Thesis for The Degree of Doctor of Philosophy

Adaptation under Distributional Shifts in Centralized and Federated settings

Adam Breitholtz

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2025

Adaptation under Distributional Shifts in Centralized and Federated settings

Adam Breitholtz

© Adam Breitholtz, 2025
except where otherwise stated.
All rights reserved.

ISBN 978-91-8103-349-6
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5806.
ISSN 0346-718X

Department of Computer Science and Engineering
Division of Data Science and AI
Healthy AI Lab
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

"You can't cross the sea merely by standing and staring at the water."
- Rabindranath Tagore

Adaptation under Distributional Shifts in Centralized and Federated settings

Adam Breitholtz

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Using large datasets to train ever larger models in machine learning have been extremely impactful for many diverse tasks. However, the challenge of ensuring and predicting model generalization under distribution shifts remains an open problem. Such shifts may occur between training and testing environments or even during the training process itself. In real-world applications, these distribution changes can severely degrade model performance, making adaptation a critical concern. This is the focus of domain adaptation (DA), a field dedicated to developing both theoretical frameworks and methods for settings with distribution shift. Domain adaptation primarily operates within the supervised learning paradigm, where access to a large, centralized dataset is assumed. However, such data availability is not always feasible due to privacy concerns or the high costs associated with data collection and storage. The federated learning (FL) setting addresses this by training models across decentralized clients coordinated by a central server. Since clients retain local data, distribution shifts, known as data heterogeneity, can arise between clients. This may potentially degrade model performance. This thesis aims to overcome some of these limitations in both the centralized and federated settings. In particular, this is achieved by (i) questioning how to measure performance under distribution shift in a practical way, (ii) proposing novel assumptions and settings where we expand the amount of information available and (iii) developing competitive methods for these settings. First, we explore the measurement of performance in domain adaptation through evaluating theoretical bounds. We survey the field of available domain adaptation bounds with an eye towards their practicality and, after selecting candidates, make empirical comparisons. Next, we consider a novel set of assumptions based on having access to privileged information which we show is both practical and empirically sound. We continue with expanding on the idea of additional information in the FL setting where we show that access to label marginals can substantially improve performance in cases where clients are meaningfully heterogeneous. Finally, we explore another aspect of heterogeneity in FL where the label sets of clients are non-identical and clients are unwilling to share them.

Keywords

Distribution shift, Domain Adaptation, Privileged Information, Federated Learning, Heterogenous clients

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **A. Breitholtz**, F. D. Johansson, *Practicality of generalization guarantees for unsupervised domain adaptation with neural networks*
Transactions on Machine Learning Research (October 2022)
Also presented at The AAAI-22 Workshop on Engineering Dependable and Secure Machine Learning Systems (March, 2022).

[**Paper II**] **A. Breitholtz**, A. Matsson, F. D. Johansson, *Unsupervised Domain Adaptation by Learning using Privileged Information*
Transactions on Machine Learning Research (September 2024)
Also presented at The Second Workshop on Spurious Correlations, Invariance and Stability, ICML (July, 2023).

[**Paper III**] E. Listo Zec, **A. Breitholtz**, F. D. Johansson, *Overcoming label shift in target-aware federated learning*
Submitted, under review. arXiv:2411.03799
Also presented at Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM), ICML (July, 2025).

[**Paper IV**] **A. Breitholtz**, E. Listo Zec, F. D. Johansson, *Federated Learning with Heterogeneous and Private Label Sets*
Springer Workshop Proceedings of ECML-PKDD 2025
Presented at 3rd Workshop on Advancements in Federated Learning (WAFL), ECML-PKDD (September, 2025).

Contribution summary

The author's contributions to the publications included in this thesis is detailed below.

[**Paper I**] Co-designed the study, performed the empirical work, and wrote most of the manuscript.

[**Paper II**] Co-designed the study, performed part of the empirical work, and wrote parts of the manuscript. The first two authors contributed equally.

[**Paper III**] Co-designed the study, performed part of the empirical work, and wrote parts of the manuscript. The first two authors contributed equally.

[**Paper IV**] Co-designed the study, performed part of the empirical work, and wrote parts of the manuscript. The first two authors contributed equally.

Acknowledgment

I would like to start by expressing my deepest gratitude towards my supervisor Fredrik Johansson. Without your seemingly endless patience, optimism, enduring support and belief in our research; this thesis would not exist. Thank you for everything.

I want to also thank my co-supervisor, Devdatt Dubhashi, and my examiner, Dag Wedelin for their feedback, support and encouragement throughout this process.

To everyone working at the DSAI division, I thank you for contributing to making the division a great place to be. Over the years, I have had the privilege of getting to know many great people ranging from professors to the administrative staff. In particular, I want to extend a special thanks the original members of the Healthy AI Lab Anton, Lena and Newton. I wish you all the best going forward. I also want to thank my collaborators Anton and Edvin. Thank you both for your great contributions to our work.

Also, I want to thank my friends that have put up with me throughout this journey. It has been an arduous march and I hope that in future I will be able to give back to you for the grace you have given me during this time. I want to particularly thank Samuel Håkansson for taking the time to proof-read a draft of this thesis.

Finally, I want to thank my family for their support. Without you I would not be writing these words here at all. A special thank you to my girlfriend Nathalie for her love and support when I needed it most.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

Abstract iii

List of Publications v

Contribution summary vii

Acknowledgement ix

I Overview 1

1 Introduction 3
1.1 Thesis contributions 5
1.2 Thesis outline 6

2 Background 7
2.1 Distributional shifts 8
2.1.1 Distributional discrepancy metrics 10
2.2 Unsupervised Domain adaptation 11
2.2.1 Multiple source domain adaptation and federated learning 12
2.2.2 Heterogeneity in federated learning 14
2.3 Guarantees and generalization bounds 15
2.3.1 The PAC learning framework 16
2.3.2 The PAC-Bayes framework 18
2.3.3 Data-dependent priors in PAC-Bayes 18
2.3.4 Bounds in MSDA and federated learning 19
2.4 Assumptions and their impact 19
2.5 Privileged information 22

3 Summary of Included Papers 25
3.1 Paper I - Practicality of generalization guarantees for unsupervised domain adaptation with neural networks 26
3.2 Paper II - Unsupervised domain adaptation by learning using privileged information 27
3.3 Paper III - Overcoming label shift in target-aware federated learning 30

3.4 Paper IV - Federated learning with heterogenous and private label sets 32

4 Concluding remarks and future directions **35**

Bibliography **39**

II Appended Publications **53**

Paper I - Practicality of generalization guarantees for unsupervised domain adaptation with neural networks

Paper II - Unsupervised Domain Adaptation by Learning using Privileged Information

Paper III - Overcoming label shift in target-aware federated learning

Paper IV - Federated Learning with Heterogeneous and Private Label Sets

Part I

Overview

Chapter 1

Introduction

Machine learning (ML) has emerged as both a popular and useful way to harness information for beneficial aims in diverse applications such as climate modeling, predicting healthcare outcomes and natural language processing. A central trend in ML is the focus on deep learning techniques, which has contributed to the explosive growth of the field. These models allow for learning intricate patterns from high-dimensional data sets through optimization methods like stochastic gradient descent. This has led to impressive results, such as large language models generating human-like text (Bubeck et al., 2023; DeepSeek-AI et al., 2025), mastering games like Go (Silver et al., 2017), and predicting medical outcomes (Ding et al., 2024). However, deep learning comes with significant limitations. It generally needs access to a large sample of data to be successful, which may not be available in all situations. In particular, sensitive settings such as healthcare may lack the large data sets which are routinely used when optimizing deep learning models. Furthermore, if the data differs materially between modeling and the target domain where it is applied, the models may fail to perform. Moreover, this can also arise *during* training where there are several distinct sources of data, e.g., in distributed training environments. This issue arises due to a shift between data distributions, which limits the generalizability of deep learning models.

To remedy this, the field of research called domain adaptation (DA) aims to offer a theoretical foundation of how to produce more generalizable methods. In DA we consider the case where we apply our model to a specific task but the underlying data distribution changes between training and deployment. The distribution over the input features and corresponding labels is called a *domain*. In real-world applications, a domain shift could be the result of e.g. a change in location where the data is collected. For example, we collect wildlife image data in Europe to predict animal species but apply the model in Africa, where the underlying data distribution is likely different. We generally refer to the data we train on as the *source* domain and the *target* domain for testing data. One of the main challenges in DA is that the underlying distribution of samples is not known, and thus a data set may be distributed differently than a test set due to several factors e.g. a limitation in the number of samples,

biased sampling or external factors changing the sampled data over time.

In scenarios where we observe distribution shift there may be further constraints that arise naturally, an important case being that information may be unavailable. An example of this is in the healthcare setting when we are trying to classify pathologies using chest X-ray images. We can train a model from historical registry data collected at Hospital A. Naturally, we may then want to apply this model at Hospital B. However, the patient cohort at Hospital B can be substantially different than the one at A. In this setting it is common that the labels corresponding to some features might be hard or even impossible to access, e.g. predicting outcomes which have yet to occur. For example, we probably do not know if a patient will develop COPD within a year from when features are collected. Therefore we can assume that we have access to features (chest X-rays) from hospital B but not the corresponding labels. This specific setting is called the *unsupervised domain adaptation* (UDA) setting, where labels are unavailable for the target domain. This setting is common in healthcare and other domains where labeling is expensive, time-consuming, or impossible.

Foundational theoretical work in UDA, such as the seminal work by Ben-David et al. (2007), and subsequent contributions (Cortes et al., 2010; Mansour et al., 2009b), aim to quantify how distribution shifts affect model performance. Since there is a lack of labeled information it does not allow us to adjust our model based on this knowledge. Furthermore, it is known that we may find guarantees of consistent learning if we do have this information (Blitzer et al., 2008). Moreover, adaptation in this setting is impossible without additional assumptions. (Ben-David & Urner, 2012) However, despite many theoretical and methodological advances, the theory of UDA often relies on assumptions which are implausible in real-world application or metrics which are impossible to quantify with the observed data. Furthermore, it is not clear what amount of information to assume access to for realistic and useful generalization guarantees. Moreover, there still remains a gap in performance compared to models that has been given access to target labels, which is not fully explained theoretically.

Distributional shifts and their impacts *during* training arises in DA when we consider multiple sources which contain their own distinct datasets. These multiple sources may give rise to heterogeneity in the classifiers learned on these disparate data sets and has been studied extensively (Mansour et al., 2008; Ben-David et al., 2010a; Hoffman et al., 2018; Peng et al., 2019).

An example of this is the *Federated* learning (FL) setting which also contains multiple sources of information with an additional constraint on communication. Federated learning is a form of decentralized learning where several clients collaborate, coordinated by a central server, to build a stronger model than they could individually. This training is based on iteratively transmitting model parameters or gradient updates to the central server, which then aggregates these into one central model which is transmitted back to the clients. This setup preserves privacy but introduces challenges due to data heterogeneity across clients.

Originally proposed as a method for distributed training under limited computational and bandwidth resources (McMahan et al., 2017), FL has

evolved into a broader paradigm focused on training models without sharing data across clients, maintaining privacy while still benefiting from all client datasets. FL thus embodies a trade-off between performance and privacy, with some works emphasizing privacy preservation, and others focusing on maximizing performance. However, distribution shifts among clients remain a key open challenge (Kairouz et al., 2021). The heterogeneity these shifts produce can degrade model performance and adversely effect convergence.

There are myriad approaches to handling different types of heterogeneity in FL based on adjusting either server-side (Reddi et al., 2021; Zeng et al., 2023; Li et al., 2023; Zhou et al., 2023) or on the client-side (Li et al., 2020; 2021; Chen et al., 2018; Jiang et al., 2019; Fallah et al., 2020; Li & Zhan, 2021) to account for the heterogeneity. However, since we generally do not know the particulars of all datasets at neither the server nor the clients we cannot make use of the theoretical approaches found in DA. Therefore, there is a need to find methods and assumptions which allow us to learn useful models in FL, even under distribution shift.

In this thesis we will investigate the issues detailed above in several ways, both in the centralized UDA and federated learning settings. We consider the following questions (i) how to predict performance under distribution shift in a practical way, (ii) what benefits are gained by introducing settings where we expand the amount of information available during training and (iii) how to develop competitive methods for these settings.

First, we explore the measurement of performance in unsupervised domain adaptation by evaluating theoretical bounds. We survey the field of available generalization bounds with an particular focus towards their practicality and, after selecting candidates, make empirical comparisons. Next, we consider a novel set of assumptions for UDA based on the having access to auxiliary privileged information which we show is both practical and empirically well-performing. We continue with expanding on the idea of additional information in the FL setting where we show that access to label marginals can substantially improve performance in cases where clients are heterogeneous with respect to their label distributions. Finally, we explore another aspect of heterogeneity in FL where the label sets of clients are non-identical and clients are unwilling to share them for privacy reasons. We propose adaptations of common FL methods and find that they incur only a small performance cost for the increased privacy. We also adapt methods from classifier coupling to align client models which achieves some promising results.

1.1 Thesis contributions

The main contributions of this thesis is summarized as follows:

- Paper I (Breitholtz & Johansson, 2022) explores whether it is possible to use the available bounds in the UDA literature for the explicit purpose of accurate performance prediction. We postulate a set of desiderata which such a bound should fulfill: that it should be tight; i.e. close to the observed performance of our model, possible to estimate from observed

data and tractably computable for realistic model classes and datasets. We survey the field of available domain adaptation bounds with regards to these desiderata and make empirical comparisons of how they perform on these three attributes. We find that PAC-Bayesian style bounds are a good fit for this and show two new such bounds.

- Paper II (Breitholtz et al., 2024) proposes a novel set of assumptions based on the concept of *privileged information* (PI), information which is auxiliary and only available during training. We argue that this set is more plausible than the more common set of assumptions used in the literature. Further, we show that these assumptions ensure consistent learning and provide a generalization bound. We then construct two model architectures, one which more closely follows our theory and a more practical end-to-end method. Extensive empirical investigation, across several datasets, show that these methods perform well compared to baselines which does not make use of PI.

- Paper III (Zec et al., 2025) investigates a similar concept to Paper II in the federated learning setting with heterogenous clients. It introduces FedPALS, a method which makes use of label marginals from clients and the target domain. It seeks to find a trade-off between having the aggregation of models be faithful to the target distribution and maintaining maximal sample efficiency. We show that our method resolves to the common federated averaging algorithm when focus on sample efficiency dominates. Empirically, we find that FedPALS performs very well with substantial performance gains compared to baselines which lack access to label marginals.

- Paper IV (Breitholtz et al., 2025) considers a more private FL setting with increased heterogeneity compared to Paper III. We consider that label sets are non-identical across clients. In addition, the clients are unwilling to share their label sets for privacy reasons; we call this the private label set setting. We adapt common FL methods to this setting as well as a tuning approach inspired by classifier coupling literature. We find that these methods perform quite well in this restricted setting with small performance impacts compared to the non-private setting.

1.2 Thesis outline

This thesis is outlined as follows: In Chapter 2 we detail background of distributional shifts, domain adaptation, federated learning and other concepts which are used in the appended papers.

Chapter 3 contain summaries of the appended papers. Chapter 4 contains a summary of the main contributions of the thesis. Further, it discusses some limitations of the methods and theory presented in the thesis and outlines several directions for future work.

Chapter 2

Background

Modeling relationships existent in large sets of data is an extremely powerful tool. This is the main objective in supervised machine learning (ML) where we aim to model a mechanism, h, which produces some desired answer Y, given an input X. These inputs are observed through a collection of samples $\{x_i, y_i\}_{i=1}^n$ which is called the *training set*. From this training set we wish to model the probability distribution $P(Y|X)$ using the data. To measure the models performance, it is subsequently applied to some unseen dataset; which we call a *test set*. This framework is widely applicable to a huge swathe of problems, and has been used to great effect in many fields such as the healthcare sector and several other industrial applications.

We formulate the core problem as a risk minimization where the quantity we seek to minimize is the *expected risk*. The risk denotes how much erroneous behavior we expect our model to have over some data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, as measured by a *loss function*, ℓ. The loss function is specified such that it measures discrepancy between the model prediction $h(X)$ and the desired outcome, Y.

The problem we solve in supervised ML can thus be stated as follows

$$\min_{h \in \mathcal{H}} R(h), \quad R(h) := \mathop{\mathbb{E}}_{(X,Y)\in\mathcal{D}} \ell(h(X), Y) \quad . \tag{2.1}$$

In general, the distribution $\mathcal{D}$ is unknown and as such we need to find a way to estimate the risk using known quantities. The general approach to this is to use *empirical risk minimization* (Vapnik, 1991). This is simply to use the samples we have observed, that are drawn from $\mathcal{D}$, and minimize the loss based on these.

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \tag{2.2}$$

The result of this optimization problem will be a model from the family of models, $\mathcal{H}$, that produces the lowest empirical risk. However, it is not guaranteed that this will be a unique solution or that this is the model which will generalize well on unseen data. To be able to say something regarding this we need to have additional assumptions on the model or the data which we

observe. A fundamental assumption which is central to the success of ML is the assumption that the data points that one is trying to model are similar to the ones that the model will encounter when applied. This assumption states that the data points are independent and identically distributed (i.i.d.) according to the same distribution $\mathcal{D}$ for both the training and test sets. This assumption is in general quite unrealistic as in real-world applications it is quite common for the distribution of data to undergo shifts of some sort. This can come up either through intentional action or inadvertently. Intentionally changing geographical location to apply models, changes in testing protocols or equipment variation are examples of the former. Unintentional shifts may arise due to time passing e.g. seasonal variation or the test population evolving. The kinds of shifts that are possible and the effects that they can have is what we will consider next.

2.1 Distributional shifts

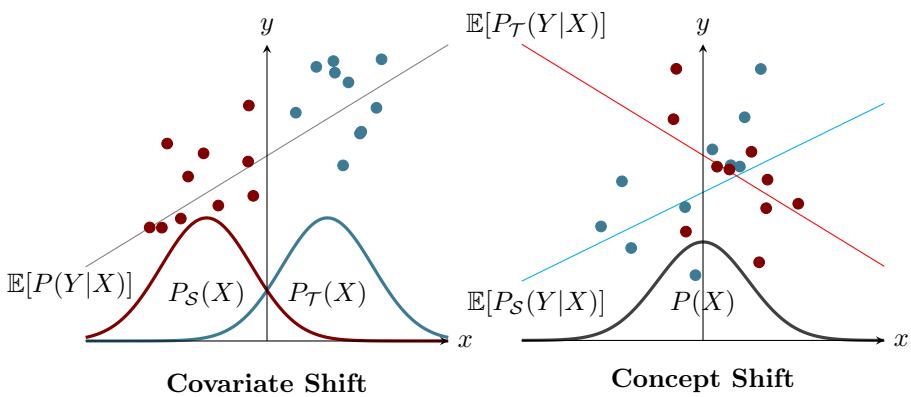


Figure 2.1: Illustration of covariate shift and concept shift when the underlying functions are linear. Label shift and concept drift can be illustrated similarly but with X taken as the dependent variable, instead of Y. Domain shift is when both shifts occur simultaneously.

In cases where the i.i.d. assumption does not hold there is some shift between the distributions that the training set and test set are drawn from. This can be due to several different types of shift. To detail the types of shift that are possible, we should first consider the decomposition of the joint probability distribution,

$$P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y) \ . \tag{2.3}$$

From this expression we can detail shifts which occur in any combination of the components of the joint distribution between the two distributions. We call the distributions $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$. Some shifts are illustrated in Figure 2.1

1. **Covariate shift**: A shift where the mechanism $P(Y|X)$ is assumed identical across domains and the distribution over features $P(X)$ varies. For example, if our task is to classify animals in images but the training and test sets are taken in different locations. If the background content of images change between the data sets but the images still depict the same types of animals, then there has been covariate shift.

2. **Label shift**: Here we assume that $P(Y)$ can vary but the $P(X|Y)$ is the same across distributions. Consider the example of sales statistics where labels are product categories. Label shift means, in this context, that the proportion of sales across product categories varies between two different retailers and the target, but that the pattern of customers who purchase items in each category ($P(X|Y)$) remain consistent.

3. **Domain shift**: This denotes a more general shift where both the conditional term, $P(Y|X)$, and marginal term, $P(X)$, can vary. Of course this is very challenging and could be impossible to model, absent other relationships between distributions. (Ben-David et al., 2010b;c) This can be the case when we model a problem where both the labeling mechanism and input distributions change. E.g. interpretation of liability from legal texts between countries. The texts may be both written in a different way ($P(X)$ changes), and their interpretation is different depending on countries legal precedent ($P(Y|X)$ changes).

4. **Concept shift**: This refers to when the function which maps inputs to outputs, $P(Y|X)$, change but the input distribution stays the same. For example, in the diagnosis of medical conditions, a set of patient features (e.g. symptoms) may indicate different diseases (labels) in different populations.

5. **Concept drift**: Similar to the shift above, with the change occurring in the mapping $P(X|Y)$ while $P(Y)$ is constant. This can be the due to a shift in appearance of labeled data caused by e.g. seasonal variation. An example is store purchase volumes increasing in some parts of the year than others, which could be driven by (possibly unmeasured) factors such as sales campaigns or stronger consumers due to tax refunds.

As we can see in the list above, there are many different possible shifts in distribution and some are more common than others. The fact remains that to solve the ML problem effectively under a distributional shift we need to consider strategies to deal with the shift. However, we first need to ask the question of how to quantify the existence and extent of a shift.

Measuring differences between distributions is not a perfectly well-defined concept, and many different metrics can be constructed. One may measure discrepancy between distributions in a specific way, and this measure may be more or less useful depending on the data and assumptions that are being made. We will detail some common choices of such metrics next.

2.1.1 Distributional discrepancy metrics

The notion of having a measure of the difference between distributions is not novel. It hails from the statistical problem to discriminate between two populations (Welch, 1939; Brown, 1950), in particular where one makes limited assumptions about their properties. Early works such as the techniques of Fisher (1936) treated this for distributions which are partly known, and more generally using the notion of hypothesis testing for unknown distributions (Neyman & Pearson, 1933).

We can define a specific metric which provides a notion of distance between distributions in many ways. However, they can be sorted into two main categories f-divergences and integral probability metrics. The f-divergences were first introduced in Rényi (1961) and the f-divergence of P from Q can be written as

$$D_f(P \parallel Q) = \int_\Omega f\left(\frac{dP}{dQ}\right) dQ, \tag{2.4}$$

where Ω is some space and f a convex function. An early example of these divergences is that of Kullback & Leibler (1951) which generalized a notion from information theory to form the Kullback-Liebler (KL) divergence. The KL-divergence from a distribution Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{x\in\mathcal{X}} \log \frac{P(dx)}{Q(dx)}\, P(dx), \tag{2.5}$$

for some measurable space $\mathcal{X}$. I.e., it is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P. One should note that this is not a metric as it is not symmetric, nor is it bounded.

Integral probability metrics (IPMs) is the second type of distances between probability distributions. Initially termed by Müller (1997), IPMs are distances on the space of distributions over a set $\mathcal{X}$ defined by a class $\mathcal{F}$ of real-valued functions on the same set. We define them as

$$D_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)\right| = \sup_{f\in\mathcal{F}} \left|Pf - Qf\right|. \tag{2.6}$$

A recent example of IPMs is one called the maximum mean discrepancy (MMD). This IPM was used to discriminate based on the moments of the distributions using a kernel function (Gretton et al., 2012). The MMD between two probability distributions P and Q is defined as

$$\mathrm{MMD}(\mathcal{F},P,Q) = \sup_{f\in\mathcal{F}} (\mathop{\mathbb{E}}_{x\sim Q}[f(x)] - \mathop{\mathbb{E}}_{y\sim P}[f(y)]), \tag{2.7}$$

for some kernel function class $\mathcal{F}$ from a reproducible kernel hilbert space.

Importantly, there is a metric which resides in the intersection between these two classes of metrics, which is the Total variation metric (Jordan, 1881). It is defined on a measurable space $(\Omega, \mathcal{F})$ as

$$\mathrm{TV}(P,Q) = \sup_{A\in\mathcal{F}} |P(A) - Q(A)| \ .$$

I.e., the metric yields the largest absolute difference between the probabilities that the two distributions assign to the same event, A. One can note that the TV metric is quite a pessimistic comparison of distributions. As an example, consider two distributions which agree completely on the probability of all events except one where they disagree with some amount Δ. Compare this to two distributions which do not agree for any events but never more than $\Delta - \epsilon$, which leads to the latter being assigned a lower metric value than the former. This highlights the need to choose metrics for comparison carefully depending upon the application and the notion of "closeness" one wishes to promote. Moreover, it is also salient to consider whether a metric can be accurately estimated from the available data during training. As we will see, this is relevant for achieving guarantees on performance under distribution shift which are not only theoretically correct, but also practically evaluable.

2.2 Unsupervised Domain adaptation

When we expect the i.i.d. assumption to be violated, we need to change our theoretical treatment to account for this. This is the focus in the field of unsupervised domain adaptation (UDA). The training samples that we have access to is assumed have been drawn from an underlying distribution $\mathcal{S}$. This distribution over the product space $\mathcal{X} \times \mathcal{Y}$ is referred to as a *domain*. In UDA, we assume that we have access to $(X, Y) \sim \mathcal{S}$ and $\tilde{X} \sim \mathcal{T}_X$; where $\mathcal{S}$ and $\mathcal{T}$ are called the source and target domains respectively and $\mathcal{T}_X$ is the marginal distribution of features in the target. These quantities will be observed through samples $S = \{x_i, y_i\}_{i=1}^n \sim (\mathcal{S})^n$ and $S'_x = \{\tilde{x}_i\}_{i=1}^m \sim (\mathcal{T}_x)^m$, where $(\mathcal{D})^N$ denotes the distribution of a sample of N datapoints drawn i.i.d. from the domain $\mathcal{D}$. The goal of UDA is to learn a model h, such that it minimizes the risk on data from the target domain. More formally, we write

$$\min_{h \in \mathcal{H}} R_{\mathcal{T}}(h), \quad R_{\mathcal{T}}(h) := \mathbb{E}_{X,Y \sim \mathcal{T}} \left[\ell(h(X), Y) \right]. \tag{2.8}$$

Note that we now have two distributions, $\mathcal{S}$ and $\mathcal{T}$, which are materially different from each other. We also do not observe full samples from the target domain, where we want to estimate model performance, which then precludes the use of ERM as in regular ML. We can think of the distributions as being somehow shifted relative to each other, and understanding this discrepancy is central to estimating $R_{\mathcal{T}}(h)$ well.

The UDA problem was initially considered in natural language processing (Hwa, 1999; Chelba & Acero, 2006; Blitzer et al., 2006) where it was observed to naturally arise. An early theoretical treatment based on maximum entropy models was done by Daumé III & Marcu (2006). The first general treatment is due to Ben-David et al. (2007) where the idea of a discrepancy metric between the source and target domains was introduced. Defining other such metrics has been a focus in several subsequent works in UDA such as e.g. Mansour et al. (2009b); Cortes & Mohri (2014). The proposal of different metrics, such as those detailed in 2.1.1, is motivated by the implicit assumptions made by their use and the fact that there are many ways to construct them.

In summary, the DA framework is a general framework which can describe the process of learning under distribution shift. DA can also lend itself fairly well to variations of different kinds; extending the setting to situations with varying amounts of information available. We will detail some of these extensions next.

2.2.1 Multiple source domain adaptation and federated learning

A natural extension of the UDA setting is to include the possibility of several different source domains. The motivation for such an approach would be that any one source domain would have a lesser probability of accurately representing the target domain. As such, the use of multiple domains would more likely lead to a higher likelihood of the target being covered in a distributional sense. This setting is called multi-source domain adaptation (MSDA) (Mansour et al., 2008; Ben-David et al., 2010a) and occurs in many real-world situations such as sentiment analysis and image classification where data originate from several different sources. Formally, MSDA assumes access to labeled data from n source domains and unlabeled data from one target domain.

In contrast to regular UDA we need to reconcile discrepancies not only between the sources and the target but also among the source domains themselves. This is commonly tackled with representational alignment of features, either by discrepancy based methods (Hoffman et al., 2018; Peng et al., 2019; Guo et al., 2018) or adversarial methods (Xu et al., 2018; Zhao et al., 2020). This type of problem generally require the use of large datasets to achieve high performance. However, to collect and store such datasets may be challenging for many reasons, privacy being a central one. Increased regulation of personal data collection as well as changing user preferences make privacy-preserving methods a valuable direction of study. In the machine learning context this has taken the form of the study of differential privacy (Dwork et al., 2006) and federated learning (McMahan et al., 2017) (FL) methods. Of course, there have been many works studying distributed machine learning before FL was introduced by McMahan et al. (2017). However, we will not provide a comprehensive survey of them in this thesis. Federated Learning is a decentralized machine learning paradigm where model training occurs across multiple clients (e.g., mobile devices, hospitals, organizations), each holding local data that remains private. I.e., the data is not transmitted outside the client to preserve the clients privacy but instead model parameters θ_i are transmitted. See Figure 2.2 for a schematic overview. A central server orchestrates the training process, which we will detail next.

FL is generally formulated as follows: assume that there are K clients, each with a dataset, $\mathcal{D}_i = \{x_k^i, y_k^i\}_{k=1}^{n_k}$, which only they have access to. For each client which takes part in the federation, they solve a local optimization problem by minimizing their risk locally

$$\min_{h \in \mathcal{H}} R_i(\theta_i), \quad R_i(\theta_i) = \mathbb{E}_{D_i}[\ell(h_i(\theta_i; x), y)] \ , \tag{2.9}$$

where θ_i are the local model parameters.

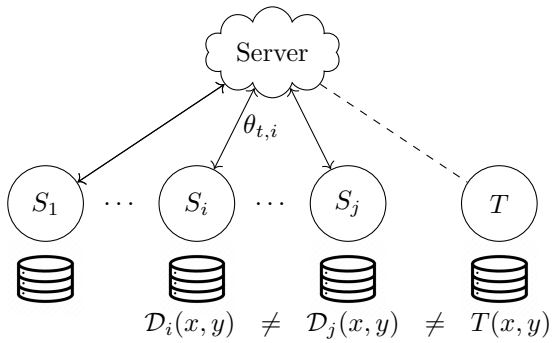


Figure 2.2: Schematic overview of federated learning (FL). Note the likeness with MSDA as each client can be viewed as its own source domain S_i where the datasets $\mathcal{D}_i$ are not necessarily distrubuted equally. The key difference between the two settings being the decentralized nature of FL, where only model parameters are shared with a central, coordinating server.

The larger objective of FL is then to train a single global model that minimizes the risk across all clients, i.e. minimizing

$$R_{fed} = \frac{1}{K} \sum_{i=1}^{K} R_i(\theta_i) \ . \tag{2.10}$$

One of the most widely used algorithms for solving the FL problem in (2.10) is federated averaging (FedAvg) introduced by McMahan et al. (2017). This algorithm first lets the clients take one or several local steps of stochastic gradient descent solving (2.9). Then the clients update their local models which are then transmitted to a central server. This server then aggregates the model parameters to create a global model. The aggregation is done by performing a weighted average of the local models, which is applied layer-wise. The weights in the aggregation is based on the sizes of the local datasets, which for some timestep t is calculated as

$$\theta^t = \sum_{i=1}^{M} \frac{n_i}{n} \theta_i^t \ , \tag{2.11}$$

where n is the total amount of samples in all clients and M is the amount of clients participating in the last round of updates. This is not necessarily equal to K as we allow for some fraction of clients to be inactive. This is motivated by applications with a very large amount of clients, where it is unlikely that all clients will participate at all times. We present an algorithmic overview of FedAvg in Algorithm 1.

The connection between MSDA and FL arises when we interpret the source domains in MSDA as clients in a federated system. In this view, each client represents a distinct data distribution (i.e., a source domain), and the system seeks to train a model that generalizes well to a new, unseen distribution (i.e.,

Algorithm 1: Federated Averaging

Data: Client set $\mathcal{S}$, learning rate η and local datasets D_i indices $\{I_k\}$
Server executes:
Initialize central parameters θ^0
for *each round* $t = 0, ..., T-1$ **do**
  Sample M clients from $\mathcal{S}$
  for *each client* $k = 1, ..., M$ **do**
    Distribute θ^t to client i
    Receive client update $\theta_i^{t+1} = \text{Client_Update}(i, \theta_t)$
  end
  $\theta^{t+1} = \sum_{i=1}^{M} \theta_i^t \frac{n_i}{n}$ where $n = \sum_{i=1}^{M} n_i$
end
return $h(\theta^{t+1})$

Function `Client_Update`(i, θ)**:**
  for *local step* $j = 1, ..., E$ **do**
    $\theta = \theta - \eta \nabla \ell(h(\theta; x), y)$ for $x, y \sim D_i$
  end
  return θ

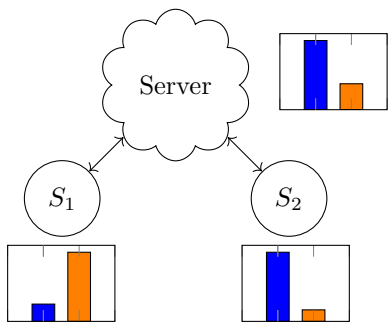


Figure 2.3: A simple example of client heterogeneity in terms of the label set distributions in clients being materially shifted w.r.t each other. Note that the aggregate of clients does not necessarily equal the server side test set distribution.

a target domain). Naturally, heterogeneity can arise in the FL clients due to distribution shifts between them, this is what we will treat next.

2.2.2 Heterogeneity in federated learning

As mentioned above, issues may arise with non-i.i.d. data across domains/clients in FL. When the client datasets differ significantly in distribution, we should not expect averaging of parameter to produce a better classifier than the individual models. Furthermore, in more adverse cases of heterogeneity, the training may

converge slower or the resulting classifier may perform worse than it would have in an i.i.d. setting. This is related to the fact that there often is an assumption that the target is some aggregate of the clients which in general may not be the case. Consider the example in Figure 2.3, where we have two clients with different label distributions over their respective datasets. Moreover, the aggregate of these clients does not conform to the distribution which is observed in the test set on the central server. This can prove challenging as the models may have differing notions of how to discriminate between the two classes and it is not clear whether aggregating the two would have a beneficial effect, especially if the aggregate of the clients is not identical to the test set. This would then lead to the problems we highlighted earlier, e.g. slow convergence or suboptimal performance.

While the problem remains open generally, there have been approaches which consider alternative objectives, e.g maximizing worst-group performance (Mohri et al., 2019). Other efforts to mitigate the effects of distributional shifts in federated learning can generally be categorized into client-side and server-side approaches. Client-side methods use techniques such as regularization techniques that penalize large deviations in client updates (Li et al., 2020; 2021), client clustering (Ghosh et al., 2020; Sattler et al., 2020; Vardhan et al., 2024) which makes models for each client cluster and meta-learning (Chen et al., 2018; Jiang et al., 2019; Fallah et al., 2020). Server-side methods focus on improving model aggregation or adjusting post-aggregation. These include optimizing aggregation weights (Reddi et al., 2021), increasing gradient diversity during updates (Zeng et al., 2023), learning adaptive aggregation weights (Li et al., 2023) and using iterative moving averages to refine the global model (Zhou et al., 2023). Another related area is personalized federated learning, which focuses on fine-tuning models to optimize performance on each client's specific local data (Collins et al., 2022; Boroujeni et al., 2024; Mclaughlin & Su, 2024). This setting, while interesting, will not be a focus in this thesis.

2.3 Guarantees and generalization bounds

In many settings, ensuring good performance of a model is critical to successful deployment. High-stakes settings have this attribute, e.g. autonomous driving and making treatment decision in healthcare settings. If the aim is to guarantee a specific level of model performance we need a bound on the target risk as specified in the previous section. Simply showing acceptable performance on held-out datasets is not a guarantee that the performance will not degrade when applied in other settings. Such a degradation in performance has been observed in e.g. the healthcare setting (Zech et al., 2018), and natural language processing (Jia & Liang, 2017).

As the quantity in (2.8) is written as an expectation over the a priori unknown distribution $\mathcal{T}$ we will have to estimate the risk somehow. Therefore, we can substitute the expectation by computing an approximation of this using

the sample average which we write as

$$\hat{R}_{\mathcal{T}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\tilde{x}_i), \tilde{y}_i) \tag{2.12}$$

for some sample $\{\tilde{x}_i, \tilde{y}_i\}_{i=1}^n \sim (\mathcal{T})^n$. However, as we assume that we do not have access to target labels we have to estimate the target risk with something that we actually can calculate. Therefore, to deal with this complication we use theory to connect the expected target risk to the expected source risk.

Further, we assumed that the distributions of the source and target data were different from each other. Therefore, we need to account for the discrepancy between the two distributions. This can be done in several different ways; to illustrate, we show a simple example of what we want to achieve. We want some way to bound the target risk with quantities which we have some hope of estimating with the available information, i.e. we want to find a bound on the following form:

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \mathcal{D}(\mathcal{S}, \mathcal{T}) \ .$$

The last term, $\mathcal{D}(\mathcal{S}, \mathcal{T})$, is measuring some distance/discrepancy between the source and target domains. We call this discrepancy the *domain shift* term which figures in some form in all domain adaptation bounds. Of course, we have to ensure that the domain shift term does not depend on knowledge of the target distribution.

After this step, we wish to bound the expected source risk with a sample average like in (2.12). However, we still need to account for the error between the expected and empirical risk. This means that a sample generalization term must be added, i.e. a term accounting for the error which we see due to our sample size being limited. One approach to estimating the sample generalization error of a classifier is to use statistical learning theory which we will detail in the next section.

Thus, if we have the tools to both relate the source risk to the target risk and connect quantities in expectation to their empirical counterparts, we can express generalization bounds on the following form:

$$R_{\mathcal{T}} \leq f(\text{Empirical source risk, Domain shift, Sample generalization error}).$$

The specific form of f and the terms it depends on is decided by the theoretical approach taken to the steps detailed above. However, the two main forms are whether the domain shift and sample generalization terms are related through addition or multiplication. We will now go into the main theoretical tools used to solve the two challenges detailed above.

2.3.1 The PAC learning framework

The most prevalent theoretical framework for reasoning about the generalization performance of non-deterministic statistical models is statistical learning theory, more specifically, a theory called Probably Approximately Correct (PAC)

learning (Valiant, 1984). This theory allows us to through assumptions on the model, task and data show that a certain task is learnable, specifically as understood through the PAC lens.

This means is that for a certain task it can be shown to be PAC learnable if we can show that given an algorithm $\mathcal{A}$ and a sample of size n, the algorithm $\mathcal{A}$ returns a model from the model class, $\mathcal{H}$, which has a small average error, ϵ, with high probability, $1-\delta$. This then amounts to that we can show that the risk for a specific model on the given data is smaller than some value $\epsilon > 0$ with confidence level $1-\delta$, where $\delta < 1$, or more formally,

$$\Pr[\mathbb{E}[\ell(h(x), y)] \leq \epsilon] \geq 1-\delta, \ \forall h \in \mathcal{H}. \tag{2.13}$$

With a formulation on this form we can then use some well known results, often based on concentration inequalities, such as e.g. Hoeffding's inequality to move from an expectation form to an empirical form. This is due to the inequality providing an upper bound on the probability that the loss deviates from its expected value by more than a certain amount. Using these kinds of techniques we can, using application of standard theory (Vapnik, 1998), get bounds like the following. For an i.i.d. sample of size m we have that the following holds with probability at least $1-\delta$ for every $h \in \mathcal{H}$:

$$R_{\mathcal{S}}(h) \leq \hat{R}_{\mathcal{S}}(h) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)}. \tag{2.14}$$

The quantity d in the above expression is the so-called Vapnik-Chervonenkis (VC) dimension. The VC dimension is a measure of how complex the family of functions, $\mathcal{H}$, are. An early example of a bound on the target risk that is achieved using this framework is the following one from Ben-David et al. (2007)

$$R_{\mathcal{T}}(h) \leq \underbrace{\hat{R}_{\mathcal{S}}(h)}_{\substack{\text{Empirical}\\ \text{risk}}} + \underbrace{\sqrt{\frac{4(d \log \frac{2em}{d} + \log \frac{4}{\delta})}{m}}}_{\text{Sample generalization}} + \underbrace{d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda}_{\text{Domain shift}}, \tag{2.15}$$

where d is the VC dimension of the $\mathcal{H}$, λ is the sum of the errors on both domains of the best performing classifier $h^* = \arg\min_{h \in \mathcal{H}}(R_{\mathcal{S}}(h) + R_{\mathcal{T}}(h))$, and $d_{\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{A \in \{\{x : h(x) = 1\} : h \in \mathcal{H}\}} |\Pr_{\mathcal{S}}[A] - \Pr_{\mathcal{T}}[A]|$ is the $\mathcal{A}$-distance for the characteristic sets of hypotheses in $\mathcal{H}$. There are many subsequent works which use this approach also use the PAC style with different choices of discrepancy terms and methods of bounding sample generalization. (Cortes et al., 2015; Acuna et al., 2021; Liu et al., 2025) Using the PAC approach we get a bound which holds uniformly over the class of hypotheses $\mathcal{H}$. This is one of the features of PAC learning, the bounds hold for all classifiers in the considered class. However, this can also be a weakness as this produces bounds which must, by definition, hold for the worst classifier imaginable from the class. Depending on the richness of the class this can be arbitrarily limiting. In response to this issue, there is an extension to the PAC framework which does not suffer the same fate which we will detail next.

2.3.2 The PAC-Bayes framework

PAC-Bayes theory is an extension of PAC theory based on using the PAC framework to understand Bayesian classifiers. This way of analyzing classifiers was initially proposed by Shawe-Taylor & Williamson (1997), with the first generalization guarantee being proved by McAllester (1998). The framework studies generalization of a posterior distribution ρ over hypotheses in $\mathcal{H}$, learned from data, in the context of a prior distribution over hypotheses, π. The generalization error in ρ may be bounded using a divergence between ρ and π as seen in the following classical result due to McAllester (2013).

For a prior π and posterior ρ on $\mathcal{H}$, a bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ and any fixed $\gamma, \delta \in (0, 1)$, we have w.p. at least $1 - \delta$ over the draw of m samples from $\mathcal{D}$, with $\mathrm{KL}(p\|q)$ denoting the Kullback-Liebler (KL) divergence between p and q,

$$\mathop{\mathbb{E}}_{h\sim\rho} R_{\mathcal{D}}(h) \leq \frac{1}{\gamma} \mathop{\mathbb{E}}_{h\sim\rho} \hat{R}_{\mathcal{D}}(h) + \frac{\mathrm{KL}(\rho\|\pi) + \ln(\frac{1}{\delta})}{2\gamma(1-\gamma)m} \ . \tag{2.16}$$

As we can see in (2.16), we now have an expression which is stated as an expectation over the posterior distribution ρ. This is in addition to the expectation over the distribution of the data. Thus the bound holds, on average, for classifiers drawn from the posterior ρ. We can interpret ρ as a distribution over the parameters of our trained classifier. The posterior would then be a distribution around the classifier, effectively covering those classifiers which are close in parameter space to our trained classifier. A prominent feature of the framework then, is that we restrict our prediction to a smaller part of the model space. The additional complexity incurred by this is by adding an expectation over the distribution ρ, i.e, the models which are close to the learned classifier. We consider an additional expectation to be an expense, as it increases computational complexity to a prospective evaluation of the bound. There are some key things to note with this formulation that are advantageous if we want to estimate the quantities in the bound. First, the shape of the prior and posterior distributions are not explicitly stated and can be chosen at will; no matter the choice, the bound will still hold. Further, the posterior ρ is something which we learn from the training data. We will next detail another way in which the PAC-Bayes formulation is preferable when the aim is to achieve tighter bounds.

2.3.3 Data-dependent priors in PAC-Bayes

The sample complexity term in (2.16), based on the KL-divergence, grows as the prior π and posterior ρ become more dissimilar. This may happen if the posterior is very sensitive to the training data or the prior is poorly chosen. The posterior should be sensitive to the data in some respect, otherwise the training of it would be ineffective. To prevent a radical divergence between π and ρ we can inform our choice of prior with some of the data we have available. This is called a data-dependent prior and was developed by the work of Ambroladze

et al. (2007) and Parrado-Hernández et al. (2012), with an extension to neural networks by Dziugaite et al. (2021).

When we use this type of prior, given that enough data has been used to inform the prior, we will observe a tightening of the resulting bound. This will be due to the KL term being smaller since the prior and posterior are now closer to each other. It is important to note that any data which is used to learn a prior must be independent of the data used to evaluate the bound. If this is not ensured the bound will not hold. However, we should also note that this restriction does not affect which data is used to learn the posterior, ρ.

2.3.4 Bounds in MSDA and federated learning

The theoretical frameworks discussed previously can be used to analyze both MSDA and FL as well. For instance, Crammer et al. (2008) and Mansour et al. (2009a) both contribute theoretical PAC-style bounds based on Rademacher complexity and Renyi Divergence respectively. More recent approaches to bounding error in MSDA from Shui et al. (2021) and Chen & Marchand (2023) make use of some labeled target data to achieve guarantees on performance. The main difference between the bounds which we obtain in regular DA, and those in MSDA is that there usually is some term which relates the target risk to some combination of the source risks and also potentially the pairwise discrepancy between sources or between some combination of sources and target.

In federated learning there are similar lines of work which show bounds on performance. The communication constraints inherent to the setting does complicate matters, and treatments which take heterogeneity of clients into account is rare. The works of Mohri et al. (2019), Fallah et al. (2021), Chen et al. (2023) and Boroujeni et al. (2025) all present bounds in the FL setting where heterogeneity is considered. However, they all make different assumptions on the boundedness of the loss, convexity of the problem or on the level of client-wise heterogeneity. This highlights the complex nature of constructing theory in such a difficult and constrained setting, as well as the importance of assumptions. The impacts, trade-offs and limitations inherent to making assumptions on the tasks we want to solve is what we will discuss next.

2.4 Assumptions and their impact

Current theoretical approaches have produced general guarantees for performance, but these are limited in their ability to provide realistic guarantees on UDA and similar problems. It is known that general adaptation is impossible and thus the question then becomes: What do we need to assume to guarantee consistent learning?

Consistent learning implies that our model will learn to solve the task at hand in the limit of infinite samples and that we will do so every time. That is, as the sample size n increases the estimates converge in probability to the value that the estimator is designed to estimate. To ensure this, a large

swathe of works make assumptions that are quite similar. First, one makes some assumption regarding a conditional probability distribution related to either One can assume that the underlying function which generates outcomes is the same between the domains. This is called the covariate shift assumption (Shimodaira, 2000), meaning that the data is allowed to change, but not the function labeling the data. We write this as follows:

Assumption 1 (Covariate shift)**.** *For domains $\mathcal{S}, \mathcal{T}$ on $\mathcal{X} \times \mathcal{Y}$, we say that* covariate shift *holds with respect to $X \in supp(\mathcal{S}, \mathcal{T})$ if*

$$\exists x : \mathcal{T}(x) \neq \mathcal{S}(x) \;\; and \;\; \forall x : \mathcal{T}(Y \mid x) = \mathcal{S}(Y \mid x) \ ,$$

where $\text{supp}(\mathcal{S}, \mathcal{T})$ is the common support of the domains. This assumption is often made and can hold in many different settings, we often have little reason to believe that the underlying labeling function will change just because the domain has done so.

Another similar approach is what we call the *label shift* assumption. (Saerens et al., 2002; Lipton et al., 2018) This is similar in nature to covariate shift, but instead focused on the conditional $\mathcal{D}(X \mid Y)$. So here, in contrast to the former, the distribution of inputs given labels is the same across domains. It can be described as follows:

Assumption 2 (Label shift)**.** *For domains $\mathcal{S}, \mathcal{T}$ on $\mathcal{X} \times \mathcal{Y}$, we say that* label shift *holds with respect to $Y \in supp(\mathcal{S}, \mathcal{T})$ if*

$$\exists y : \mathcal{T}(y) \neq \mathcal{S}(y) \;\; and \;\; \forall x : \mathcal{T}(X \mid y) = \mathcal{S}(X \mid y) \ .$$

One might be tempted to think that any of the two assumptions above would be enough, however, this is unfortunately not the case. As shown in Ben-David et al. (2010b) we also need a assumption of coverage of the target domain to have a guarantee of consistent learning. Therefore, we also need the overlapping support assumption to be able to guarantee consistent learning.

Assumption 3 (Domain overlap)**.** *A domain $\mathcal{S}$ overlaps another domain $\mathcal{T}$ with respect to a variable Z on $\mathcal{Z}$ if*

$$\forall z \in \mathcal{Z} : \mathcal{T}(Z = z) > 0 \implies \mathcal{S}(Z = z) > 0 \ .$$

This assumption states that if some datapoint is possible to observe in the target domain we also have a non-zero probability to observe it in the source domain. As should be evident, this is quite a strong assumption that may not hold in realistic scenarios. We illustrate this phenomenon in figure 2.4.

To exemplify how assumptions result in limitations on theory we will present some examples from the literature. We start with the following bound due to Cortes et al. (2010)

$$R_{\mathcal{T}} \leq \hat{R}^{w}_{\mathcal{S}} + 2^{5/4} \sqrt{d_2(\mathcal{T} \| \mathcal{S})} \sqrt[3/8]{\frac{d \log \frac{2ne}{d} + \log \frac{4}{\delta}}{n}}.$$

This bound is an example of an importance weighting bound which bounds the target risk using a weighted empirical source risk, $\hat{R}^{w}_{\mathcal{S}}$. In this term we re-weight

the loss function according to the density ratio $w(x) = \frac{\mathcal{T}(x)}{\mathcal{S}(x)}$ of each sample. For this style of bound we run into issues when the overlap assumption does not hold. Consider the density ratio above; if there is a lack of overlap we may have a data point which only has non-zero density in the target domain. This leads to a division by zero in w and the bound immediately becomes vacuous. The issue is that it is very simple to violate overlap in practice, e.g. learning from black and white images and applying to color images. This inability to handle the non-overlapping case is a weakness we would like to avoid.

So we might come to the conclusion that we should avoid the importance weighting type bounds but still keep the assumptions. This often yields something akin to the bound we stated in (2.15). We will state a similar bound here due to Ben-David et al. (2010a):

$$R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + 4\sqrt{\frac{2(d \log 2m + \log \frac{2}{\delta})}{m}} + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda, \tag{2.17}$$

where

$$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2\left(1 - \min_{h,h' \in \mathcal{H}}\left[\frac{1}{m}\sum_{\substack{x \sim (\mathcal{T}_X)^m: \\ h(x) \neq h'(x)}} \mathbb{1}[x] - \frac{1}{m}\sum_{\substack{x \sim (\mathcal{S}_X)^m: \\ h(x) \neq h'(x)}} \mathbb{1}[x]\right]\right).$$

This bound has some qualities that we might take issue with; these are mainly related to the way domain shift is measured. First, the $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ term measures the discrepancy between how much two distinct hypotheses will disagree on the source and target. Intuitively this accounts for the difference between the source and target, but this quantity is not easy to calculate as it requires a minimization over the hypothesis class, $\mathcal{H}$. This is difficult to evaluate as the class can be very large, which is the case for neural network classifiers. This type of of quantity figures in many other works. (Ben-David et al., 2007; Blitzer et al., 2008; Ben-David et al., 2010a; Morvant et al., 2012; Mansour et al., 2009b; Redko et al., 2019; Cortes & Mohri, 2014; Cortes et al., 2015; Yin et al., 2024; Koc et al., 2025).

Second, the λ term, which accounts for the joint optimal error of the best classifier, is not possible to observe with the data available. So if we want a bound that is tight we have to assume that this quantity is small. This non-observable quantity or ones like it is quite common in the literature. (Kuroki et al., 2019; Redko, 2015; Long et al., 2015; Redko et al., 2017; Johansson et al., 2019; Zhang et al., 2019; Dhouib et al., 2020; Shen et al., 2018; Courty et al., 2017; Germain et al., 2013; Dhouib & Redko, 2018; Acuna et al., 2021; Nguyen et al., 2022; Huang et al., 2025; Koc et al., 2025; Liu et al., 2025) As we have seen in this section, the limitations of the current literature are not insubstantial. We can therefore see that there is a need to develop theoretical results which do not suffer from these limitations.

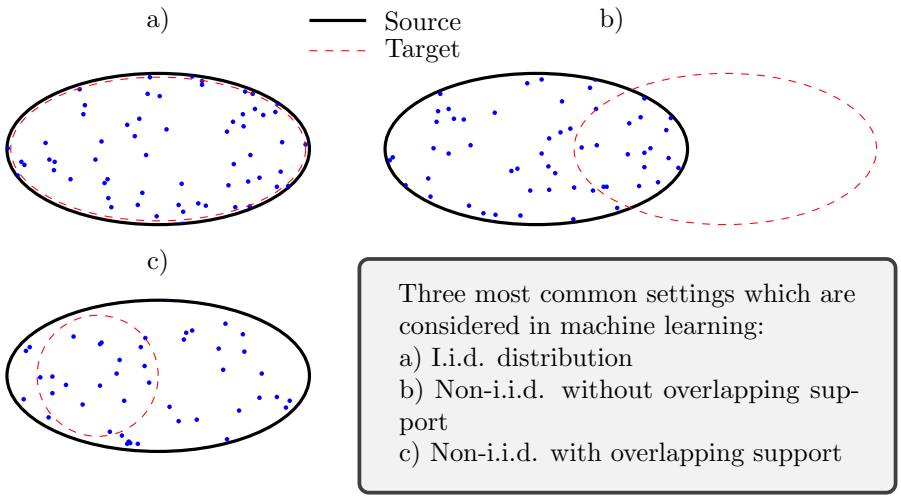


Figure 2.4: Illustration of different settings including shifts in distribution and how this relates to overlapping support of the distributions.

2.5 Privileged information

Learning using privileged information (LUPI) is a framework that was first introduced by Vapnik & Vashist (2009). In this setting we assume we have access to some additional information when training our model. This auxiliary data is in addition to the features and labels available in regular supervised learning. Furthermore, a key assumption is that we only have access to this information at the training stage and not when performing inference.

This additional information is called *privileged information* (PI) and can be many different things: medical journals, object segmentations, bounding boxes etc. The main goal in this framework is to accelerate the pace of learning. The motivation for why this would be achieved is that in real-world learning we often have students being taught by a teacher. This teacher has better knowledge about what material, and how it should be presented to the students in order for them to learn the concepts faster. This may be specific explanations or interjections made to facilitate the students learning. It is this process that the PI framework seeks to imitate. So, using this analogy further there would be some data (x, y) generated by the world and the privileged information would then be generated by the teacher using the conditional distribution $P(w|x)$ which is assumed to be unknown. A related field is that of knowledge distillation (Hinton et al., 2015) which has been noted to be compatible with PI (Lopez-Paz et al., 2016).

Formally, we assume the existence of some PI, $W \in \mathcal{W}$, which is related to the data through $P(w|x)$. Thus the problem is the following: Given tuples $\{(x_i, w_i, y_i)\}_{i=1}^N$ we seek to learn a function f which predicts the outcomes y_i given the data x_i. In contrast to regular supervised learning, we only have access to the PI w_i during training. Since we do not assume that we have access

to this data at test time, the resulting model, f, cannot explicitly depend on the PI, w, as an input. Note, we do not assume any specific form or other properties of the PI. Since explicit model dependence on privileged information is prohibited, one approach is that PI is used in the loss function to learn the model. (Vu et al., 2019) Thereby guiding the optimization based on the information available in the PI. Another is to construct a particular architecture which makes use of the PI. (Xie et al., 2020)

The LUPI paradigm have been applied in settings such as image recognition (Vu et al., 2019; Hoffman et al., 2016), healthcare (Shaikh et al., 2020), finance (Silva et al., 2010), clustering (Feyereisl & Aickelin, 2012). In the context of distribution shifts we have approaches which study the use of PI with SVMs (Li et al., 2022; Sarafianos et al., 2017). Vu et al. (2019) considered using scene depth as PI in semantic segmentation using neural networks. Regarding theory in the DA setting, Xie et al. (2020) provide some theoretical results for linear classifiers and Motiian (2019) investigated using PI for DA using the information bottleneck method.

Chapter 3

Summary of Included Papers

The general problem of adaptation under distributional shift is impossible to guarantee a solution for. (Ben-David & Urner, 2012) However, with the use of assumptions on problem structure it is possible to successfully adapt to such shifts. In this thesis, we search for ways in which we might achieve successful domain adaptation while *still* having useful guarantees on performance. Both of these goals are attainable in some form for the neural network model class, but seldom together. We have shown in section 2.4 that there are limitations of the current theory that inhibit us from achieving this goal at present. Our first paper deals with illustrating the issue with achieving tractably computable and tight generalization bounds for neural network classifiers. One can often find a neural network classifier that performs well on a UDA task, however, there are no realistic guarantees on performance. The second paper focuses on the second and third questions from Chapter 1 and considers practical ways to use privileged information to achieve consistent learning in DA where such information is available. As part of this we propose a novel set of assumptions which we show lead to consistent learning and extensively empirically validate this. In paper III we consider the federated learning setting with heterogeneous clients, where we have a fixed target domain. We show that allowing the server access to label marginals from clients as well as the intended target leads to large potential gains in performance. This performance comes at the cost of some privacy as clients are required to share this information with the central server. Paper IV considers FL with increased privacy characteristics, where we assume that the individual label sets of clients are considered sensitive. In this setting, which we call the private label set setting, we show a way to adapt common FL methods to this setting. We also present a method of tuning the global model for representational alignment, which we show some theoretical guarantees for.

3.1 Paper I - Practicality of generalization guarantees for unsupervised domain adaptation with neural networks

In high-stakes scenarios, like the healthcare setting, we would ideally want to have some guarantees on how well our models are going to perform before deployment. The most straightforward way of achieving this is by upper bounding the error on the target domain. This can be achieved theoretically in many different ways with varying degrees of utility. We can trivially state that the risk is less than or equal to 1 for any task, you cannot be more wrong than all of the time. However, this bound is not particularly informative so we would like to find better bounds which give us further insights into the expected performance of models. Ideally, we would like to have guarantees which are tightly linked with the performance we observe in application while also being possible to evaluate a priori. In Paper I we search the domain adaptation literature for existing bounds which have three main properties:

1. Tightness – Is the bound a poor approximation of the true risk?
2. Estimability – Can we estimate the bound from the observed data?
3. Computability – Can we tractably compute the bound for real-world data sets and hypothesis classes?

After conducting an extensive search we arrive at the conclusion that most available bounds do not fulfill these desiderata. Our final selection contains three types of bounds: importance weighting (IW) bounds, bounds containing integral probability metrics (IPMs) and PAC-Bayesian bounds. To enable easier comparison we adapt the IW and IPM methods to the PAC-Bayesian framework, thereby creating two novel corollaries to a theorem due to McAllester (2013). These bounds, along with two existing ones due to Germain et al. (2020), are the ones we choose to computationally evaluate. One of them requires access to target labels to compute and is included for comparison.

We find that without further modification our evaluation results in vacuous bounds due to the sample generalization terms being too large. To remedy this we apply the practice of learning data-dependent priors which entails sacrificing a part of the sample to inform the choice of prior. This tightens the bounds as the sample generalization term, which measures the difference between the prior and posterior distributions, is smaller using this.

We then compute the four different bounds for two image classification tasks, one based on digit classification and one based on X-ray classification. We find that the bound which requires target labels is the tightest, followed by our IW bound which is computable without such information. The other bounds struggle to remain tight even for the simpler digit classification task. We conclude that in cases where our assumptions hold, an importance weighting strategy works well for bounding the error tightly. Further, we conjecture that changing current assumptions will be a way towards a more complete theory explaining out-of-distribution generalization.

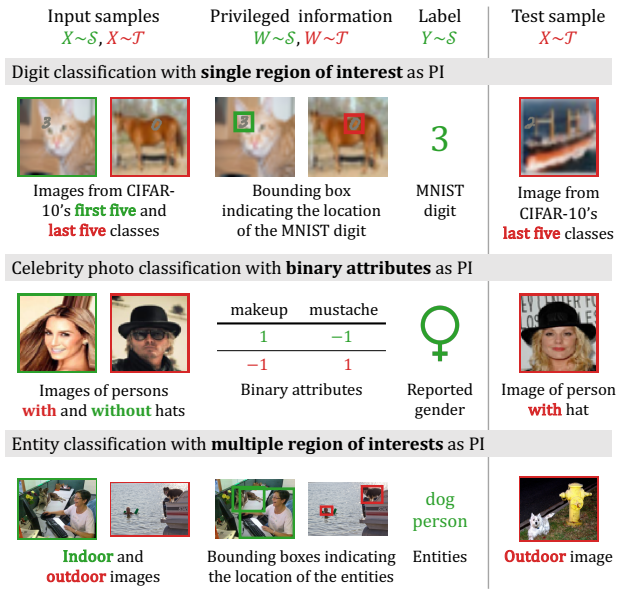
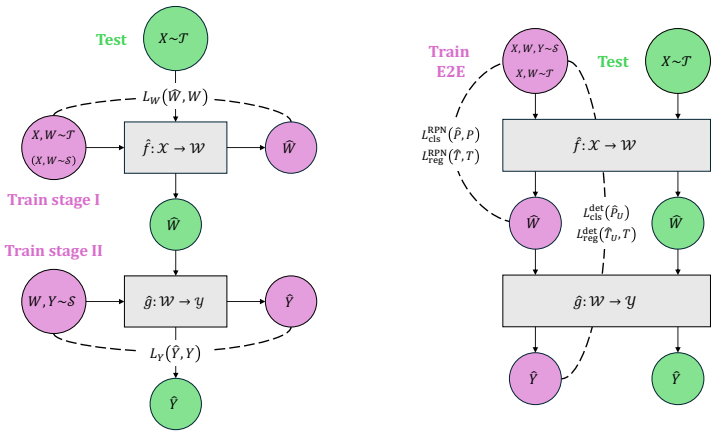


Figure 3.1: An illustration of the data available in the domain adaptation by learning using privileged information (DALUPI) setting. During training, input samples X and privileged information W are available from both source and target domains. Labels Y are only available for inputs from the source domain. At test time, a target sample X is observed.

3.2 Paper II - Unsupervised domain adaptation by learning using privileged information

In this paper we propose some changes to one of the standard sets of assumptions commonly used in UDA, which we do by using the concept of privileged information (PI). We put forward a version of UDA where we assume access to some privileged information, where the PI is assumed to be available in both the source and target domains during training time. Furthermore, at test time, we only assume that we have access to target features as in regular UDA. We call this setup Domain Adaptation by Learning Using Privileged Information (DALUPI). The general structure of our setting is illustrated in 3.1 with some examples. We construct theory based on this novel structure which ensures consistent learning without the reliance on the overlapping support assumption in the input space. The overlap assumption is often violated in practice and as such we would prefer not to build UDA theory using it. The DALUPI setting enables a novel and straightforward way of transferring the model from the source to the target. We simply learn two separate mappings, one from input features to PI and one from PI to the label or outcome. This also gives us a simple way to make theory which conforms to this structure, we just learn one mapping after the other. To avoid the overlap assumption in the input



(a) Two-stage estimator. (b) End-to-end estimator.

Figure 3.2: A schematic representation of the train and test flow for DALUPI using (a) the two-stage estimator and (b) an end-to-end architecture based on Faster R-CNN.

space we instead assume overlap w.r.t the PI, which we argue is more plausible. Additionally, we assume covariate shift w.r.t the PI similar to what is used in other assumption sets. If we add the additional assumption that PI is sufficient for predicting the outcome we show that will have consistent learning. We also propose a bound on the target risk for this setting.

We conduct experiments on four different tasks; a synthetic experiment where we investigate how well a model, which is the composition of two separate mappings learned independently, performs when the amount of overlap is varied. The dataset is constructed by inserting a digit in a larger image and having the bounding box around the digit as PI. It shows that a model based on our framework outperforms all other models. We also adapt an experiment from another work (Xie et al., 2021), based on the CelebA dataset, to compare our method to theirs. Upon comparison, We find that we achieve similar performance to the method proposed by Xie et al. (2021).We then perform two other experiments, one entity classification task based on the MS-COCO dataset and pathology classification from chest X-rays. For these experiments we also propose an end-to-end model, based on the Faster R-CNN architecture (Ren et al., 2015). A schematic overview of the two model architectures we use are presented in Figure 3.2.

In all experiments, except the one based on CelebA, the PI we consider are bounding boxes around the region(s) which are informative for the labeling. In the CelebA experiment, the PI used is a vector of binary appearance attributes present in the pictures. See Figure 3.1 for clarification of the settings we consider.

From the entity classification task we learn that our method outperforms both the UDA baseline well as performs on par with the model which has been given access to target labels. From the X-ray classification task we learn

that the use of PI can yield increased sample efficiency, in-line with previous observations. However, the sufficiency of the PI in this task is not obvious, nor guaranteed. We conclude therefore that the DALUPI setting can be beneficial, even when our assumptions are unlikely to hold. In addition, we note that a domain expert will likely be needed to be able to judge the sufficiency of PI for tasks like the X-ray classification task we considered.

3.3 Paper III - Overcoming label shift in target-aware federated learning

Distributional shifts are not only a cause for concern in centralized settings such as regular DA. Many of the same issues arise in federated learning where we consider a collaborative effort between several clients which contribute to the production of a global model. This training is administrated by a central server which aggregates the model weights which are transmitted from each client to construct the global model. When the clients datasets differ in some way between them we may need to account for this. Furthermore, it is not generally guaranteed that the aggregate of clients is equal to the test data distribution.

In this paper we aim to find a method to remedy this when the clients label distributions differ, but the labels conditioned on inputs is static, i.e . we assume label shift. In this setting there are methods to account for such a shift, however, these do not consider the case that the test domain may not be i.i.d. to the aggregate of all clients. This is an implicit assumption made in most works which propose to mitigate label distribution heterogeneity between clients; we will not make this assumption. Instead, we will consider a target-aware setting where the central server has access to label marginals from the clients as well as the fixed target domain. If we allowed for this information to be shared across clients one could make use of an importance weighting strategy, however, this is not feasible here due to the client's privacy.

We propose to use this information to aggregate client parameters according to the solution of the following simple optimization problem

$$\alpha^{\lambda} = \underset{\alpha \in \Delta^{M-1}}{\arg\min} \|T(Y) - \sum_{i=1}^{M} \alpha_i S_i(Y)\|_2^2 + \lambda \sum_i \frac{\alpha_i^2}{n_i} \ , \tag{3.1}$$

for a given parameter λ and aggregate client parameters as $\theta_{t+1}^{\lambda} = \sum_{i=1}^{M} \alpha_i^{\lambda} \theta_{i,t}$. We refer to this strategy as Federated learning with Parameter Aggregation for Label Shift (FedPALS).

This strategy finds the optimal trade-off between the model aggregation closely aligning with the target label distribution, and minimizing the variance due to weighting using the inverse of the effective sample size (ESS) Kong (1992).

FedPALS also has the nice property that in the limit of $\lambda \to \infty$ we recover the regular FedAvg aggregation, which is equivalent to maximizing the effective sample size.

We perform experiments on several different datasets where we show that FedPALS perform quite well with its awareness of label marginals. We perform classification experiments where we on CIFAR-10 and Fashion-MNIST perform both sparsity sampling and dirichlet sampling. We also construct an experiment based on PACS that modify the dataset to consist of three clients, each missing a label that is present in the other two. Additionally, one client is reduced to one-tenth the size of the others, and the target distribution is made sparse in the same label as that of the small client. As a more real-world experiment we

Table 3.1: Comparison of mean accuracy and standard deviation (±) across different algorithms. The reported values are over 8 independent random seeds for the CIFAR-10 and Fashion-MNIST tasks, and 3 for PACS. C indicates the number of labels per client and β the Dirichlet concentration parameter. M is the number of clients. The *Oracle* method refers to a FedAvg model trained on clients whose distributions are identical to the target.

Data set	Label split	M	FedPALS	FedAvg	FedProx	SCAFFOLD	AFL	FedRS	Oracle
Fashion-MNIST	$C = 3$	10	$\mathbf{92.4 \pm 2.1}$	67.1 ± 22.0	66.9 ± 20.8	69.5 ± 19.3	78.9 ± 14.7	85.3 ± 13.5	97.6 ± 2.1
	$C = 2$		$\mathbf{80.6 \pm 23.7}$	53.9 ± 36.2	52.9 ± 35.7	54.9 ± 36.8	78.6 ± 20.0	63.14 ± 20.2	97.5 ± 4.0
CIFAR-10	$C = 3$	10	$\mathbf{65.6 \pm 10.1}$	44.0 ± 8.4	43.5 ± 7.2	43.3 ± 7.4	53.2 ± 0.9	44.0 ± 8.0	85.5 ± 5.0
	$C = 2$		$\mathbf{72.8 \pm 17.4}$	46.7 ± 15.8	47.7 ± 15.6	46.7 ± 14.9	54.7 ± 0.1	49.4 ± 9.5	89.2 ± 3.9
	$\beta = 0.1$		$\mathbf{62.6 \pm 17.9}$	40.8 ± 9.2	41.9 ± 9.7	43.5 ± 10.5	53.4 ± 11.5	57.1 ± 11.2	79.2 ± 3.7
PACS	$C = 6$	3	$\mathbf{86.0 \pm 2.9}$	73.4 ± 1.6	75.3 ± 1.3	73.9 ± 0.3	74.5 ± 0.9	76.1 ± 1.6	90.5 ± 0.3

evaluate performance on the iWildCam dataset which is sampled sparsely as detailed in the benchmark from Bai et al. (2024).

We compare the performance of FedPALS to similar approaches which are designed to combat client heterogeneity: FedProx, SCAFFOLD and AFL. We also compare to regular FedAvg and an oracle classifier for which the i.i.d. assumption is true to understand how much the introduction of label marginals help performance.

We show the results of these experiments in Table 3.1. We clearly see that the awareness of label marginals of FedPALS is beneficial and see quite substantial gains in many tasks. We also observe that there seems to be a bias-variance trade-off as the FedPALS results experiencing increased variance. This indicates that the addition of target label distribution faithfulness increases the variance of the resulting estimator. Further, we show in an ablation on the PACS task, that increasing the noise level in the target label marginal decreases the performance we observe from FedPALS which is expected.

Overall, we find that FedPALS balances the trade-off between matching the target label distribution and minimizing variance in the model updates. In particular, FedPALS perform well in challenging scenarios where label sparsity and client heterogeneity hinder the performance of other methods.

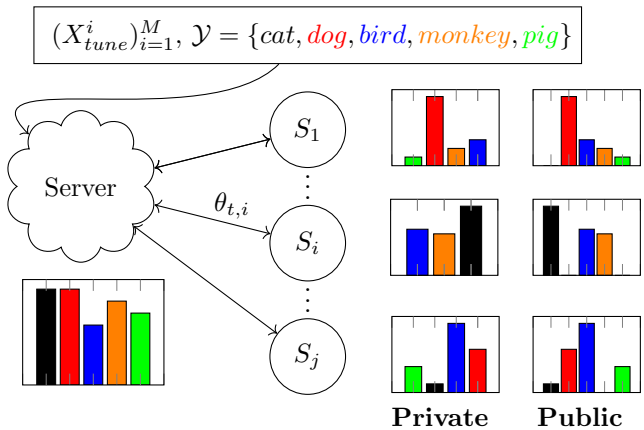


Figure 3.3: A schematic view of the two settings which we consider in Paper IV. The private setting where the clients are unaware of the full label set and the public setting where this is known. Note that the server has access to the information about which clients have which label sets.

3.4 Paper IV - Federated learning with heterogenous and private label sets

Considering the interest in handling heterogeneity in client datasets, it is reasonable to investigate how to handle this in cases where client label sets are considered sensitive. With current methods not sharing this would be unavoidable as the clients need to all agree on the full label set which they are constructing a classifier for. If the label set includes sensitive information, clients will be unwilling to share which labels they have available to other clients.

To ameliorate this issue, we propose the *private label set* setting. We adapt the regular FL setup to allow for client label sets being kept private. This is done by trusting only the server with the information about which labels a client has in its dataset. We illustrate the setup and how it compares to the regular setup, which we call *public*, in Figure 3.3.

Considering this settings challenges, We propose two main approaches to learning a model. First we propose an adaptation of FedAvg which is possible to extend to other FL approaches. This consists of two main modifications, in each round t, each client k is sent the full set of current encoder parameters θ^t_ϕ and the subset of current classifier parameters corresponding to their label set, $\theta^t_\psi[\mathcal{Y}_k] := [\theta^t_\psi(y) : y \in \mathcal{Y}_k]^\top \in \mathbb{R}^{|\mathcal{Y}_k|}$. Clients then proceed with local updates as normal.

On the server side, the server receives parameter updates every round $(\theta^t_{\phi,k}, \theta^t_{\psi,k})$ from each client k and averages the classifier parameters for each label y based on models of clients which have the label in their label set, weighted according to their sample size (see Algorithm 2). Encoder parameter updates $\theta^t_{\phi,k}$ are averaged as we would with regular FedAvg. Second, we

Algorithm 2: FedAvg with private label sets

Data: Client label sets $\{\mathcal{Y}_k\}$ and reverse indices $\{I_k\}$
Result: Classifier $h(x) = \sigma(\theta_\psi^\top \phi(x))$
Initialize central parameters $\theta^0 = (\theta_\phi^0, \theta_\psi^0)$
for *each round* $t = 0, ..., T-1$ **do**
  for *each client* $k = 1, ..., m$ **do**
    Distribute $(\theta_\phi^t, \theta_\psi^t[\mathcal{Y}_k])$ to client k
    Receive client update $(\theta_{\phi,k}^t, \theta_{\psi,k}^t)$
  end
  $\theta_\phi^{t+1} = \sum_{k=1}^m \theta_{\phi,k}^t \frac{n_k}{n}$ where $n = \sum_{k=1}^m n_k$
  for *each label* $y \in \mathcal{Y}$ **do**
    $\theta_\psi^{t+1}(y) = \sum_{k:y\in\mathcal{Y}_k} \theta_{\psi,k}^t(I_k(y)) \frac{n_k}{n'_y}$ where $n'_y = \sum_{k:y\in\mathcal{Y}_k} n_k$
  end
end
Return classifier with parameters $\theta = (\theta_\phi^T, \theta_\psi^T)$

propose a tuning strategy in conjunction with the above aggregation. This is motivated by earlier work in the classifier coupling literature (Hastie & Tibshirani, 1997; Wu & Weng, 2004). This is motivated by the fact that the first strategy will not be guaranteed to work well when the representation is suboptimal for one or more clients.

In classifier coupling approaches, methods were developed to combine several *binary* classifiers, e.g., support vector machines, on different pairs of labels into a single multi-class classifier. This technique not commonly used as multi-class classifiers are trained routinely using neural networks with softmax outputs or similar models. However, in federated learning with heterogenous and private label sets, we face a similar problem since no client nor the server has access to labeled data from all classes.

Taking inspiration from these approaches we show that a tuning approach using an unlabeled dataset on the central server is a viable option. One which, under the right conditions, could produce an optimal classifier given optimal client models. However, while this particular special case is unlikely to occur, it is not unreasonable to believe that such a method could be empirically useful. We use a tuning approach with two main loss functions:

$$\underset{h\in\mathcal{H}}{\text{minimize}} \sum_{k=1}^{m} w_k \sum_{\substack{y,y'\in\mathcal{Y}_k \\ y\neq y'}} \mathbb{E}_{\boldsymbol{X}} \left[\left(h_k(y \mid \boldsymbol{X}) h(y' \mid \boldsymbol{X}) - h_k(y' \mid \boldsymbol{X}) h(y \mid \boldsymbol{X}) \right)^2 \right], \tag{3.2}$$

which we call the *pairwise* loss which align model prediction in a pairwise manner. We also consider the direct MSE loss

$$\underset{h\in\mathcal{H}}{\text{minimize}} \sum_{k=1}^{m} w_k \sum_{y\in\mathcal{Y}_k} \mathbb{E}_{\boldsymbol{X}} \left[\left(h_k(y \mid \boldsymbol{X}) - h(y \mid \boldsymbol{X}) \right)^2 \right]. \tag{3.3}$$

We perform classification experiments on both Fashion-MNIST and CIFAR-10 where we vary the number of labels that a client has access to from 2-10. We choose a random set of labels for each client which they then get distributed from the dataset equally. To combat the the heterogeneous amount of examples in clients resulting from this we perform a subsampling step, which results in each client having an equal amount of samples.

We observe from our experiments that our adaptation of regular FL methods to the private setting perform quite well. This suggests that the simple per-class aggregation is sufficient to achieve adequate performance. Further, this indicates that alignment of representation may not always be necessary. However, the pairwise tuning outperforms our adaptations of FedAvg and FedProx on CIFAR-10 in the public setting. This suggests that the tuning of the representation can be of use in this setting also.

In summary, paper IV shows that federated learning is still possible in a private label set scenario. Furthermore, our proposed approaches successfully enable learning in this setting. We observe empirically that these methods work well with a minimal performance drop compared to performance in the public setting.

Chapter 4

Concluding remarks and future directions

In this thesis we have presented the challenges posed by distributional shifts in machine learning. We have explored how to both predict performance as well as adapt to these shift in both centralized and federated scenarios. We performed a survey of generalization bounds with the express purpose to find ones which can practically measure and assess the target performance of models. We also proposed a method for leveraging the structure inherent in privileged information which can be available during training. In the federated learning context, we introduced the concept of target-aware federated learning and proposed a method to leverage the server-side knowledge of the target. Finally, we consider a setting where label sets are too sensitive to share across clients. We have formalized two novel approaches to performing federated learning with both heterogeneous clients and private label sets. Overall, we have shown that there are significant benefits from exploring novel settings and modifying the assumptions of what data is available or their statistical properties.

The methods and theory we propose all come with their own strengths and limitations. These are all discussed in each paper which is appended in Part II of the thesis, however, some are worth discussing here. Most methods and settings we investigate assume that we have access to more information or that the underlying data has certain structure. We have also generally focused on image classification datasets in our work. Of course, this is limiting as there are many other types of data which we could have considered, e.g tabular or time-series data, each with their own distinct properties. However, this is motivated by the large dimensionality of images, and the models used to process them. This is a complicating factor in most theoretical treatments and why we chose to work with these.

In future, the focus should be to further attempt to solve the distributional shift problem for specific scenarios where different amounts of information is available. We have made several inroads in these areas, but several promising avenues still remain to explore further in this respect. Similar to the setting we propose in paper II one might explore how the addition of other modalities

of data could impact learning both on a theoretical and empirical level. Furthermore, in the context of paper III there are many interesting settings to consider. One could attempt to further explore to what extent the information in target-aware federated learning can be inferred from model updates from clients. If this inference is both plausible to conduct and accurate enough for guiding learning one could lessen the amount of information clients would need to explicitly share. An alternative would be that we could transmit the information in a compressed form. For instance, we could achieve this through sharing some representation of the label distribution similar to the approach in Mclaughlin & Su (2024).

Building on paper IV, one could easily envision a setting with even further privacy gains. One would not share the model parameters at all but allow there to be a querying system between the clients and the server. Perhaps they could be answers to how the models would classify certain, publicly available, inputs. The global and local models could then be updated based on the responses to these queries, making use of a similar alignment method as we introduce in our work.

Furthermore, it is of great interest to develop novel theoretical perspectives to make use of the underlying structure of problems explicitly. This may be done by presupposing a specific underlying structure of the data and treating the problem based on this assumption. For example, there have been works that incorporate the notion of a taxonomy in their data and tailor their architecture to exploit this hierarchical structure (Liu et al., 2023). This raises the compelling question of which structures are most beneficial, and how they can be effectively integrated into learning frameworks.

Historically, architectural innovations such as Convolutional Neural Networks (Lecun, 1998) and Transformers (Vaswani et al., 2017) have achieved remarkable success by embedding strong inductive biases directly into the model design. These approaches represent embedding structural assumptions at the architectural level. However, it remains an open question whether similar benefits can be achieved through purely methodological approaches, i.e., without altering model architecture. Crucially, any such assumptions must be justifiable and, ideally, verifiable a priori. If the assumed structure cannot be assessed from the available data, its practical utility is limited. Although, it may be permissible if the empirical gains derived from making the assumption is substantial enough. Whether these structural assumptions can be made general across tasks or must be tailored to specific applications remains an open and important question. If the latter holds, considerable effort will be required to identify and formalize appropriate structures for each application or task.

On a related note, the recent success of large pretrained foundation models suggests that transferability can, to some extent, be achieved simply through access to vast and diverse datasets e.g. LAION (Schuhmann et al., 2022). This brings to light a salient question: in such large-scale settings, is the overlapping support assumption, central to many theoretical generalization guarantees, more likely to hold due to the breadth and diversity of the training data?

Taken together, the work presented in this thesis highlights the importance

of revisiting core assumptions in machine learning. By proposing novel methods and theoretical perspectives in both centralized and federated settings, we have shown that addressing these issues directly can lead to more robust and adaptable models. Future work should continue by investigating how structure in data sets can be leveraged in both model design and method development. Additionally, further theoretical grounding is needed to understand the implications of such assumptions and to formalize generalizable principles across tasks and domains.

As machine learning increasingly moves toward real-world deployment, addressing distributional shift will remain a key concern. This work has provided some initial steps toward this goal, and with future efforts we will hopefully be able to produce powerful, but also reliable and adaptable models, in the face of real-world complexity.

Bibliography

David Acuna, Guojun Zhang, Marc T. Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/acuna21a.html`.

Amiran Ambroladze, Emilio Parrado-hernández, and John Shawe-taylor. Tighter pac-bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Ruqi Bai, Saurabh Bagchi, and David I Inouye. Benchmarking algorithms for federated domain generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pp. 139–153. Springer, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility Theorems for Domain Adaptation. In *International Conference on Artificial Intelligence and Statistics*, 2010b.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. *Journal of Machine Learning Research*, 9:129–136, 2010c. URL `http://www.jmlr.org/proceedings/papers/v9/david10a.html`.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, pp. 120, Sydney, Australia, 2006. Association for Computational Linguistics. ISBN 978-1-932432-73-2.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning Bounds for Domain Adaptation. *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp. 8, 2008.

Mahrokh Ghoddousi Boroujeni, Andreas Krause, and Giancarlo Ferrari-Trecate. Personalized federated learning of probabilistic models: A pac-bayesian approach. *ArXiv*, abs/2401.08351, 2024.

Mahrokh Ghoddousi Boroujeni, Andreas Krause, and Giancarlo Ferrari-Trecate. Personalized federated learning of probabilistic models: A PAC-bayesian approach. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=ZMliWjMCor`.

Adam Breitholtz and Fredrik Daniel Johansson. Practicality of generalization guarantees for unsupervised domain adaptation with neural networks. *Transactions on Machine Learning Research*, 2022.

Adam Breitholtz, Anton Matsson, and Fredrik D. Johansson. Unsupervised domain adaptation by learning using privileged information. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=saV3MPH0kw`.

Adam Breitholtz, Edvin Listo Zec, and Fredrik D. Johansson. Federated learning with heterogenous and private label sets. In *3rd Workshop on Advancements in Federated Learning (WAFL), ECML-PKDD*, 2025.

George W. Brown. Basic principles for construction and application of discriminators. *Journal of Clinical Psychology*, 6(1):58–61, 1950. doi: https://doi.org/10.1002/1097-4679(195001)6:1⟨58::AID-JCLP2270060112⟩3.0.CO;2-B. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-4679%28195001%296%3A1%3C58%3A%3AAID-JCLP2270060112%3E3.0.CO%3B2-B`.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. March 2023. URL `https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/`.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2005.05.005. URL `https://www.sciencedirect.com/science/article/pii/S0885230805000276`.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv:1802.07876*, 2018.

Qi Chen and Mario Marchand. Algorithm-dependent bounds for representation learning of multi-source domain adaptation. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10368–10394. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/chen23h.html`.

Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J. Su. Minimax estimation for personalized federated learning: an alternative between fedavg and local training? *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, January 2014. ISSN 0304-3975.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning Bounds for Importance Weighting. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23, pp. 442–450. Curran Associates, Inc., 2010.

Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation Algorithm and Theory Based on Generalized Discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178, Sydney NSW Australia, August 2015. ACM. ISBN 978-1-4503-3664-2.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint Distribution Optimal Transportation for Domain Adaptation. *arXiv:1705.08848 [cs, stat]*, October 2017. arXiv: 1705.08848.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(57):1757–1774, 2008. URL `http://jmlr.org/papers/v9/crammer08a.html`.

Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research (JAIR)*, 26:101–126, 2006. URL `http://hal3.name/docs/#daume06megam`.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui

Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL `https://arxiv.org/abs/2412.19437`.

Sofien Dhouib, Ievgen Redko, and Carole Lartizien. Margin-aware adversarial domain adaptation with optimal transport. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2514–2524. PMLR, 13–18 Jul 2020.

Sofien Dhouib and Ievgen Redko. Revisiting (ϵ , γ, τ)-similarity learning for domain adaptation. *NeurIPS*, pp. 7408–7417, 2018.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology, 2024. URL `https://arxiv.org/abs/2411.19666`.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel M. Roy. On the role of data in PAC-Bayes bounds. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, October 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: recurring and unseen tasks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Jan Feyereisl and Uwe Aickelin. Privileged information for data clustering. *Information Sciences*, 194:4–23, 2012.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers. In *International Conference on Machine Learning*, pp. 738–746. PMLR, May 2013. ISSN: 1938-7228.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. PAC-Bayes and Domain Adaptation. *Neurocomputing*, 379:379–397, February 2020. ISSN 09252312. arXiv: 1707.05712.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4694–4703, Brussels, Belgium,

October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1498. URL https://aclanthology.org/D18-1498/.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, NIPS'97, pp. 507–513, Cambridge, MA, USA, 1997. MIT Press.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with Side Information through Modality Hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 826–834, Las Vegas, NV, USA, June 2016. IEEE.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 8256–8266, Red Hook, NY, USA, 2018. Curran Associates Inc.

Min Huang, Zifeng Xie, Bo Sun, and Ning Wang. Multi-source unsupervised domain adaptation with prototype aggregation. *Mathematics*, 13(4), 2025. ISSN 2227-7390. doi: 10.3390/math13040579. URL https://www.mdpi.com/2227-7390/13/4/579.

Rebecca Hwa. Supervised grammar induction using training data with limited constituent information. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pp. 73–79, 1999.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

C. Jordan. On Fourier series. *C. R. Acad. Sci., Paris*, 92:228–230, 1881. ISSN 0001-4036.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner,

Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Far inaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, M ariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Y ang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL `http://dx.doi.org/10.1561/2200000083`.

Okan Koc, Alexander Soen, Chao-Kai Chiang, and Masashi Sugiyama. Domain adaptation and entanglement: an optimal transport perspective. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL `https://openreview.net/forum?id=ZDyi1BeTu7`.

Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348:14, 1992.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 00034851. URL `http://www.jstor.org/stable/2236703`.

Seiichi Kuroki, Nontawat Charoenphakdee, Han Bao, Junya Honda, Issei Sato, and Masashi Sugiyama. Unsupervised Domain Adaptation Based on Source-Guided Discrepancy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4122–4129, July 2019.

Yann Lecun. Gradient-Based Learning Applied to Document Recognition. *proceedings of the IEEE*, 86(11):47, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, 2020.

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

Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021.

Yanmeng Li, Huaijiang Sun, and Wenzhu Yan. Domain adaptive twin support vector machine learning using privileged information. *Neurocomputing*, 469: 13–27, 2022.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Tianyi Liu, Zihao Xu, Hao He, Guang-Yuan Hao, Guang-He Lee, and Hao Wang. Taxonomy-structured domain adaptation. In *International Conference on on Machine Learning*, 2023.

Yiling Liu, Juncheng Dong, Ziyang Jiang, Ahmed Aloui, Keyu Li, Michael Hunter Klein, Vahid Tarokh, and David Carlson. Understanding and robustifying sub-domain alignment for domain adaptation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=oAzu0gzUUb`.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *arXiv:1502.02791 [cs]*, May 2015. arXiv: 1502.02791.

David Lopez-Paz, Leon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain Adaptation with Multiple Sources. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp. 1041–1048, January 2008.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009, pp. 367–374. AUAI Press, 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. In *Proceedings of the Conference on Learning Theory*, February 2009b.

David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pp. 230–234, New York, NY, USA, July 1998. Association for Computing Machinery. ISBN 978-1-58113-057-7.

David A. McAllester. A PAC-Bayesian Tutorial with A Dropout Bound. *arXiv e-prints*, 1307:arXiv:1307.2118, July 2013.

Connor Mclaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=Wl2optQcng`.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/mohri19a.html`.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, November 2012. ISSN 0219-1377, 0219-3116.

Saeid Motiian. *Domain Adaptation and Privileged Information for Visual Recognition*. phdthesis, https://researchrepository.wvu.edu/etd/6271, 2019. Graduate Theses, Dissertations, and Problem Reports. 6271.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 00018678. URL `http://www.jstor.org/stable/1428011`.

J Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY OF LONDON SERIES A-CONTAINING PAPERS OF A MATHEMATICAL OR PHYSICAL CHARACTER*, 231:289–337, MAR 1933. ISSN 0264-3952. doi: 10.1098/rsta.1933.0009.

A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip Torr, and Atilim Gunes Baydin. KL guided domain adaptation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=0JzqUlIVVDd`.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13(1):3507–3531, dec 2012. ISSN 1532-4435.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2019. doi: 10.1109/ICCV.2019.00149.

Sashank Reddi, Zachary Burr Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Brendan McMahan (eds.). *Adaptive Federated Optimization*, 2021.

Ievgen Redko. *Nonnegative matrix factorization for transfer learning*. PhD thesis, Paris North University, 2015.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical Analysis of Domain Adaptation with Optimal Transport. *arXiv:1610.04420 [cs, stat]*, July 2017. arXiv: 1610.04420.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 849–858. PMLR, 16–18 Apr 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new aprioriprobabilities: A simple procedure. *Neural Computation*, 14(1):21–41, January 2002. doi: 10.1162/089976602753284446. URL `http://dx.doi.org/10.1162/089976602753284446`.

Nikolaos Sarafianos, Michalis Vrigkas, and Ioannis A. Kakadiaris. Adaptive svm+: Learning with privileged information for domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32 (8):3710–3722, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Tawseef Ayoub Shaikh, Rashid Ali, and M. M. Sufyan Beg. Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Machine Vision and Applications*, 31(1):9, February 2020.

John Shawe-Taylor and Robert C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational*

Learning Theory, COLT '97, pp. 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein Distance Guided Representation Learning for Domain Adaptation. *arXiv:1707.01217 [cs, stat]*, March 2018. arXiv: 1707.01217.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. ISSN 0378-3758.

Changjian Shui, Zijian Li, Jiaqi Li, Christian Gagné, Charles X Ling, and Boyu Wang. Aggregating from multiple target-shifted sources. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9638–9648. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/shui21a.html`.

Catarina Silva, Armando Vieira, Antonio Gaspar-Cunha, and Joao Carvalho das Neves. Financial distress model prediction using SVM+. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–7, July 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017. doi: 10.1038/nature24270. URL `http://dx.doi.org/10.1038/nature24270`.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, pp. 1134–1142, 1984.

V. Vapnik. Principles of risk minimization for learning theory. In *Proceedings of the 5th International Conference on Neural Information Processing Systems*, NIPS'91, pp. 831–838, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 9 1998. ISBN 0471030031.

Harsh Vardhan, Avishek Ghosh, and Arya Mazumdar. An improved federated clustering algorithm with model-based clustering. *Transactions on Machine Learning Research*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7363–7372, 2019.

B. L. Welch. (ii) note on discriminant functions. *Biometrika*, 31(1-2):218–218, 07 1939. ISSN 0006-3444. doi: 10.1093/biomet/31.1-2.218. URL `https://doi.org/10.1093/biomet/31.1-2.218`.

Ting-Fan Wu and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5: 975–1005, 2004.

Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. *arXiv preprint arXiv:2012.04550*, 2020.

Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*, 2021.

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Wenzhe Yin, Shujian Yu, Yicong Lin, Jie Li, Jan-Jakob Sonke, and Efstratios Gavves. Domain adaptation with cauchy-schwarz divergence. In *Conference on Uncertainty in Artificial Intelligence*, 2024. URL `https://api.semanticscholar.org/CorpusID:270123453`.

Edvin Listo Zec, Adam Breitholtz, and Fredrik D. Johansson. Overcoming label shift in targeted federated learning. In *Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM)*, 2025. URL `https://openreview.net/forum?id=ZUlrLeW82p`.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Dun Zeng, Zenglin Xu, Yu Pan, Qifan Wang, and Xiaoying Tang. Tackling hybrid heterogeneity on federated optimization via gradient diversity maximization. *arXiv preprint arXiv:2310.02702*, 2023.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I. Jordan. Bridging Theory and Algorithm for Domain Adaptation. *arXiv:1904.05801 [cs, stat]*, April 2019. arXiv: 1904.05801 version: 1.

Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12975–12983, April 2020. doi: 10.1609/aaai.v34i07.6997.

Tailin Zhou, Zehong Lin, Jinchao Zhang, and Danny H. K. Tsang. Understanding and improving model averaging in federated learning on heterogeneous data. *IEEE Transactions on Mobile Computing*, 2023.