samples.
The problem at hand is classification but the lower bounds are derived for regression with least squares loss. In order to address the problem, we use one-hot encoded labels and consider least squares loss. Fig 1 plots the lower bound in terms of L2 target generalization error as a function of the number of target samples. Since the source and target are in the low similarity regime, the lower bound does not depend on the number of source samples. Fig 2 plots an upper bound obtained by simply minimizing the empirical risk over target samples.

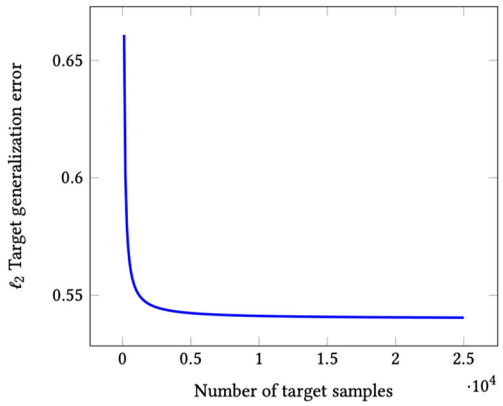


Figure 1. *Theoretical lower bound on target generalization error (measured in Euclidean distance between the output of the algorithm and the one-hot encoded labels) as a function of the number of target samples for DomainNet dataset.*

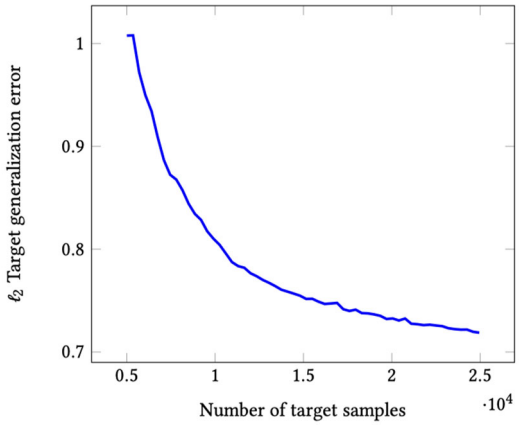


Figure 2. *Upper bound on target generalization error (measured in Euclidean distance between the output of the algorithm and the one-hot encoded labels) as a function of the number of target samples for DomainNet dataset. The upper bound is obtained by simply minimizing the empirical risk over target samples.*

4.1.2 Contribution II: Minimax lower bound for classification with linear models and Gaussian distributions

In [2], we focus on transfer learning in binary classification with linear models and Gaussian features and develop statistical minimax lower bounds in terms of the number of source and target samples and an appropriate notion of similarity between source and target tasks. Similar to the previous result, lower bounds consist of different regimes. When the distance of source and target tasks is high the corresponding lower bound is only a function of the target samples indicating that the target error is determined by the number of target samples and source samples are useful only up to a

point. On the other hand, in the regime where the distance between the source and the target is low the target error depends on both the number of source and target samples which demonstrates that source samples are useful when source is similar to the target.

Next, we evaluate the derived lower bounds on the DomainNet dataset. By plotting the theoretical lower bounds as well as upper bounds obtained by weighted empirical risk minimization, we investigate the sharpness of the lower bounds. Furthermore, we investigate that the defined semantic transfer distance conforms with the dataset.

We first pick three pairs of source and target tasks as described in Table 1. Then we extract features of dimension 2048 by passing the raw images through a ResNet50 network pre-trained on Imagenet. Then we train linear networks separately for source and target tasks. Using the estimated parameters, we calculate the semantic distance for each pair as shown in Table 1. To find the corresponding upper bounds, we run weighted empirical risk minimization.

Table 1. *Three pairs of source and target tasks along with corresponding semantic distance*

Tasks	Transfer distance
Target: Clock vs. Ambulance (Clipart)	-
Source1: Clock vs. Ambulance (Sketch)	0.35
Source2: Clock vs. apple (Sketch)	0.41
Source3:apple vs. animal-migration (Sketch)	0.48

As Table 1 shows, the pair (Source1, Target) has the lowest transfer distance among other pairs since both the source and target share the same objects, namely Clock and Ambulance. The semantic distance of pair 2 is less than that of pair 3 because in pair 2 the source and target share at least one common object which is Clock.

Fig 3 demonstrates that pairs with small semantic distance have lower target generalization error when the number of target samples is small. Because source samples would be more useful and compensate for the target samples.

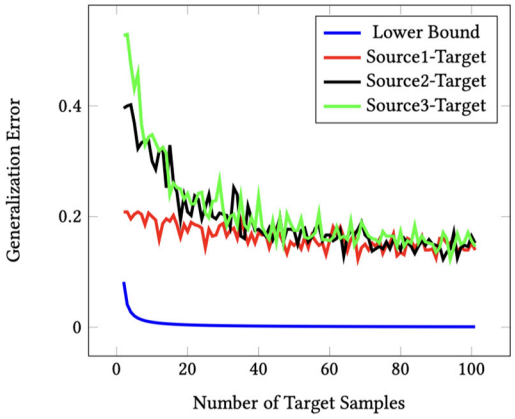


Figure 3. *Theoretical lower bound along with upper bounds for three pairs of source and target obtained by weighted empirical risk minimization*.

4.1.3 Contribution III: Minimax lower bound for classification with a general hypothesis class and general class of distributions

In summary our key contributions are as follows:

- We develop a novel statistical minimax lower bound on the generalization error that can be achieved for binary classification by any transfer learning algorithm as a function of the amount of source and target samples and a natural notion of similarity between source and target tasks.
- We also develop extensions of our result for multiclass classification problems based on a generalization of the VC dimension called Natarajan dimension.
- A key feature of our lower bound (including our notion of similarity) is that it can be easily computed on real world data sets. Furthermore, our lower bounds hold for any source/target distribution and apply with minimal assumptions to a wide variety of contemporary learning models including deep neural networks.
- We investigate the sharpness of our lower bounds and demonstrate their utility via experiments on action recognition and image classification.

Experiments on LWLL challenge data in binary classification:

Plotting the lower bounds requires estimating some parameters such as the semantic transfer distance, VC dimension, etc. By estimating the parameters for different pairs of tasks, we first plot the lower bounds and then by running weighted empirical risk minimization investigate the sharpness of the bounds. We also investigate the effectiveness of different source tasks with different transfer distances on the target generalization error.

Binary Action Recognition. We first perform experiments on the UCF101 action recognition data set. We pick CricketBowling and TableTennis videos from UCF101 as the target task as well as three different pairs of classes as the source tasks: 1- CricketBowling and BaseballPitch, 2- Cricketshot and Archery, 3- BasketballDunk and Basketball. We pass the videos through an i3d network pre-trained on kinetics400 with the fully connected top classifier removed and extract the corresponding features of dimension 2048 from the raw videos. We then work with the extracted features instead of the raw videos.

We then train a one hidden layer neural network with 15 hidden units and ReLU activation functions for each pair of data sets. Table 2 consists of test accuracies on the target task, i.e., CricketBowling vs. TableTennis, for three networks trained on source tasks. We use these accuracies for deriving the corresponding lower bounds. Furthermore, we run weighted empirical risk minimization as a simple transfer learning approach to find some upper bounds on the target generalization error.

Table 2. *Three pairs of source and target tasks from UCF101 data set on action recognition along with corresponding transfer distances.*

Task	Target test accuracy for source network	ρ(Source, Target)
Target: CricketBowling vs. TableTennis	-	-
Source1: CricketBowling vs. Baseball Pitch	0.946	0.054
Source2: Cricketshot vs. Archery	0.61	0.39
Source3: BasketballDunk vs. Basketball	0.52	0.48

Results on binary action recognition. As it can be observed in Table 2, the pair of Source1 and Target has the lowest transfer distance among other pairs since both of the source and target tasks share the same class which is CricketBowling. Furthermore, this column determines which pairs are more suitable for transferring the source knowledge to the target. Then in Figure 4b we plot the lower bounds along with the upper bounds obtained via empirical risk minimization. Fig 4b shows that when the distance of a source from the target is small it would be more effective in achieving small target generalization error.

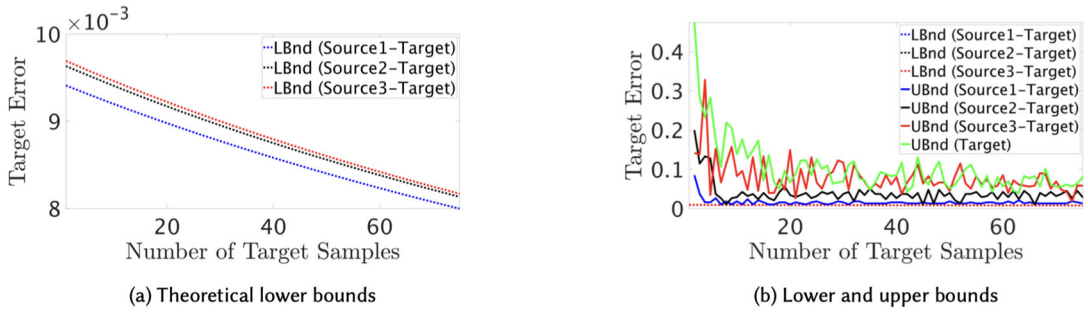


Figure 4. *(a) depicts the lower bounds for three pairs of source and target tasks on action classification. (b) depicts the lower bounds along with the upper bounds obtained via weighted empirical risk minimization.*

Binary Image Classification. Here we focus on image classification tasks and recognize appropriate pairs of tasks that are suitable for transfer learning. We choose three pairs of source and target tasks based on Table 3 and then extract the features using a ResNet50 network pre-trained on Imagenet. We train a one hidden layer neural network with 15 hidden units and ReLU activation functions for each of pairs of the tasks. Table 3 demonstrates the results. We also run weighted empirical risk minimization for finding upper bounds for the pairs of the source and target.

Table 3. *Three pairs of source and target tasks from DomainNet data set on image classification. The second column consists of target test accuracies for source networks. The third column consists of transfer distances for each pair of source/target.*

Task	Target Test Acc	Transfer Distance
Target: Clock vs. Ambulance (Sketch)	-	-
Source1: Clock vs. Ambulance (Clipart)	0.916	0.084
Source2: Clock vs. Crow (Clipart)	0.7	0.3
Source3: Crow vs. House (Clipart)	0.65	0.35

Results on binary image classification. We plot the lower bounds in Fig 5a and the corresponding upper bounds obtained by weighted empirical risk minimization in Fig 5b. One can see that sources that are closer to the target according to our notion of distance are more effective in achieving small target generalization error.

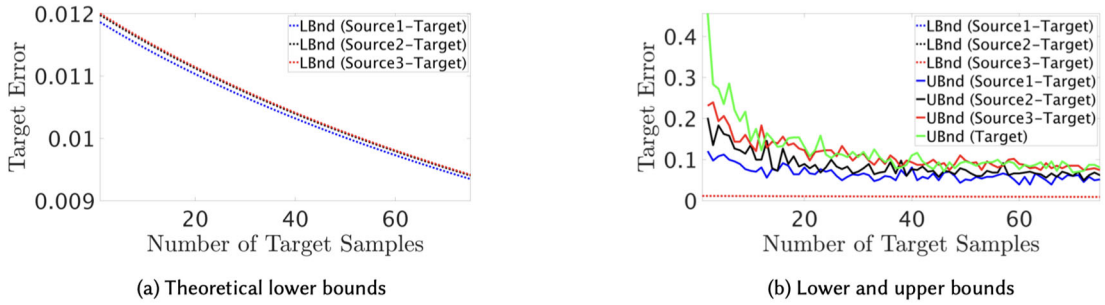


Figure 5. *(a) depicts the lower bounds for three pairs of source and target tasks on image classification. (b) depicts the lower bounds along with the upper bounds obtained via weighted empirical risk minimization*.

4.2 Approach II: Geometry-based understanding and optimization of practical deep learning systems

4.2.1 Contribution I: NNK Graph Construction

Our idea is based on framing the graph construction as a signal representation problem, where enforcing orthogonality among the atoms used in the approximation (the set of neighbors) is equivalent to removing connections. This idea is illustrated in Figure 6, where among the neighbors of a specific data point *i,* only those points that are closest along the same direction are connected, i.e., the blue nodes remain connected, while the orange nodes are disconnected.

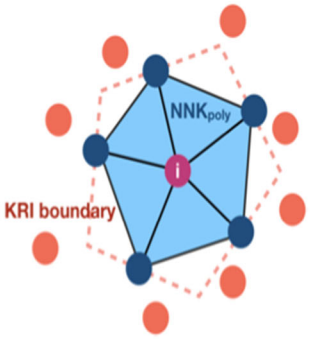


Figure 6. *Framing the graph construction as a signal representation problem, where enforcing orthogonality among the atoms used in the approximation (the set of neighbors) is equivalent to removing connections.*

In our work, we use multiple representations for each data point, and, in particular, we construct separate graphs corresponding to the same set of data points with their representation at the outputs of each layer in a neural network. This allows us to

compare how the relative positions of the points evolve during training (across epochs) and as part of the feature extraction (across layers). The key benefit of the NNK construction is its geometric interpretation, which allows us to compare graphs obtained from the same data in different embedded spaces, while also providing insights into the intrinsic dimensions of the dataset. Specifically, since redundant connections (see Figure 6) are removed the number of NNK neighbors given a sufficiently large value *k* in the KNN initialization is a function of the intrinsic dimension of the data (given large enough *k* more NNK neighbors means higher intrinsic dimension). More details on the NNK construction can be found in [12].

4.2.2 Contribution II: Local graph-based label interpolation and generalization

When applied to the penultimate layer of a neural network (before the fully connected layer), NNK provides for each point a list of its closest non-redundant neighbors along with their respective weights. We have proposed a local label interpolation along with its theoretical analysis. The basic idea is to use the labels of the NNK neighbors of one point to estimate the label of that point. Our theoretical model relates the performance of this technique to properties of the NNK neighborhood, e.g., the diameter of the NNK polytope. As a practical application, we develop a leave-one-out performance estimator, where we determine if the (known) label of the point at the center of a polytope can be correctly interpolated from its neighbors' labels. Our experimental results show that this leave-one-out technique provides more reliable estimates of generalization performance [13]. This result can be used to improve training and reduce overfitting. In particular, when comparing the error resulting from using the fully connected layer for classification to the error resulting from the leave-one-out local interpolation, we have observed that the two errors are similar in general, but the estimated leave-one-out error is larger under overfitting conditions. Thus, one can change training conditions or network parameters when overfitting is detected using this method, without a need to use a separate validation set. Additionally, local graph constructions can be applied channel-wise to help understand whether channels in convolutional neural network capture redundant information [14].

The key insight behind this result can be understood by noting that the fully connected layers use many parameters to describe class boundaries. Thus, even if the data embedding does not clearly separate data points from different labels (e.g., a node has NNK neighbors corresponding to multiple different labels), a class boundary with a sufficiently large number of parameters can achieve good performance on the training set. In contrast, local interpolation can directly characterize the class separation achieved with a given embedding since it has no additional parameters optimized for classification. Instead, the estimated label is based on the relative position of a node with respect to its neighbors and their respective labels. Thus, the labels obtained from

local interpolation are more stable under perturbations of the input than those obtained from a complex optimized boundary, and we may expect better estimates of generalization performance from local interpolation.

4.2.3 Contribution III: NNK-Means clustering and outlier detection

A major challenge in understanding the performance of modern ML systems is the lack of structure in the training of datasets used for training. These datasets are increasingly large (millions of objects) and diverse (high-level semantic classes with high variation in their content). Thus, while increasing the amount of training data continues to be a sound design strategy, there may be underlying biases in the data that are hard to detect. For example, in a pre-trained system exposed to real data, there can be drifts in the real input that are hard to detect, other than by observing degradation in classification performance.

To address this problem, we have developed a clustering method, NNK-Means [15], based on the NNK graph construction. In NNK-Means, given the current set of cluster centers, an input is assigned to a subset of clusters using the NNK graph construction. Because the NNK weights are non-negative and the selected clusters are not geometrically redundant, the resulting clusters provide a better approximation than K-Means clustering (each point assigned to more than one cluster center) while selecting cluster centers that remain in the data space (unlike in the case of dictionary learning). One concrete practical use of these clusters is in outlier detection for cases where data may be experiencing drift. In this case, we assign new data received in real time to the pre-design clusters. The resulting assignment weights can be interpreted as a measurement of how well the new data fits the existing model, represented by the cluster centers.

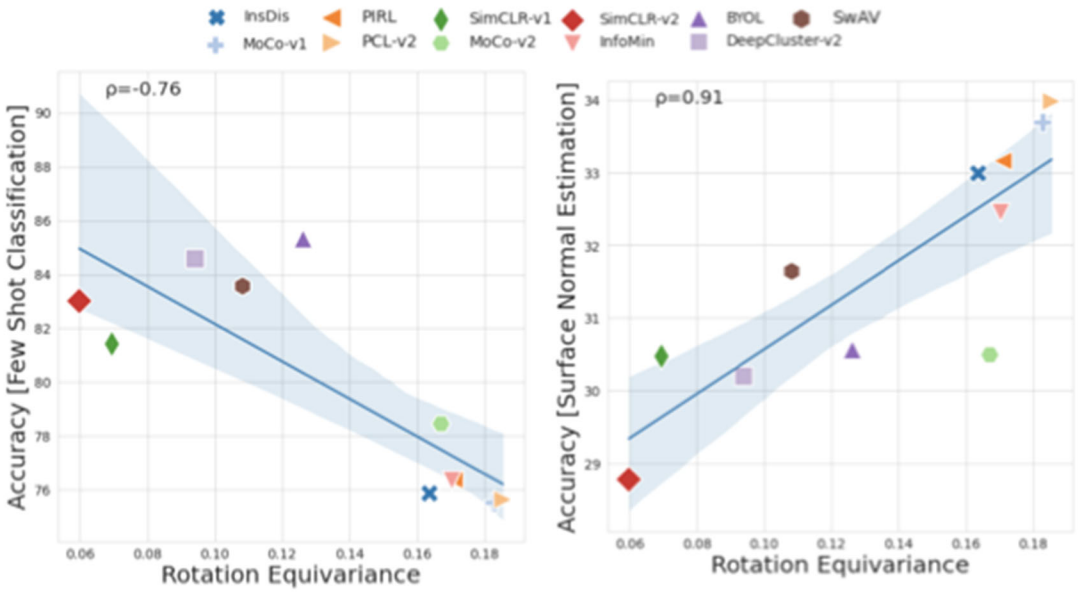
4.2.4 Contribution IV: Manifold Graph Metrics (MGM), intrinsic dimension and invariance

We have developed a set of manifold graph metrics (MGMs) based on the NNK graph construction. These are local measurements, such as number of neighbors or polytope diameter, that vary depending on the intrinsic dimension of the embedded space [16]. By intrinsic dimension we mean the local dimension of the data manifold, which is generally (much) lower than the dimension of the ambient space (the dimension of the feature vectors).

We have applied these ideas to analyze self-supervised learning (SSL) systems [17]. In an SSL system, a network is trained without labels, using instead data augmentations. For example, for an image processing task, one can use images and rotated versions of

the original images. The network is trained to discriminate among different images while keeping the distance between an image and its augmentation as small as possible. This type of training aims to ensure that the network has built-in invariance with respect to the augmentation (e.g., rotation).

In the example in Figure 7 below, we show how the NNK construction can shed some light on this process. We create NNK polytopes with an image and all its augmentations, all of them represented in the embedded space obtained via SSL. The diameter of these NNK polytopes captures the degree of invariance of the network (smaller diameter implies more invariance or less equivariance). As can be seen in Figure 7, in the few-shot object classification (where invariance to rotation is important) those SSL networks with better invariance (lower equivariance) perform better. This is in contrast with a surface normal estimation task, where rotation should affect the result and thus better performance can be achieved with less invariance to rotation.



.

Figure 7. *Accuracy versus rotation equivariance in few-shot object classification and surface normal estimation tasks.*

4.3 Approach III: Analyzing the performance of popular heuristics

In this project, we also rigorously analyzed the performance of popular heuristics for data reduction, feature learning, and transfer learning. Our specific contributions/results are described below.

4.3.1 Contribution I: Feature learning via gradient descent with applications to transfer learning

Significant theoretical work has established that in specific regimes, neural networks trained by gradient descent behave like kernel methods. However, in practice, it is known that neural networks strongly outperform their associated kernels by learning data-dependent features. In [18] we explain this gap by demonstrating that there is a large class of functions that cannot be efficiently learned by kernel methods but can be easily learned with gradient descent on a two-layer neural network outside the kernel regime by learning representations that are relevant to the target task. We also demonstrate that these representations allow for efficient transfer learning, which is impossible in the kernel regime.

Specifically, we consider the problem of learning polynomials which depend on only a few relevant directions, i.e., of the form $f_*(x) = g(Ux)$ where $U: \mathbb{R}^d \mapsto \mathbb{R}^r$ with $d \gtrapprox r$. When the degree of f_* is p, it is known that $n \propto dp$ samples are necessary to learn f_* in the kernel regime. Our primary result is that gradient descent learns a representation of the data which depends only on the directions relevant to f_*. This results in an improved sample complexity of $n \propto dr^2 + drp$. Furthermore, in a transfer learning setup where the data distributions in the source and target domain share the same representation U but have different polynomial heads, we show that a popular heuristic for transfer learning has a target sample complexity independent of *d*.

4.3.2 Contribution II: Feature learning in transformers via prompt-tuning

Prompt-tuning is an emerging strategy to adapt large language models (LLM) to downstream tasks by learning a (soft-)prompt parameter from data. Despite its success in LLMs, there is limited theoretical understanding of the power of prompt-tuning and the role of the attention mechanism in prompting. In [19], we explore prompt-tuning for one-layer attention architectures and study contextual mixture models where each input token belongs to a context-relevant or -irrelevant set. We isolate the role of prompt-tuning through a self-contained prompt-attention model. Our contributions are as follows: (1) We show that softmax-prompt-attention is provably more expressive than softmax-self-attention and linear-prompt-attention under our contextual data model. (2) We analyze the initial trajectory of gradient descent and show that it learns the prompt and prediction head with near-optimal sample complexity and demonstrate how the prompt can provably attend to sparse context-relevant tokens. (3) Assuming a known prompt but an unknown prediction head, we characterize the exact finite sample performance of prompt-attention which reveals the fundamental performance limits and the precise benefit of the context information. We also provide experiments that verify our theoretical insights on real datasets and demonstrate how prompt-tuning enables the model to attend to context-relevant information.

4.4 Approach IV: Collaboration between Annotated and Unannotated Data Silos

In this project, we also studied the performance of transfer learning, and semi-supervised learning in a data heterogeneous setting of Federated Learning (FL). Our specific contributions/results are described below.

4.4.1 Contribution I: Transfer Learning in FL for non-IID data distribution

Pre-training is a well-explored technique for training machine learning models in Centralized Learning (CL) settings [3]. Model initialization with a pre-trained model has been shown to enhance the generalizability and accuracy of models in CL. In our recent work [4], we consider it for the challenging setting of FL where data can be scarce and non-IID across silos such as medical datasets. We evaluate the benefit of pre-trained model initialization on two naturally partitioned medical datasets, KiTS19 [5] and FETS2021 [6], and show that pre-training closes the accuracy gap between federated Learning and its counterpart centralized learning by a significant margin.

4.4.2 Contribution II: Federated Alternate training to leverage Unannotated Data Silos in FL

In recent years, Federated Learning (FL) has been widely explored for medical applications [7]. However, most current works focus on supervised federated learning where all silos have pixel-wise annotations available. In practical scenarios, pixel-level label acquisition for massive medical imaging datasets requires a radiologist expert and therefore, can be time-consuming and expensive, so not all silos can afford it. Examples are silos from rural regions with limited expert resources. It has motivated us to study the research question: How can a server leverage unannotated data silos, that have no labeled data, along with a few labeled data silos in a realistic non-independent and identical (non-IID) data distribution-based FL regime to improve the global model performance. Further, we focus on a more realistic scenario where the number of unannotated data silos can be larger than the annotated data silos.

To leverage the unannotated data silos to improve modeling, most existing works [8] use variants of federated averaging [9] where all silos participate at each training round. Contrary to these approaches, we propose an alternate training-based framework, Federated Alternate Training (FAT), that alters training between annotated data silos and unannotated data silos. This idea is inspired by work [10], where they assume the server has access to IID annotated data. However, in our work, we use a more realistic and challenging scenario for medical imaging where the server does not have access to any data, but a few silos may have labeled but non-IID data available. In FAT, as shown in Figure 8, annotated data silos exploit high-quality annotations to learn a reasonable global segmentation model. Once the server receives the model weights from the annotated data silos $\{\theta_1, \theta_2, \ldots, \theta_S\}$, it aggregates the model weights obtained from the supervised silos, $\sum_{k=1}^{S} \frac{N_k}{\sum_{i=1}^{S} N_i}(\theta_k)$, and send it to unannotated data silos. Meanwhile, unannotated data silos use the global segmentation model as a target model to generate pseudo labels for self-supervised learning. For self-supervised learning, we leverage mixup data augmentation, $x' = \lambda\, x_1 + (1 - \lambda)\, x_2$, to perturb the two input images $x_1$and x_2, where $\lambda \in (0,1)$ and is a hyperparameter. We use Dice and cross-entropy loss as our consistency loss. To keep self-supervised learning stable, we also leverage the exponential moving average where the global model (θ) is updated via the exponential moving average of the local student model (ξ), $\theta = \tau\,\theta + (1 - \tau)\,\xi$. After the unannotated data silos send the model weights to the server, the server aggregates the model weights, $\sum_{k=S+1}^{K} \frac{N_k}{\sum_{i=S+1}^{K} N_i}(\theta_k)$, and sends them to annotated data silos for further training.

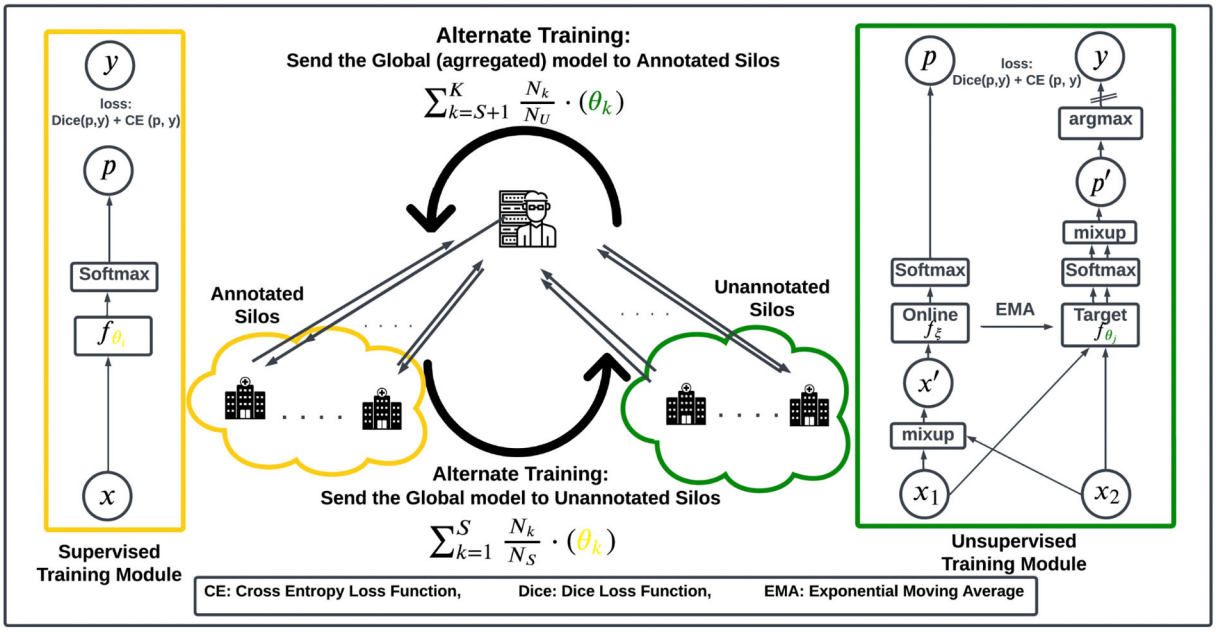


Figure 8. *The proposed Federated Alternate Training (FAT) framework, where we alternate training between Annotated Data Silos and Unannotated Data Silos. The Annotated Data Silos follow a supervised training module with ground truth labels available. The Unannotated Data Silos follow a bootstrapping-based self-supervised training module where the target model generates pseudo labels, y, for the self-supervised learning and uses exponential moving average (EMA) for the model updates.*

Results. To evaluate our proposed framework, we used a federated version of KiTS19 dataset given in [7] work, where we used 6 silos as training silos and the rest of the silos were used as test silos. To report transfer learning results, we compare random model initialization versus pre-trained model initialization. We use a model pretrained on LiTS dataset [11]. As shown in Figure 9a, we observe that the tumor dice score gap between centralized learning and federated learning drops from 3.3% to 1.1% on the tumor segmentation task. This highlights the significance of transfer learning in the data heterogeneous setting of FL. Next, we evaluate our proposed framework FAT in a setting where two silos have annotations available and four silos do not have annotations. First, we compare our results to the setting where we only exploit the two silos that have annotations available. As can be seen from Figure 9b, we achieve a gain of 17.7% by leveraging unannotated data silos. Further, we compare our method to the state-of-the-art algorithm [8] to leverage unannotated data silos. We outperform the SOTA method [8] by achieving a 10.2% higher tumor dice score on the tumor segmentation task. This shows the significance of alternate training in a label-heterogeneous setting of FL.

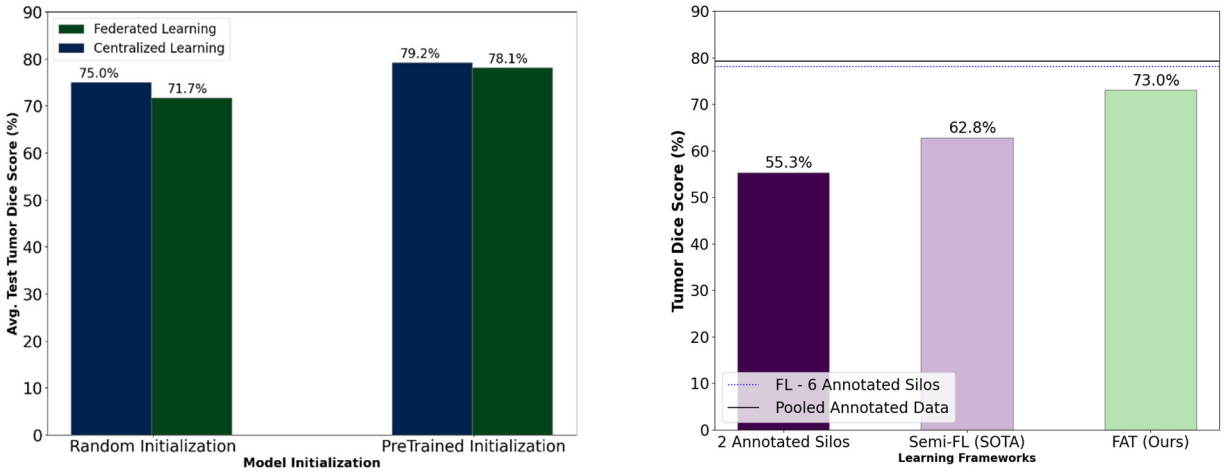


Figure 9. *Transfer learning via PreTrained Model initialization, and Federated Alternate Training results on KiTS19 dataset for tumor segmentation task. a) Random versus PreTrained Model Initialization Results. b) Proposed Federated Alternate Training Framework Results*

4.5 Approach V: Statistical Query Complexity and Beyond

Our specific contributions/results are described below.

4.5.1 Contribution I: Statistical Query Lower Bounds for Supervised Learning with Noise

SQ Lower Bounds for Learning Simple Neural Networks

In [20], we studied the fundamental problem of learning one-hidden layer neural networks (without noise), where the coefficients and the weight vectors are unknown. We focused on the simple case where the distribution over features is well-behaved, namely the features follow the Gaussian distribution. This is arguably the simplest possible distributional assumption that appeared to be amenable to efficient algorithms. This work established two main results. On the lower bound side, we gave an SQ lower bound suggesting that any algorithm for this problem requires time at least d^k, where d is the dimension and k is the number of hidden units. On the positive side, we gave an efficient algorithm that beats our lower bound if the coefficients in the linear combination are assumed to be non-negative.

SQ Lower Bounds for List-decodable Linear Regression

In [21], we showed near-optimal SQ lower bounds for the problem of list-decodable linear regression. The setup is as follows: We are given a dataset of size *n* in *d* dimensions a parameter 0< alpha<1/2 with the promise that alpha-fraction of these points are i.i.d. samples from an unknown (target) linear regression model. No assumptions are made on the remaining points, which could be selected even adversarially. This setting models situations where the clean data is the minority of the given dataset, and has been extensively studied recently due to its applications (e.g., to data poisoning attacks in ML) and its connection to mixture models. In the list-decodable setting, the goal is to output a small list of candidate solutions with the guarantee that at least one of them is a good approximation to the target. Prior work had developed algorithms for list-decodable linear regression with sample and time complexity $d^{1/alpha}$. On the other hand, *d*/*alpha* samples information-theoretically suffice to solve the problem (ignoring runtime). We show that this sample-time tradeoff is essentially optimal, but establishing a matching SQ lower bound for the problem.

SQ and Cryptographic Hardness for Learning with Adversarial Label Noise

In a sequence of works, we established optimal SQ lower bounds and related cryptographic hardness for learning various simple supervised models with adversarial label noise. Interestingly, the cryptographic hardness reductions were inspired by the SQ-hard instances. In [22], we showed that the problems of learning linear threshold functions and ReLUs to optimal accuracy of OPT+eps in the agnostic learning model (i.e., with adversarial label noise) requires time $d^{1/eps}$, even if the distribution on features is Gaussian. These lower bounds qualitatively matched known algorithms for these problems. Subsequently, in [23], we developed a convex duality framework that allows us to show the following: for essentially {\em any} class of functions, polynomial regression is optimal (within the class of SQ algorithms) for agnostic learning under the Gaussian distribution. In [24], we developed super-polynomial SQ lower bounds for agnostically learning a single neuron in the distribution-free setting. Importantly, these lower bounds rule out even {\em approximate} learners. Inspired by these SQ lower bound constructions, in [25] we showed that these problems are computationally hard under the widely believed hardness of the Learning with Errors (LWE) problem.

SQ and Cryptographic Hardness for Learning with Semi-Random Label Noise

The preceding works established SQ hardness for basic learning problems under the assumption that a small constant fraction of the labels has been corrupted adversarially. This adversarial contamination may be too pessimistic in some scenarios. Hence, it is natural to ask how the complexity of these problems changes in the presence of semi-random noise. For example, suppose that the label of each example is perturbed independently with some small instance-dependent probability. While this noise model

may appear innocuous, it turns out that computational limits show up for a number of basic problems. In [26], we give the first super-polynomial SQ lower bounds for learning halfspaces in this semi-random noise model. Our lower bound essentially matches previous algorithms established by the co-PI Diakonikolas and collaborators. More recently, in [27], we leveraged this SQ-hard construction to obtain a computational hardness result for this problem, under the assumed hardness of LWE. These two works [26, 27] nearly characterize the complexity of learning halfspaces with semi-random label noise, resolving a longstanding open problem in the theory of machine learning. These ideas also led to qualitatively similar hardness results for more general neurons, including ReLUs.

4.5.2 Contribution II: Efficient Gradient-descent-based Algorithms

In [28, 29], we gave a gradient-descent-based algorithm for agnostically learning ReLUs within error O(OPT}+eps under all log-concave distributions. Recall that achieving error OPT+\eps is computationally hard (as shown in our aforementioned work). Essentially what we show here is that if we relax the desired error by a small constant factor, efficient algorithms exist. Our approach leveraged a natural non-convex relaxation of the problem. Finally, in [30] we gave an efficient algorithm ReLU regression with semi-random noise in the distribution-free setting.

5 CONCLUSIONS

In the Statistical Minimax Bound Analysis approach, we developed minimax lower bounds for regression with one-hidden layer neural networks; for classification with linear models and Gaussian distributions; and for classification with a general hypothesis class and general class of distributions.

In the Geometry-based Understanding and Optimization of Practical Deep Learning Systems approach, our contributions are NNK graph construction, local graph-based label interpolation and generalization, NNK-Means clustering and outlier detection, and developing manifold graph metrics (MGMs) based on the NNK graph construction.

In the Analyzing the Performance of Popular Heuristics approach, we rigorously investigated the performance of popular heuristics for data reduction, feature learning, and transfer learning.

In the Collaboration between Annotated and Unannotated Data Silos approach, we studied the performance of transfer learning, and semi-supervised learning in a data heterogeneous setting of Federated Learning (FL).

In the Statistical Query Complexity and Beyond approach, we analyzed the Statistical Query complexity of a range of fundamental learning problems, most notably in the presence of various types of noise. For some of these problems, we also developed nearly matching efficient algorithms based on gradient descent.

6 REFERENCES

[1] Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. 2020. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. Advances in Neural Information Processing Systems 33 (2020), 1959–1969.

[2] Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. 2022. Statistical Minimax Lower Bounds for Transfer Learning in Linear Binary Classification. In 2022 IEEE International Symposium on Information Theory (ISIT). IEEE, 282–287.

[3] Hendrycks, Dan, Kimin Lee, and Mantas Mazeika. "Using pre-training can improve model robustness and uncertainty." In *International conference on machine learning*, PMLR, 2019, pp. 2712-2721.

[4] Mushtaq, Erum, Yavuz Faruk Bakman, Jie Ding, and Salman Avestimehr, "Federated Alternate Training (FAT): Leveraging Unannotated Data Silos in Federated Segmentation for Medical Imaging." *International Symposium on Biomedical Imaging (ISBI),* 2023.

[5] Heller, Nicholas, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg et al. "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes." *arXiv preprint arXiv:1904.00445* (2019).

[6] Pati, Sarthak, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G. Anthony Reina, Patrick Foley et al. "The federated tumor segmentation (fets) challenge." *arXiv preprint arXiv:2105.05874* (2021)

[7] Ogier du Terrail, Jean, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold et al., "FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings," *Advances in Neural Information Processing Systems,* 35, 2022, pp. 5315-5334.

[8] Yang, Dong, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R. Roth, Stephanie Harmon, Sheng Xu et al., "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Medical image analysis,* **70**, 2021, pp. 101992.

[9] McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," In *Artificial intelligence and statistics*, *PMLR*, 2017, pp. 1273-1282.

[10] Diao, Enmao, Jie Ding, and Vahid Tarokh, "SemiFL: Semi-supervised federated learning for unlabeled clients with alternate training," *Advances in Neural Information Processing Systems,* **35** 2022, pp. 17871-17884.

[11] Bilic, Patrick, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin et al., "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis,* **84**, 2023, pp. 102680.

[12] Shekkizhar, S., and A. Ortega (2019). Neighborhood and Graph Constructions using Non-Negative Kernel Regression. *ArXiv*. /abs/1910.09383

[13] S. Shekkizhar and A. Ortega, "Revisiting Local Neighborhood Methods in Machine Learning," *2021 IEEE Data Science and Learning Workshop (DSLW)*, Toronto, ON, Canada, 2021, pp. 1-6, doi:10.1109/DSLW51110.2021.9523409.

[14] D. Bonet, A. Ortega, J. Ruiz-Hidalgo and S. Shekkizhar, "Channel Redundancy and Overlap in Convolutional Neural Networks with Channel-Wise NNK Graphs," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 4328-4332, doi: 10.1109/ICASSP43922.2022.9746186.

[15] S. Shekkizhar and A. Ortega, "NNK-Means: Data summarization using dictionary learning with non-negative kernel regression," *2022 30th European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022, pp. 2161-2165, doi: 10.23919/EUSIPCO55093.2022.9909928.

[16] Cosentino, R., Shekkizhar, S., Soltanolkotabi, M., Avestimehr, S., & Ortega, A. (2022). The Geometry of Self-supervised Learning Models and its Impact on Transfer Learning. *ArXiv*. /abs/2209.08622

[17] C. Hurtado, S. Shekkizhar, J. Ruiz-Hidalgo and A. Ortega, "Study of Manifold Geometry Using MultiscaleNon-Negative Kernel Graphs," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095956.

[18] A. Damian, J. Lee, and M. Soltanolkotabi. 2022. Neural networks can learn representations with gradient descent. In Conference on Learning Theory. PMLR, 5413–5452.

[19] S. Oymak, A. Rawat, M. Soltanolkotabi, and T. Christos. 2023. On the role of attention in prompt-tuning. In International Conference on Machine Learning.

[20] Diakonikolas, Ilias and Kane, Daniel M and Kontonis, Vasilis and Zarifis, Nikos. Algorithms and SQ Lower Bounds for PAC Learning One-Hidden-Layer ReLU Networks, Conference on Learning Theory, 1514--1539, 2020.

[21] I. Diakonikolas, D. Kane, A. Pensia, T. Pittas, A. Stewart. Statistical Query Lower Bounds for List-Decodable Linear Regression, NeurIPS 2021, 3191--3204, 2021.

[22] I. Diakonikolas, D. Kane, N. Zarifis. Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals. Advances in Neural Information Processing Systems (NeurIPS 2020)

[23] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, Nikos Zarifis. The Optimality of Polynomial Regression for Agnostic Learning under Gaussian Marginals in the SQ Model. Conference on Learning Theory, COLT 2021, 1552-1584, 2021.

[24] I. Diakonikolas, D. Kane, P. Manurangsi, L. Ren. Hardness of Learning a Single Neuron with Adversarial Label Noise. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022).

[25] Ilias Diakonikolas, Daniel M. Kane, and Lisheng Ren. Near-Optimal Cryptographic Hardness of Agnostically Learning Halfspaces and ReLU Regression under Gaussian Marginals. ICML 2023, 7922-7938.

[26] I. Diakonikolas, D. Kane. Near-Optimal Statistical Query Hardness of Learning Halfspaces with Massart Noise. Proceedings of the 35th Annual Conference on Learning Theory (COLT 2022).

[27] I. Diakonikolas, D. Kane, P. Manurangsi, L. Ren. Cryptographic Hardness of Learning Halfspaces with Massart Noise. Advances in Neural Information Processing Systems (NeurIPS 2022).

[28] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2022. Learning a Single Neuron with Adversarial Label Noise via Gradient Descent. In Conference on Learning Theory, 2022, 4313-4361.
[29] Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. 2022. Learning General Halfspaces with Adversarial Label Noise via Online Gradient Descent. In ICML 2022, PMLR, 5118–5141.
[30] Ilias Diakonikolas, Jong Ho Park, and Christos Tzamos. 2021. ReLU Regression with Massart Noise. In Advances in Neural Information Processing 2021, 25891–25903.

Papers / Journal Articles / Conference Presentations or Proceedings

Mousavi Kalan, M.,Fabian, Z.,Avestimehr, S.,Soltanolkotabi, M., “Minimax Lower Bounds for Transfer Learning with Linear and One-hidden Layer Neural Networks”, *Advances in Neural Information Processing Systems,* **33**, 2020

Mousavi Kalan, M., Soltanolkotabi, M., Avestimehr, S., “Statistical Minimax Lower Bounds for Transfer Learning in Linear Binary Classification”, *IEEE International Symposium on Information Theory,* 2022, 282-287

Mushtaq, Erum, Yavuz Faruk Bakman, Jie Ding, and Salman Avestimehr, "Federated Alternate Training (FAT): Leveraging Unannotated Data Silos in Federated Segmentation for Medical Imaging." *International Symposium on Biomedical Imaging (ISBI),* 2023.

Mushtaq, Erum, Jie Ding, and Salman Avestimehr. "What If Kidney Tumor Segmentation Challenge (KiTS19) Never Happened." *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 1740-1747.

URLs

Mousavi Kalan, M.,Fabian, Z.,Avestimehr, S.,Soltanolkotabi, M., “Minimax Lower Bounds for Transfer Learning with Linear and One-hidden Layer Neural Networks”, URL: https://proceedings.neurips.cc/paper/2020/file/151d21647527d1079781ba6ae6571ffd-Paper.pdf

Mousavi Kalan, M., Soltanolkotabi, M., Avestimehr, S., “Statistical Minimax Lower Bounds for Transfer Learning in Linear Binary Classification”, URL: https://ieeexplore.ieee.org/abstract/document/9834760

Mushtaq, Erum, Yavuz Faruk Bakman, Jie Ding, and Salman Avestimehr, "Federated Alternate Training (FAT): Leveraging Unannotated Data Silos in Federated Segmentation for Medical Imaging." *International Symposium on Biomedical Imaging (ISBI),* 2023. URL: https://arxiv.org/abs/2304.09327

Mushtaq, Erum, Jie Ding, and Salman Avestimehr. "What If Kidney Tumor Segmentation Challenge (KiTS19) Never Happened." *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022, pp. 1740-1747. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10069820

Lists of Symbols, Abbreviations and Acronyms

CL	Centralized Learning
EMA	Exponential Moving Average
FAT	Federated Alternate Training
FETS2021	Federated Tumor Segmentation Challenge 2021
FL	Federated Learning
IID	Independent and Identically Distributed
KiTS19	Kidney Tumor Segmentation Challenge 2019
KNN	k Nearest Neighbor
L2	Least Square Error Loss
LLM	Large Language Model
LWE	Learning with Errors
LWLL	Learning with Less Labeling
MGM	Manifold Graph Metrics
ML	Machine Learning
NNK	Non-Negative Kernel Regression
PI	Principle Investigator
ReLU	Rectified Linear Unit
SOTA	State-of-the-art
SQ	Statistical Query
SSL	Self-Supervised Learning
UCF101	Human Actions Dataset
VC	Vapnik–Chervonenkis Dimension