Beyond Discrepancy: A Closer Look at the Theory of Distribution Shift

Robi Bhattacharjee ROBI.BHATTACHARJEE@UNI-TUEBINGEN.DE
University of Tübingen and Tübingen AI Center

Nick Rittler NRITTLER@UCSD.EDU
University of California- San Diego

Kamalika Chaudhuri KAMALIKA@UCSD.EDU
University of California- San Diego

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Learning theory of distribution shift generally bounds performance on the target distribution as a function of the discrepancy between the source and target, rarely guaranteeing high target accuracy. Instead of relying on the discrepancy, we adopt an assumption inspired by Invariant Risk Minimization, where the source and target distributions are unified by an unknown feature projection. Under this assumption, we show that a learner can leverage the relationship between the source and target distributions to greatly reduce the number of required target samples to achieve high accuracy. To quantify this effect, we introduce a new combinatorial complexity measure—the distance dimension—and derive bounds for linear maps and neural networks.

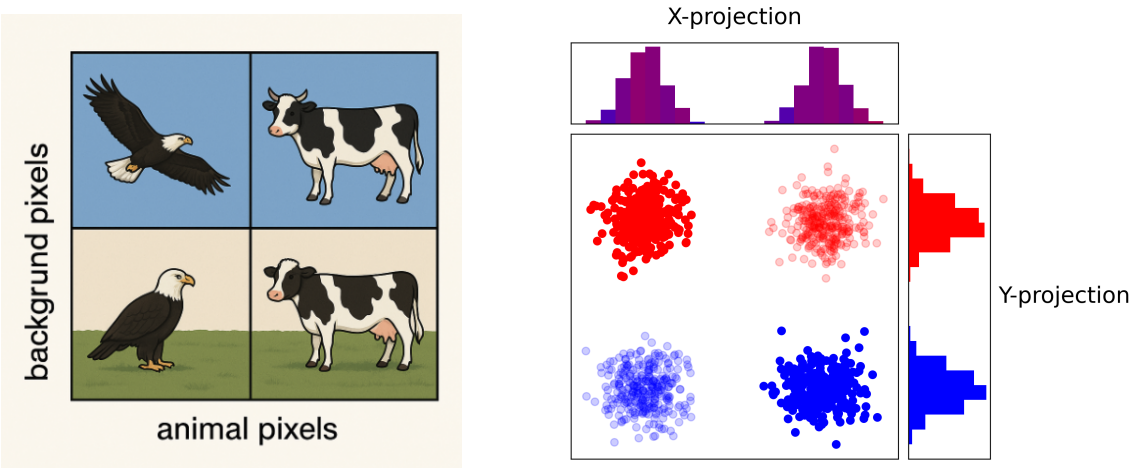
Keywords: transfer learning, out-of-distribution generalization, distribution shift, k-nearest neighbors.

1. Introduction

Classical learning theory operates within the statistical learning framework, in which the training and testing datasets are assumed to be drawn from the same distribution Valiant (1984). However, this assumption is rarely met in practice, where models often succeed in ever-changing real world environments rarely matching the precise conditions of their training data. This motivates the study of distribution shift, in which a learner trains solely or primarily on a source distribution, with the goal of generalizing well over a distinct target distribution.

Thus far, the theory of distribution shift has often taken a worst-case approach, typically bounding generalization error in terms of some notion of discrepancy between the source and target distributions (Ben-David et al., 2006, 2010; Mansour et al., 2012). In cases where the source and target distributions are completely unrelated, or the source provides little information about the decision boundary on the target, discrepancy-based analyses correctly capture the difficulty of generalization.

As a motivating example, consider the problem of classifying animals based on source data that comprises natural training images. For example (as shown in Figure 1 (a)), we might have images of cows in fields and eagles in the sky. At test time, however, we might be asked to classify eagles in the field (of which few source examples exist) or flying cows. This "target" distribution is quite far from the source distribution, based on traditional discrepancy measures, because it inhabits a different part of the input space's support. However, in practice, modern classifiers would correctly learn a representation where (for the purposes of animal classification), only the pixels corresponding to



(*a*) (*b*)

Figure 1: Panel (a) illustrates an animal classification task where source data comprises of animals featured in their natural backgrounds to target data where animals are in unnatural backgrounds. Panel (b) illustrates a case where the source data (bold) and the target data (faded) have disjoint supports. In both cases, the source and target are related through a relevant feature map – in the first case, this map captures the "animal pixels" where as in the second case this is the linear projection onto the Y-axis. Furthermore, both cases *require* some access to the target distribution in order to determine which features are relevant, as otherwise the learner might classify animals based on the background in panel (a) and simply use the x-coordinate in panel (b).

the animal itself are relevant. After extracting such a representation, the source and target are no longer disparate and are quite clearly related.

Motivated by this observation, in this work we investigate the prospect of generalization via the selection of a feature map that appropriately connects the source distribution $\mathcal{D}_s$ and target distribution $\mathcal{D}_t$. We do this under a specific realizability assumption inspired by Invariant Risk Minimization (IRM) (Arjovsky et al., 2020), which centers around the search for feature map ψ^* and a classifier over feature space h^* such that h^* is the optimal classifier for data representations arising from ψ^* across all source and target domains Arjovsky et al. (2020).

We adapt this assumption to one tailored to an investigation of convergence to the target Bayes risk, stipulating the existence of a feature map $\phi^* \in \Phi$ mapping points from the target $\mathcal{D}_t$ close to those from the source $\mathcal{D}_s$ while retaining information sufficient for optimal prediction. We call this the Statistical IRM assumption (Definition 4).

To achieve convergence towards that Bayes-optimal, we utilize k_n-nearest neighbors (k_n-NN) in a chosen feature space. The strong convergence guarantees of k_n-NN allow us to focus our attention on the amount of target data needed to identify feature map leading to target generalization. We begin by showing (Theorem 5) that under the Statistical IRM assumption, applying a k_n-nearest neighbors classifier trained on source data converges towards the Bayes risk for the *target* distribution as long as we first apply the appropriate feature map, ϕ^*.

Next, we investigate the problem of *learning* the feature map ϕ^*. We are left with a natural problem: given two distributions $\mathcal{D}_s$ and $\mathcal{D}_t$ that satisfy our assumption, and given a class of feature maps Φ which contains the unknown map ϕ^*, the learner must converge towards the Bayes optimal of $\mathcal{D}_t$ using predominantly samples from $\mathcal{D}_s$ and a small number of samples from $\mathcal{D}_t$.

To solve this, we introduce a combinatorial complexity measure on an embedding class Φ called the *distance dimension*, and use it to upper bound the amount of labeled target data needed for generalization in this setting. Our result (Theorem 9) implies that convergence towards the target Bayes optimal can be achieved over the target distribution using a nearly distribution-independent number of target samples. In many cases, this is far fewer samples than would be required for directly converging towards the Bayes optimal using pure target data.

We also include upper bounds on our complexity measure for linear feature maps (Proposition 7 and relu-activated neural networks (Proposition 8). We additionally note that when coupled with Theorem 9, our result gives generalization bounds for learning feature maps with linear and neural network maps. To our knowledge, these are the first such bounds for classifiers that compose nearest neighbors with such maps.

2. Related Work

Transfer learning – wherein the learner attempts to generalize through the supplementation of source data with small amounts of labeled target data – has recently been a topic of theoretical interest. The past few years have seen minimax rates established in various important settings, including covariate shift and posterior drift, and under the assumption that source and target share an optimal predictor in a fixed VC class Kpotufe and Martinet (2018); Cao et al. (2010); Maity et al. (2020); Hanneke and Kpotufe (2020); Hanneke et al. (2023). Despite the interest in the problem, we are not aware of a line of theoretical work studying the possibility of converging towards the target Bayes-optimal via representation learning across a single source and target. The most closely related vein of literature of which we are aware is that of ''few-shot representation learning" Maurer et al. (2016); Du et al. (2021); Tripuraneni et al. (2020); Watkins et al. (2023); Tripuraneni et al. (2020), where the goal is to use aggregate data on a set of source tasks in a way that allows for generalization to a related target task with a small amount of labeled target data.

The backbone of many successful applied techniques for unsupervised domain adaptation – where the learner has access to both labeled source and unlabeled target data – is so-called "marginal alignment" Nguyen et al. (2022); Gretton et al. (2012); Glorot et al. (2011); Liu et al. (2022); Li et al. (2018); Sun et al. (2016); Ganin et al. (2016); Shen et al. (2018). The unifying idea in this vein of work, similar to the learning rule we propose in Section 7, is to learn a feature representation under which the source and target distributions' marginals over instance space align, and so implicitly, these strategies rely on the existence of a unifying feature map. The theory of marginal alignment has motivated this approach by bounding the target risk in terms of discrepancy measures between source and target in representation space Shen et al. (2018); Nguyen et al. (2022); Wang and Mao (2023) – our results in the unsupervised domain adaptation setting differ in that we are interested in characterizing when identifying a good feature representation is possible, given the lack of access to labeled target data.

A significant amount of work on representation learning stems from the study of "domain generalization", wherein the learner tries to generalize to a set of testing environments using samples from a smaller set of source environments Blanchard et al. (2011); Muandet et al. (2013); Zhao et al.

(2017); Arjovsky et al. (2020); Bellot and van der Schaar (2020); Mahajan et al. (2021). It is under this model that IRM prescribes searching for a feature map ψ^* under which the classification task is invariant" across environments Arjovsky et al. (2020). Theoretical work has been somewhat critical of IRM as a training routine, focusing on domain generalization problems under which it fails to learn the relevant invariance Rosenfeld et al. (2020); Kamath et al. (2021). By contrast, our work considers a realizability assumption inspired by the desired invariance condition of IRM (Arjovsky et al., 2020), and focuses on settings outside of the scope of domain generalization.

The use of k_n-NN in out-of-distribution generalization problems has been explored by both the applied and theoretical literatures. Berlind and Urner (2015) introduces a k-NN-based algorithm with guarantees for active domain adaptation in the covariate shift setting. Their work does not consider the selection of feature representations, and instead endows the learner with the power to selectively query the labels of target examples, something we do not consider in this work. The application of nearest neighbors in the embedding space of a pre-trained neural network has shown empirical promise in out-of-distribution detection Sun et al. (2022).

3. Preliminaries

Let the instance space $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space, and $\mathcal{Y}$ be a finite label set. A data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is defined as a pair (μ, η) where μ is a Borel measure over $\mathcal{X}$ and η is a conditional probability distribution $\eta(y|x)$.

For a classifier $h : \mathcal{X} \to \mathcal{Y}$, we define its **risk** $R(h, \mathcal{D})$ over $\mathcal{D}$ as the probability it missclassifies. We denote this by $R(h, \mathcal{D}) := \Pr_{(X,Y)\sim\mathcal{D}}[h(X) \neq Y]$. The classifier with the lowest possible risk is called the **Bayes optimal classifier**, defined as $g_{\mathcal{D}}(x) = \arg\max_{y\in\mathcal{Y}} \eta(y|x)$.

3.1. Problem Statement and Goal

In this work, we are interested in the problem of distribution shift, in which the goal is to build a classifier with low risk over a target distribution $\mathcal{D}_t = (\mu_t, \eta_t)$, primarily using data from a source distribution, $\mathcal{D}_s = (\mu_s, \eta_s)$. We denote the Bayes risk on source and target via R_s^* and R_t^*.

The challenge in this setting is that μ_s and μ_t can put mass in drastically different regions in $\mathcal{X}$ making direct generalization from the source distribution to the target distribution difficult or impossible in the worst case. We will also assume that μ_s, μ_t both have compact support.

3.2. Feature Maps

We consider classification after first applying a transformation into a feature space $(\mathcal{Z}, d_{\mathcal{Z}})$, also a metric space.

We assume we are given Φ, a class of feature maps $\phi : \mathcal{X} \to \mathcal{Z}$. Here, each $\phi \in \Phi$ represents a potential feature map under which the source and target distributions could plausibly be connected. We also let d_ϕ denote the distance metric induced on $\mathcal{X}$ by ϕ, i.e. $d_\phi(x, x') = d_{\mathcal{Z}}(\phi(x), \phi(x'))$.

If $\mathcal{D} = (\mu, \eta)$ is a distribution over $\mathcal{X} \times \mathcal{Y}$, then any feature map ϕ induces a distribution $\mathcal{D}^\phi = (\mu^\phi, \eta^\phi)$ over $\mathcal{Z} \times \mathcal{Y}$ as follows: a sample from $\mathcal{D}^\phi$ is generated by sampling $(x, y) \sim \mathcal{D}$ and then outputting $(\phi(x), y)$. We include a rigorous definition of the measure and conditional distribution corresponding to $\mathcal{D}^\phi$ in Section A of the appendix.

We now define two important examples of classes of feature maps, linear maps and neural networks.

Definition 1 (Linear Feature Maps) *Let* $\mathrm{Lin}_{D,K}$ *denote the set of all linear maps from* $\mathbb{R}^D \to \mathbb{R}^K$ *where* $\mathbb{R}^D$ *and* $\mathbb{R}^K$ *are metrized with the Euclidean metric.*

Definition 2 (Neural Network Feature Maps) *Let* $\mathrm{Nnet}_{D,D_1,\ldots,D_l,K}$ *denote the set of all ReLu activated neural networks with input layer having size* D*, hidden layers having sizes* $D_1, \ldots, D_l$*, and output layer having size* K*.*

3.3. Nearest Neighbors

Let $\mathcal{D}$ be a data distribution, and $S \sim \mathcal{D}^n$ an i.i.d training sample of n points. We let $\mathcal{N}_S : \mathcal{X} \to \mathcal{Y}$ denote the k_n-nearest neighbor classifier arising from a sample S and a metric over the instances, where ties are broken arbitrarily. It is well known that under mild regularity conditions, $k_n/n \to 0$ and $k_n \to \infty$ imply that k_n-nearest neighbors will converge to the Bayes optimal classifier (Chaudhuri and Dasgupta, 2014). Throughout this work, we will fix k_n as any sequence that satisfies these conditions, and we also assume ties are broken arbitrarily and independently of the map ϕ (note that this holds for the tie-breaking mechanism in (Chaudhuri and Dasgupta, 2014)).

Because we consider classification in feature space, we will often consider the composition of k_n-NN with maps $\phi \in \Phi$. To this end, we let $\mathcal{N}_S^\phi : \mathcal{X} \to \mathcal{Y}$ denote the map defined by

$$\mathcal{N}_S^\phi(x) = \mathcal{N}_{\{(\phi(x),y):(x,y)\in S\}}\big(\phi(x)\big).$$

3.4. Margin Conditions

It will be useful to characterize data distributions in which Bayes optimal classification is clearly non-ambiguous, and regions in which Bayes-optimal predictions differ are separated by a margin. We formalize this as follows.

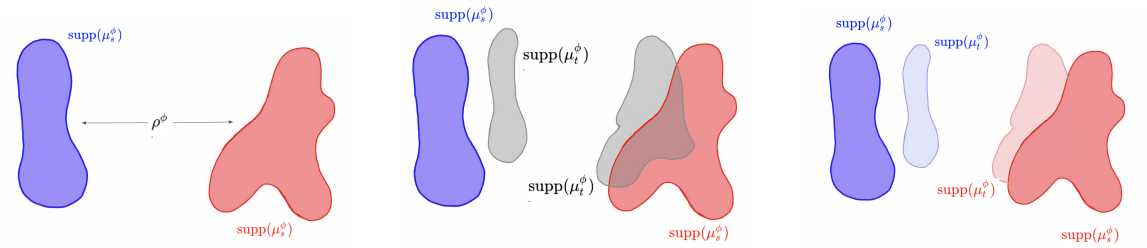
Definition 3 *A data distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$ *is* (ρ, Δ) $-$ ***separated*** *if there exist* $\rho, \Delta > 0$*, and disjoint subsets of* $\mathcal{X}$*,* $\{\mu^y : y \in \mathcal{Y}\}$*, so that the following hold:*

1. $\mathrm{supp}(\mu) = \cup_{y\in\mathcal{Y}} \mu^y$.
2. *If* $y \neq y'$*, then* $\forall x \in \mu^y$*,* $\eta(y|x) > \eta(y'|x) + \Delta$.
3. $\min_{y\neq y'} d(\mu^y, \mu^{y'}) = \rho$.

When $\mathcal{D}$ is (ρ, Δ)-separated, we say that $\mathcal{D}$ has **margin** ρ, and **label margin** Δ. For convenience, we say that $\mathcal{D}$ is **separable** if it is (ρ, Δ)-separated for some $\rho, \delta > 0$. The conditions of separable distributions are met in most practical cases, where classification is rarely ambiguous, and arbitrarily close examples are usually classified identically.

4. The Statistical IRM Assumption

Generalizing from source data in a feature space induced by some $\phi \in \Phi$ is only possible if Φ contains a map that appropriately unifies the classification tasks on $\mathcal{D}_s$ and $\mathcal{D}_t$. Invariant Risk Minimization (Arjovsky et al., 2020) proposes such a condition by positing the existence of a feature map $\psi^* : \mathcal{X} \to \mathcal{Z}$ and an "invariant predictor" $h^* : \mathcal{Z} \to \mathcal{Y}$ for which h^* is the optimal predictor on all projected training and testing environments $\phi^*(\mathcal{X})$. In this work, we will consider a similar



(*a*) Data from $\mathcal{D}_s$ is well-separated under ϕ with differently labeled points having distance at least ρ^ϕ from each other.

(*b*) Points from the target distribution lie close to regions from the source distribution relative to the margin, ρ^ϕ. This allows all target points to have a consistent labeling for their nearest neighbor in the source data.

(*c*) The correct labels of the target distribution match the labels of the corresponding nearest neighbors in the source distribution.

Figure 2: An illustration of the three conditions of Definition 4, with panels (a), (b), and (c) corresponding to conditions 1, 2, and 3 respectively.

condition that is tailored to allow for optimal prediction under an appropriate feature map using the k_n-nearest neighbors.

We begin with the following definition:

Definition 4 (Statistical IRM Assumption) *We say that $\phi \in \Phi$ satisfies the Statistical IRM assumption w.r.t. $(\mathcal{D}_s, \mathcal{D}_t)$ if the following hold:*

1. *Both $\mathcal{D}_s$ and $\mathcal{D}_s^\phi$ are separable and share the same Bayes-risk. that is,*

$$R(g_{\mathcal{D}_s^\phi}, \mathcal{D}_s^\phi) = R_s^*.$$

2. *Let ρ^ϕ denote the margin of $\mathcal{D}_s^\phi$. Then all points in the support of μ_t^ϕ have distance strictly less than $\frac{\rho^*}{2}$ from the support of μ_s^ϕ. That is,*

$$\sup_{x_t \in \text{supp}(\mu_t)} \inf_{x_s \in \text{supp}(\mu_s)} d_\phi(x_t, x_s) < \frac{\rho^\phi}{2}.$$

3. *Points from μ_t have labels that match their nearest neighbor (under ϕ) in μ_s. That is, for all $x_t \in \text{supp}(\mu_t)$,*

$$g_{\mathcal{D}_t}(x_t) = g_{\mathcal{D}_s}\left(\underset{x_s \in \text{supp}(\mathcal{D}_s)}{\arg\min}\ d_\phi(x_t, x_s)\right).$$

We let $\Phi^*(\mathcal{D}_s, \mathcal{D}_t)$ denote the set of all maps in Φ that satisfy the Statistical IRM assumption with respect to $(\mathcal{D}_s, \mathcal{D}_t)$.

Here, Condition 1 implies that the feature map ϕ must preserve all the necessary information from the input needed for accurate classification. Furthermore, ϕ induces a separable distribution $\mathcal{D}_s^\phi$, which implies that under ϕ, $\mathcal{D}_s$ is particularly suitable to classification using k_n-nearest neighbors.

Condition 2 implies that all points from the target distribution are mapped relatively closely to the source distribution, and condition 3 implies that this mapping is consistent with the correct labeling of $\mathcal{D}_t$ under the Bayes-optimal.

In essence, the three conditions together state that for an appropriate choice of ϕ, the induced Bayes-optimal classifier of $\mathcal{D}_s^\phi$ can be naturally extended to correctly classify both $\mathcal{D}_s$ and $\mathcal{D}_t$. We illustrate these conditions in panels (a), (b), and (c) respectively of Figure 2.

We now formalize our intuition that these conditions enable good out-of-distribution generalization with the following result.

Theorem 5 *Suppose $\phi^* \in \Phi^*(\mathcal{D}_s, \mathcal{D}_t)$ realizes the Statistical-IRM assumption w.r.t. $(\mathcal{D}_s, \mathcal{D}_t)$. Then applying k_n-nearest neighbors to source data after applying ϕ^* converges towards the Bayes-optimal over the target distribution. That is, for all $\epsilon, \delta > 0$, there exists N such that for all $n \geq N$, with probability $\geq 1 - \delta$ over $S \sim \mathcal{D}_s^n$,*

$$R(\mathcal{N}_S^{\phi^*}, \mathcal{D}_t) \leq R_t^* + \epsilon.$$

Note that this result crucially assumes that ϕ^* is known to the learner. This does not hold in general, and indeed one of the core challenges of this setting is *learning* such a feature map.

The main idea of proving Theorem 5 is quite straightforward – given enough labels from $\mathcal{D}_s$, for any point $x \sim \mathcal{D}_t$, its image under ϕ^* will eventaully only draw its nearest neighbors from points that have an identical Bayes-optimal labeling (Condition 3). We defer a full proof to Appendix B.

We now consider the task of learning a classifier over $\mathcal{D}_t$ using data from $\mathcal{D}_s$ under the Statistical IRM-assumption. With no restrictions on ϕ, this task is clearly intractable. Thus, we assume the learner is given prior knowledge that $\phi \in \Phi$ for some known class of feature maps Φ. Furthermore, in many cases we will require at least *some* data from $\mathcal{D}_t$. As a motivating example, recall Figure 1 where in both panels, data purely from the source distribution is insufficient for determining which feature map (X or Y projections) contains the relevant information for classification.

Thus, in our setting the learner is given access to n i.i.d samples from $\mathcal{D}_s$, m i.i.d samples from $\mathcal{D}_t$, and a class of feature maps Φ and must output a classifier that performs well over $\mathcal{D}_t$.

While a trivial solution would be to simply ignore source data and build a classifier from the target sample, we stress that in our setting target data is scare while source data is abundant reflecting typical real-life use cases. Furthermore, as we will later see, it is possible to drastically improve the performance of our classifier with just a small number of samples from $\mathcal{D}_t$.

As a simple intuition for this, observe panel (b) of Figure 1. Given enough source data, all we need is a *single point* of target data in order to infer that the X-projection is disastrous leaving the Y-projection as the only possibility. More generally, the data from $\mathcal{D}_t$ can be thought of as useful for choosing the correct feature map, while the data from $\mathcal{D}_s$ can be thought of as useful for actually building the classifier.

5. The Distance Dimension

Unfortunately, succeeding in the learning problem outlined in the previous section is not possible with no further assumptions on the feature class Φ. For example, if Φ were allowed to comprise of

all possible functions, then there would be a myriad of possibilities for ϕ that satisfy the Statistical IRM-assumption, none of which would be computable let alone learnable. To address this, we introduce a novel complexity measure of a class of feature maps called the *distance dimension.*

Definition 6 (Distance Dimension) *For $\phi : \mathcal{X} \to \mathcal{Z}$, let $\Delta_\phi : \mathcal{X}^4 \to \{0,1\}$ denote*

$$\Delta_\phi(x_1, x_2, x_3, x_4) = \mathbb{1}\left(d_{\mathcal{Z}}\left(\phi(x_1), \phi(x_2)\right) \geq d_{\mathcal{Z}}\left(\phi(x_3), \phi(x_4)\right)\right),$$

and let $\Delta\Phi = \{\Delta_\phi : \phi \in \Phi\}$. Then the distance dimension, $\partial(\Phi)$, is the VC dimension of $\Delta\Phi$.

The main idea behind the distance dimension is that it bound the number of possible ways mapping by $\phi \in \Phi$ can affect the in the distance comparisons in feature space can shake out for a given training set. As we will later see, this will play particularly nicely with nearest neighbors which is a fully comparison based algorithm.

We now show bound the distance dimension for the two examples of feature maps shown earlier.

Proposition 7 *Let $\text{Lin}_{D,K}$ be the class of linear feature maps defined in Definition 1. Then $\partial(\text{Lin}_{D,K}) \leq D^2$.*

Proof We will handle the two sets of feature maps separately.

We will construct two maps, $\alpha : \text{Lin}_{D,K} \to \mathbb{R}^{D^2}$, and $\beta : (\mathbb{R}^D)^4 \to \mathbb{R}^{D^2}$ such that for any $\phi \in \text{Lin}_{D,K}$ and $x_1, x_2, x_3, x_4 \in (\mathbb{R}^D)$,

$$\Delta\phi(x_1, x_2, x_3, x_4) = sgn\left(\langle \alpha(\phi), \beta(x_1, x_2, x_3, x_4)\rangle\right).$$

This will immediately imply the result as it is well known that the set of all linear classifiers over $\mathbb{R}^{D^2}$ has VC-dimension at most D^2.

Letting A_ϕ be the $D \times D$ matrix associated with ϕ, we have

$$\begin{aligned}
\Delta\phi(x_1, x_2, x_3, x_4) &= sgn\left(d(\phi(x_1), \phi(x_2))^2 - d(\phi(x_3), \phi(x_4))^2\right) \\
&= sgn\left(||A_\phi x_1 - A_\phi x_2||^2 - ||A_\phi x_3 - A_\phi x_4||^2\right) \\
&= sgn\left((x_1 - x_2)^t A_\phi^t A_\phi (x_1 - x_2) - (x_3 - x_4)^t A_\phi^t A_\phi (x_3 - x_4)\right) \\
&= sgn\left(\langle A_\phi^t A_\phi, (x_1 - x_2)(x_1 - x_2)^t\rangle - \langle A_\phi^t A_\phi, (x_3 - x_4)(x_3 - x_4)^t\rangle\right) \\
&= sgn\left(\langle A_\phi^t A_\phi, (x_1 - x_2)(x_1 - x_2)^t - (x_3 - x_4)(x_3 - x_4)^t\rangle\right)
\end{aligned}$$

Thus, letting $\alpha(\phi) = A_\phi^t A_\phi$ (cast as a vector in $\mathbb{R}^{D^2}$) and $\beta(x_1, x_2, x_3, x_4) = (x_1 - x_2)(x_1 - x_2)^t - (x_3 - x_4)(x_3 - x_4)^t$ suffices.

■

Proposition 8 *Let $\text{Nnet}_{D,D_1,\ldots,D_l,K}$ be the class of ReLu activated neural networks defined in Definition 2. $\partial(\text{Nnet}_{D,D_1,\ldots,D_l,K}) \leq O(w^4)$, where w denotes the total number of parameters comprising a neural network from this class.*

Proof

The key idea is to apply the proof technique used in (Karpinski and Macintyre, 1997) to analyze the VC-dimension of sigmoid activated neural networks. Their proof idea implies that if, for a family of functions $F = \{x \mapsto f(\theta, x), \theta \in \mathbb{R}^w\}$, $f(\theta, x)$ can be computed using at most t operations of basic arithmetic, inequality comparisons, and applications of the exponential function, then F has VC-dimension at most $O(t^2w^2)$.

In our case, it thus suffices to show that this holds for Δ_ϕ when $\phi \in \Phi$. Observe that Δ_ϕ is parametrized by w parameters (as ϕ has w parameters). Next, observe that outputting $\phi(x_i)$ can be done using at most $O(w)$ operations (by simply doing a forward pass on the neural network corresponding to ϕ). By doing this 4 times, and then applying a distance comparison (which takes clearly at most $O(K)$ operations where K is the size of the last layer), we see that in total we apply $O(w)$ operations. Thus, applying the result from (Karpinski and Macintyre, 1997) implies our result. ■

6. A learning rule for transfer learning

Armed with this notion of complexity, we now give a learning rule for the transfer learning setting. Our idea (Algorithm 1) is simple: given samples $S \sim \mathcal{D}_s$ And $T \sim \mathcal{D}_t$, we compute $\phi \in \Phi$ that, when coupled with k_n-nearest neighbors over S, minimizes the empirical risk over $T \sim \mathcal{D}_t^m$. As we will see, the bounded distance dimension of Φ prevents overfitting and allows us the following generalization bound.

Theorem 9 *Suppose Φ realizes the Statistical IRM assumption. Then for every $\epsilon, \delta > 0$, there exists N such that if*

$$n \geq N, m \geq \Omega\left(\frac{\partial(\Phi)\log\left(n + \partial(\Phi)\right) + \log\frac{1}{\delta}}{\epsilon^2}\right),$$

then with probability at least $1 - \delta$ over $S \sim \mathcal{D}_s^n$, $T \sim \mathcal{D}_t^m$,

$$R(\mathcal{N}_S^{\hat{\phi}}, \mathcal{D}_t) \leq R_t^* + \epsilon,$$

where $\mathcal{N}_S^{\hat{\phi}}$ is output of MAP_VALIDATE(S, T).

Thus, the SIRM assumption allows a learner to allocate the "heavy lifting" of classifier construction to the source distribution. The amount of labeled target data required for generalization can be largely controlled through the distance dimension. We say "largely" given that m, the amount of data required from $\mathcal{D}_t$, has a logarithmic dependence on n, the amount of data drawn from $\mathcal{D}_s$. This implies a near distributional-independence between source and target in the sample complexity.

Theorem 9 implies *significant* savings in the amount of target data required for good generalization. A direct application of k_n-nearest neighbors to the target distribution would demand enough samples to cover its support—a quantity often exponential in the intrinsic dimension. By contrast, Theorem 9 shows that if a comparable amount of *source* data is available, then only a relatively small number of additional target samples are needed.

Observe that k_n-nearest neighbors typically requires data sufficient to "reasonably cover" the support of the distribution, which for many distributions scales exponentially with the intrinsic dimension. Theorem 9 yields savings in two distinct ways:

Algorithm 1 Transfer Learning through Feature Selection: Target Loss Validation

procedure MAP_VALIDATE($S \sim \mathcal{D}_s^n, T \sim \mathcal{D}_t^m$)

  $\hat{\phi} = \arg\min_{\phi \in \Phi} \frac{1}{m} \sum_{(x,y) \in T} \mathbb{1}\left[\mathcal{N}_S^{\phi} \neq y\right]$

  return $\mathcal{N}_S^{\hat{\phi}}$

end procedure

1. By applying nearest neighbors in feature space, we can work in a potentially much lower dimension than the intrinsic dimension of the data in the original space.

2. The number of target samples needed grows only logarithmically in the amount of source data required to cover the source distribution.

Thus, rather than needing enough target data to cover the entire target support, it suffices to use an amount proportional to the logarithm of the source data required to cover the source support.

Proof of Theorem 9 Let ϕ^* satisfy the Statistical-IRM assumption.

$$E_1 = \mathbb{1}\left(R(\mathcal{N}_S^{\phi^*}, \mathcal{D}_t) - R(g_{\mathcal{D}_t}, \mathcal{D}_t) < \frac{\epsilon}{2}\right),$$

and

$$E_2 = \mathbb{1}\left(\sup_{\phi \in \Phi} \left| R(\mathcal{N}_\phi^S, \mathcal{D}_t) - \frac{1}{m} \sum_{(x,y) \in T} \mathbb{1}\left(\mathcal{N}_\phi^S(x) \neq y\right) \right| < \frac{\epsilon}{2}\right).$$

Intuitively, E_1 is the event that k_n-nearest neighbors performs well over $\mathcal{D}_t$ when composed with the *correct* projection, $\phi^* \in \Phi$, whereas E_2 is the event that empirical risks of such classifiers over T are representative of their true risk.

Our goal is to show that E_1 and E_2 jointly hold with probability at least $1-\delta$ – this would imply that our learned classifier has risk at most $R(g_{\mathcal{D}_t}, \mathcal{D}_t) + \epsilon$, as desired. By Theorem 5, E_1 holds with probability at least $1 - \frac{\delta}{2}$, so it suffices to show that E_2 holds with probability at least $1 - \frac{\delta}{2}$ as well.

Fix any set of n points, $\hat{S}$. It suffices to show that $\Pr_{T \sim \mathcal{D}_t^m}[E_2 = 1 | S = \hat{S}] \geq 1 - \delta$, as integrating over all possibilities of S would give the desired result. Consider the hypothesis class, $H_{\hat{S}} : \{h_\phi : \phi \in \Phi\}$ where $h_\phi : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ is defined as

$$h_\phi(x, y) = \mathbb{1}\left(\mathcal{N}_{\phi(\hat{S})}(x) \neq y\right).$$

Observe that $h \in H_{\hat{S}}$ is a binary classifier over its domain. It follows that given $S = \hat{S}$,

$$\begin{aligned} E_2 &= \mathbb{1}\left(\sup_{\phi \in \Phi} \left| R(\mathcal{N}_\phi^S, \mathcal{D}_t) - \frac{1}{m} \sum_{(x,y) \in T} \mathbb{1}\left(\mathcal{N}_\phi^S(x) \neq y\right) \right| < \frac{\epsilon}{2}\right) \\ &= \mathbb{1}\left(\sup_{h \in H_{\hat{S}}} \left| \mathbb{E}_{(x,y) \sim \mathcal{D}_t}[h(x, y)] - \frac{1}{m} \sum_{(x,y) \in T} h(x, y) \right| < \frac{\epsilon}{2}\right). \end{aligned}$$

To analyze the latter quantity, it suffices to show that $vc(H_{\hat{S}}) \leq O\left(\partial(\Phi)\log\left(n+\partial(\Phi)\right)\right)$, as standard application of the fundamental theorem of statistical learning (see Shavel-Schwartz and Ben-David) would imply that E_2 holds with probability $1-\frac{\delta}{2}$ provided that $m \geq \Omega\left(\frac{vc(H_{\hat{S}})+\ln\frac{1}{\delta}}{\epsilon^2}\right)$.

To this end, suppose,$H_{\hat{S}}$ shatters a set V of v points in $\mathcal{X}\times\mathcal{Y}$, $V=\{(x_1,y_1),\ldots,(x_v,y_v)\}$. Let $\hat{S}=\{(x'_1,y'_1),\ldots,(x'_n,y'_n)\}$. The key observation is that for any $h_\phi \in H_{\hat{S}}$, the way h_ϕ labels a given point (x,y) is determined by the k_n-nearest neighbors of $\phi(x)$ in $\{\phi(x_1),\ldots,\phi(x_n)\}$. Furthermore, these labels are full determined by the set of all $\binom{n}{2}$ comparisons,

$$\left\{\mathbb{1}\left(d(\phi(x,x_i)\geq d(\phi(x,x_j)\right):1\leq i<j\leq n\right\}.$$

This is precisely the definition of a distance comparer (Definition 6). It follows that the number of distinct ways that $H_{\hat{S}}$ can label V is at most the number of ways $\Delta\Phi$ can label all $v\binom{n}{2}$ possible comparisons, $\{(x_i,x_j,x_k):1\leq i\leq v, 1\leq j<k\leq n\}$. Since by definition, $vc(\Delta\Phi)=\partial(\Phi)$, By Sauer's Lemma, the number of ways $H_{\hat{S}}$ can label V is at most $\left(v\binom{n}{2}\right)^{\partial(\Phi)}$. However, since $H_{\hat{S}}$ shatters V, there exist precisely 2^v such labelings. It follows that $v\leq\log\left(\left(v\binom{n}{2}\right)^{\partial(\Phi)}\right)$. From here, straightforward algebra implies that $v=O\left(\partial(\Phi)\log\left(n+\partial(\Phi)\right)\right)$, as desired. ■

Acknowledgments

We would like to thank National Science Foundation NSF (CIF-2402817, CNS-1804829), SaTC-2241100, CCF-2217058, ARO-MURI (W911NF2110317), and ONR under N00014-24-1-2304 for research support. Robi Bhattacharjee would like to additionally thank the German Research Foundation through the Cluster of Excellence "Machine Learning - New Perspectives for Science" (EXC 2064/1 number 390727645) and the Carl Zeiss Foundation through the CZS Center for AI and Law.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Alexis Bellot and Mihaela van der Schaar. Generalization and invariances in the presence of unobserved confounding. *stat*, 1050:6, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 05 2010. doi: 10.1007/s10994-009-5152-4.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Bin Cao, Sinno Jialin Pan, Yu Zhang, Dit-Yan Yeung, and Qiang Yang. Adaptive transfer learning. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 407–412, 2010.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3437–3445, 2014.

Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. ICML'11, 2011.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *CoRR*, abs/2002.04747, 2020. URL https://arxiv.org/abs/2002.04747.

Steve Hanneke, Samory Kpotufe, and Yasaman Mahdaviyeh. Limits of model selection under transfer learning. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5781–5812. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/hanneke23c.html.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *J. Comput. Syst. Sci.*, 54(1):169–176, 1997. doi: 10.1006/JCSS.1997.1477. URL https://doi.org/10.1006/jcss.1997.1477.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886. PMLR, 2018.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives, 2022.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem. *arXiv preprint arXiv:2003.10443*, 1, 2020.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the renyi divergence, 2012.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning, 2016.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation, 2013.

A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip H. S. Torr, and Atılım Güneş Baydin. Kl guided domain adaptation, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors, 2022.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

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

Ziqiao Wang and Yongyi Mao. Information-theoretic analysis of unsupervised domain adaptation, 2023.

Austin Watkins, Enayat Ullah, Thanh Nguyen-Tang, and Raman Arora. Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36:2207–2251, 2023.

Han Zhao, Shanghang Zhang, Guanhang Wu, João P. Costeira, José M. F. Moura, and Geoffrey J. Gordon. Multiple source domain adaptation with adversarial training of neural networks, 2017.

Appendix A. Induced Distributions

In this section, we rigorously define the conditional data distribution of $\mathcal{D}^\phi$. Recall that if $(X, Y) \sim \mathcal{D}$ denote the random variables corresponding to $\mathcal{D}$, then $\mathcal{D}^\phi$ is defined as the data distribution $(\phi(X), Y)$, where $\phi : \mathcal{X} \to \mathcal{Z}$ is a feature map. We write $\mathcal{D} = (\mu, \eta)$, where μ denotes the measure corresponding to X over $\mathcal{X}$, and η is the conditional data distribution, $p(y|X)$. Our goal in this section is to similarly write $\mathcal{D}^\phi = (\mu^\phi, \eta^\phi)$.

First, observe that for any measurable subset $B \subseteq \mathcal{Z}$, $\mu^\phi(B) = \mu\left(\phi^{-1}(B)\right)$. This directly follows from the definition of the random variable $\phi(X)$.

Next, to define η^ϕ, first recall that $\eta(y|x)$ denotes the probability that $Y = y$ given that $X = x$. By assumption this is well defined for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and moreover for any $y \in \mathcal{Y}$ the function $\mathcal{X} \to [0, 1]$ defined by $x \mapsto \eta(y|x)$ is measurable. To define η^ϕ, we first define υ^y for all $y \in \mathcal{Y}$ as follows.

Definition 10 *υ^y is a measure over $\mathcal{Z}$ so that for all measurable sets B,*

$$\upsilon^y(B) = \int_{\phi^{-1}(B)} \eta(y|x) d\mu(x).$$

The fact that υ^y is a well-defined measure follows directly from the rules of integration. In essence, $\upsilon^y(B)$ is the probability. of observing (X, Y) with $\phi(X) \in B$ and $Y = y$. We now show the following:

Lemma 11 *υ^y is absolutely continuous with respect to μ^ϕ for all y*

Proof This immediately follows from the fact that $\eta(y|x) \leq 1$ for all y, x. Thus for any measurable set B,

$$\upsilon^y(B) = \int_{\phi^{-1}(B)} \eta(y|x) d\mu(x) \leq \int_{\phi^{-1}(B)} d\mu(x) = \mu(\phi^{-1}(B)) = \mu^\phi(B).$$

Thus for any $\epsilon > 0$, we can simply choose $\delta = \epsilon$ so that $\mu^\phi(B) < \delta \implies \upsilon^y(B) < \epsilon$. ■

We now use the Radon-Nikoym theorem on υ^y to define η^ϕ.

Lemma 12 *For all $y \in \mathcal{Y}$, there exists a measurable function $f^y : \mathcal{Z} \to [0, 1]$ such that*

$$\upsilon^y(B) = \int_B f^y(z) d\mu^\phi(z),$$

for all measurable sets B.

Proof This directly follows from the Radon-Nikoym theorem. ■

We then define η^ϕ using these functions, f^y.

Definition 13 *For all $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, we define $\eta^\phi(y|z) = f^y(z)$.*

Appendix B. Proof of Theorem 5

First, we characterize areas of $\mathcal{X}$ that are likely to be correctly classified by composing nearest neighbors with ϕ.

Definition 14 *Let ϕ be a feature map that preserves $\mathcal{D}_s$. Let $0 < p < 1$, and let $r > 0$ be a distance. We let $\mathcal{X}^{\phi}_{p,r}$ denote the set of all points x such that there exists x' for which the following hold.*

1. $d_{\phi}(x, x') < \frac{\rho^{\phi}}{2} - \frac{r}{2}$.
2. $\mu_s\left(B_{\phi}(x', r)\right) \geq p$.

Here, ρ^{ϕ} denotes the margin of $\mathcal{D}^{\phi}_s$, p represents a small amount of mass that must be close to x, and x' and r determine a region in which that mass is concentrated. The idea will be that x can be accurately classified using points sampled from $B(x', r)$. We now formalize this with the following lemma.

Lemma 15 *Suppose that $\phi \in \Phi$ source preserves $\mathcal{D}_s$, and let Δ denote the label margin of $\mathcal{D}_s$. Let $p, r > 0$, and let $x \in \mathcal{X}^{\phi}_{p,r}$ be an arbitrary point. For all $\delta > 0$, if*

$$\frac{np}{2} > k_n > \max\left(\frac{\log\frac{2}{\delta}}{p}, \frac{2\log\frac{2|\mathcal{Y}|}{\delta}}{(\Delta)^2}\right),$$

then with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n_s$,

$$\mathcal{N}^{\phi}_S(x) = g_{\mathcal{D}^{\phi}_s}\left(\underset{z \in \text{supp}(\mathcal{D}^{\phi}_s)}{\arg\min}\ d_{\mathcal{Z}}(z, \phi(x))\right).$$

Here, $\arg\min_{z \in supp(\mathcal{D}^{\phi}_s)} d_{\mathcal{Z}}(z, \phi(x))$ denotes the closest point in the support of $\mathcal{D}^{\phi}_s$ to x.

Proof Let x' be the point in $\mathcal{X}^{\phi}_{p,r}$ that corresponds to x based on Definition 14. Since ϕ source-preserves $\mathcal{D}_s$, $\mathcal{D}^{\phi}_s$ is separated, and there exist (ρ^{ϕ}, Δ) and regions $\{\mu^y_s : y \in \mathcal{Y}\}$ satisfying the conditions of Definition 3. Here note that μ^y_s denotes a subset of $\mathcal{Z}$, as we are using the separation of the induced distribution, $\mathcal{D}^{\phi}_s$. We also let $\mu^{y'}_s$ denote the region that contains x'. By a simple application of the triangle inequality, along with the definition of margin, it follows that

$$y' = g_{\mathcal{D}^{\phi}_s}\left(\underset{z \in \text{supp}(\mathcal{D}^{\phi}_s)}{\arg\min}\ d_{\mathcal{Z}}(z, \phi(x))\right).$$

Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be drawn i.i.d from $\mathcal{D}^n_s$. Observe that it is equivalent to consider S as being generated by first selecting $X = \{x_1, \ldots, x_n\} \sim \mu^n_s$, and then generating the labels $y_1, \ldots, y_n$ using the conditional distribution, $\eta_s(y|x_i)$. This point of view implies that the labels of the k_n-nearest neighbors of x are drawn independently from each other even after conditioning on X. Our strategy will be to leverage this and show that

1. The k_n nearest neighbors of x are extremely likely to all be in $(\mu_s^{\phi})^{y'}$, which is the region corresponding to y' defined in Definition 3.

2. As a consequence, the plurality of the neighbors is highly likely to match the "correct" label of x, $\mathcal{N}_S^{\phi}(x)$.

The key observation for proving the first point is that if $B_\phi(x', r)$ contains at least k_n points from X, then the k_n nearest neighbors (according to d_ϕ) of x will all be drawn from $\mu_s^{y'}$. This is a direct consequence of the triangle inequality; if any x_i satisfies $\phi(x_i) \notin \mu_s^{y'}$, then we have

$$\begin{aligned} d_\phi(x, x_i) &\geq d_\phi(x', x_i) - d_\phi(x, x') \\ &> \rho^\phi - \left(\frac{\rho^\phi}{2} - \frac{r}{2}\right) \\ &= \frac{\rho^\phi}{2} + \frac{r}{2} \\ &= \left(\frac{\rho^\phi}{2} - \frac{r}{2}\right) + r \\ &\geq d_\phi(x, x') + r \\ &\geq \sup_{x'' \in B_\phi(x', r)} d_\phi(x, x''). \end{aligned}$$

This implies that all points in $B_\phi(x', r)$ are closer to x than any points outside of $\mu_s^{y'}$.

Based on this, we now bound the probability of observing at least k_n points from $B_\phi(x', r)$. By Definition 14, $B_\phi(x', r)$ has probability mass at least q under $\mathcal{D}_s$. Applying Hoeffding's inequality and noting that $\frac{np}{2} > k_n > \frac{\log \frac{2}{\delta}}{p}$ and, we see that

$$\begin{aligned} \Pr\left[|X \cap B_\phi(x', r)| > k_n\right] &\geq \Pr\left[|X \cap B_\phi(x', r)| > \frac{np}{2}\right] \\ &\geq 1 - \exp\left(-\frac{n^2p^2}{2n}\right) \\ &\geq 1 - \exp\left(-k_n p\right) \\ &\geq 1 - \frac{\delta}{2}. \end{aligned}$$

Next, suppose that this event occurs. We now select the labels for our points. As per our discussion above, the labels of each nearest neighbor of x are selected independently based on the condition distribution, η_s. Let the label of the ith nearest neighbor of x be denoted as y_i.

For all $y \neq y'$, define

$$J_i^y = \begin{cases} 1 & y_i = y' \\ -1 & y_i = y \\ 0 & \text{otherwise} \end{cases}.$$

It follows that $\mathcal{N}_S^{\phi}(x) = y$ if and only if $\sum_{i=1}^{k_n} J_i^y > 0$ for all $y \neq y'$ as this will imply that y' is the plurality choice.

Recall that $\mathcal{D}_s^\phi$ is separated with label margin Δ^ϕ. Because ϕ source-preserves $\mathcal{D}_s$, it follows that $\Delta^\phi \geq \Delta$ – otherwise this would imply that points from μ_s are misclassified by the Bayes-optimal over μ^ϕ . Since we are conditioning on $|X \cap B_\phi(x', r)| > k_n$, it follows that all k_n nearest neighbors are from the regions $\mu_s^{y'}$. Thus, the definition of label margin, it follows that J_i^y is a random variable bounded in $[-1, 1]$ that has expected value at least Δ. It follows by Hoeffding's inequality, that

$$\begin{aligned}
\Pr[\sum_{i=1}^{k_n} J_i^{y'} > 0] &\geq 1 - \exp\left(\frac{-2(\Delta)^2 k_n^2}{4k_n}\right) \\
&= 1 - \exp\left(\frac{-(\Delta)^2 k_n}{2}\right) \\
&\geq 1 - \exp\left(-\log\frac{2|\mathcal{Y}|}{\delta}\right) \\
&= 1 - \frac{\delta}{2|\mathcal{Y}|}.
\end{aligned}$$

Here, the last inequality holds because $k_n > \frac{2\log\frac{2|\mathcal{Y}|}{\delta}}{(\Delta)^2}$. Thus taking a union bound over all $y \in \mathcal{Y} \setminus \{y\}$ along with including the probability that $|X \cap B_\phi(x', r)| > k_n$ in the first place, we see that with probability $1 - \delta$,

$$\mathcal{N}_S^\phi(x) = y' = g_{\mathcal{D}_s^\phi}\left(\operatorname*{arg\,min}_{z \in \text{supp}(\mathcal{D}_s^\phi)} d_{\mathcal{Z}}(z, \phi(x))\right),$$

as desired. ■

Next, we show that if ϕ satisfies the Statistical-IRM assumption, then $\mathcal{D}_t$ is covered by $\mathcal{X}_{p,r}^\phi$ for appropriate choices of p, r. For technical reasons, we will compute precise bounds on p and r based on the following definition.

Definition 16 *Suppose $\phi \in \Phi$ satisfies the Statistical-IRM assumption w.r.t. $(\mathcal{D}_s, \mathcal{D}_t)$. Let ρ^ϕ denote the margin of $\mathcal{D}_s^\phi$. We say that ϕ has **source-target margin** α^ϕ, where*

$$\alpha^\phi = \rho^\phi - 2\left(\sup_{x_t \in \text{supp}(\mu_t)} \inf_{x_s \in \text{supp}(\mu_s)} d_\phi(x_t, x_s)\right).$$

The source-target margin can be thought of as a measure of how much slack is left over within the margin of $\mathcal{D}_s^\phi$ after it engulfs the entire target distribution, $\mathcal{D}_t^\phi$. Note that it must always be strictly positive by the definition of the Statistical-IRM assumption. We now show how to cover $\mathcal{D}_t$ using a set $\mathcal{X}_{p,r}^\phi$.

Lemma 17 *Suppose $\phi \in \Phi$ satisfies the Statistical-IRM assumption with respect to $(\mathcal{D}_s, \mathcal{D}_t)$ and has source-target margin α^ϕ. Define r, p with*

1. $r = 0.99\alpha^\phi$.

2. $p = \inf_{x' \in supp(\mu_s)} \mu_s\left(B_{\mathcal{X}}\left(x', \frac{r}{L}\right)\right)$ *where* L *is the global Lipschitz factor of* Φ.

Then $r, p > 0$, *and* $\text{supp}(\mu_t) \subseteq \mathcal{X}^{\phi}_{p,r}$.

Proof

The fact that $r > 0$ immediately follows from Definition 16 along with the definition of the Statistical-IRM assumption.

To show $p > 0$, recall that $\text{supp}(\mu_s)$ is compact by assumption. Take an open cover of μ_s by balls of radius $\frac{r}{2L}$. Then it has a finite sub-cover. Each of these balls have positive mass under μ_s, and furthermore every ball $B_{\phi}\left(x, \frac{r}{L}\right)$ where $x \in \text{supp}(\mu_s)$ must fully contain at least one of these balls. Thus $\mu_s(B(x, \frac{r}{L})) \geq q$, where $q > 0$ is the minimum mass of one of these balls. Since $q > 0$, it follows that $p > 0$, as desired.

Finally, we show that $\text{supp}(\mu_t) \subset \mathcal{X}^{\phi}_{p,r}$. Let $x \in \text{supp}(\mu_t)$. Because μ_s has compact support, there exists $x' \in \text{supp}(\mu_s)$ such that

$$d_{\phi}(x, x') = \inf_{x_s \in \text{supp}(\mu_s)} d_{\phi}(x, x_s).$$

It follows that

$$\begin{aligned} d_{\phi}(x, x') &\leq \sup_{x_t \in \text{supp}(\mu_t)} \inf_{x_s \in \text{supp}(\mu_s)} d_{\phi}(x_t, x_s) \\ &= \frac{\rho^{\phi} - \alpha^{\phi}}{2} \\ &< \frac{\rho^{\phi} - r}{2}. \end{aligned}$$

Since $x' \in \text{supp}(\mu_s)$, and since Φ has Lipschitz factor L, it follows that

$$\mu_s\left(B_{\phi}(x', r)\right) \geq \mu_s\left(B_{\mathcal{X}}\left(x', \frac{r}{L}\right)\right) \geq p.$$

Thus we have established the two criterion of Definition 14, which completes the proof. ■

We are now prepared to prove Theorem 5. To do so, we prove a more precise statement that clearly immediately implies the theorem.

Lemma 18 *Suppose* ϕ *realizes the Statistical IRM assumption with respect to* $(\mathcal{D}_s, \mathcal{D}_t)$ *and has source-target margin* α^{ϕ}. *Let* Δ *denote the label margin of* $\mathcal{D}_s$, *and let*

$$p = \inf_{x \in \text{supp}(\mu_s)} \mu_s\left(B_{\mathcal{X}}\left(x, \frac{0.99\alpha^{\phi}}{L}\right)\right),$$

and suppose that n *satisfies*

1. $\frac{k_n}{n} < \frac{p}{2}$.
2. $k_n > \max\left(\frac{8}{p^3}, \frac{2(\log|\mathcal{Y}|+2)}{\Delta^3}\right)$.

Then with probability at least $1 - \exp\left(-\frac{k_n^{2/3}}{2}\right)$ *over* $S \sim \mu_s^n$,

$$R(\mathcal{N}_S^\phi, \mathcal{D}_t) \leq R_t^* + \exp\left(-\frac{k_n^{2/3}}{2}\right).$$

Proof Let $\delta = \exp(-k_n^{2/3})$. It follows that conditions 1. and 2. imply that

$$\frac{np}{2} > k_n > \max\left(\frac{\log\frac{2}{\delta}}{p}, \frac{2\log\frac{2|\mathcal{Y}|}{\delta}}{(\Delta)^2}\right).$$

Lemma 17 implies that $\text{supp}(\mu_t) \subseteq \mathcal{X}^\phi_{p,0.99\alpha^\phi}$. Thus, by Lemma 15, we have

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}_s^n}[R(\mathcal{N}_S^\phi, \mathcal{D}_t)] &= \mathbb{E}_{S\sim\mathcal{D}_s^n}\mathbb{E}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}\left(\mathcal{N}_S^\phi(x_t) \neq y_t\right)\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}_s^n}\mathbb{E}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}\left(\mathcal{N}_S^\phi(x_t) \neq g_{\mathcal{D}_t}(x_t)\right) + \mathbb{1}\left(g_{\mathcal{D}_t}(x_t) \neq y_t\right)\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}_s^n}\mathbb{E}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}\left(\mathcal{N}_S^\phi(x_t) \neq g_{\mathcal{D}_t}(x_t)\right)\right] + \mathbb{E}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}\left(g_{\mathcal{D}_t}(x_t) \neq y_t\right)\right] \\
&\leq \delta + R_t^*.
\end{aligned}$$

It follows by Markov's inequality that

$$\begin{aligned}
\Pr[R(\mathcal{N}_S^\phi, \mathcal{D}_t) - R_t^* > \sqrt{\delta}] &\leq \frac{\mathbb{E}_{S\sim\mathcal{D}_s^n}[R(\mathcal{N}_S^\phi, \mathcal{D}_t)]}{\sqrt{\delta}} \\
&\leq \frac{\delta}{\sqrt{\delta}} = \sqrt{\delta}.
\end{aligned}$$

Here we are using the fact that R_t^* is the minimum possible risk incurred by any classifier over $\mathcal{D}_t$. Our claim then follows by substituting the value of δ.

■

We now prove Theorem 5.

Proof (Theorem 5) Fix $\epsilon, \delta > 0$. Let $\Delta > 0$ be the label margin of $\mathcal{D}_s$, α^ϕ denote the source-target margin of ϕ, and

$$p = \inf_{x\in\text{supp}(\mu_s)} \mu_s\left(B_\mathcal{X}\left(x, \frac{0.99\alpha^\phi}{L}\right)\right).$$

Lemma 17 implies that $p > 0$. Since $k_n \to \infty$ and $\frac{k_n}{n} \to 0$, there exists N such that for all $n > N$,

1. $\exp\left(-\frac{k_n^{2/3}}{2}\right) < \epsilon, \delta$,
2. $\frac{k_n}{n} < \frac{p}{2}$.
3. $k_n > \max\left(\frac{\log\frac{2}{\delta}}{p}, \frac{2\log\frac{2|cY|}{\delta}}{(\Delta)^2}\right)$.

Choose any such n. It follows that with probability at least $1 - \exp\left(-\frac{k_n^{2/3}}{2}\right)$ over $S \sim \mathcal{D}_s^n$, $R(\mathcal{N}_S^\phi, \mathcal{D}_t) \leq R_t^* + \exp\left(-\frac{\sqrt{k_n}}{2}\right)$. Since $\exp\left(-\frac{k_n^{2/3}}{2}\right) < \epsilon, \delta$, the result follows. ■