Domain Generalization by Marginal Transfer Learning

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, Clayton Scott

Journal of Machine Learning Research 22 (2021) 1-55 Submitted 11/17; Published 1/21

Domain Generalization by Marginal Transfer Learning

Gilles Blanchard BLANCHARD@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay

Aniket Anand Deshmukh ANIKETDE@UMICH.EDU
Microsoft AI & Research

Ürun Dogan URUNDOGAN@GMAIL.COM
Microsoft AI & Research

Gyemin Lee GYEMIN@SEOULTECH.AC.KR
Dept. Electronic and IT Media Engineering
Seoul National University of Science and Technology

Clayton Scott CLAYSCOT@UMICH.EDU
Electrical and Computer Engineering, Statistics
University of Michigan

Editor: Arthur Gretton

Abstract

In the problem of domain generalization (DG), there are labeled training data sets from several related prediction problems, and the goal is to make accurate predictions on future unlabeled data sets that are not known to the learner. This problem arises in several applications where data distributions fluctuate because of environmental, technical, or other sources of variation. We introduce a formal framework for DG, and argue that it can be viewed as a kind of supervised learning problem by augmenting the original feature space with the marginal distribution of feature vectors. While our framework has several connections to conventional analysis of supervised learning algorithms, several unique aspects of DG require new methods of analysis.

This work lays the learning theoretic foundations of domain generalization, building on our earlier conference paper where the problem of DG was introduced (Blanchard et al., 2011). We present two formal models of data generation, corresponding notions of risk, and distribution-free generalization error analysis. By focusing our attention on kernel methods, we also provide more quantitative results and a universally consistent algorithm. An efficient implementation is provided for this algorithm, which is experimentally compared to a pooling strategy on one synthetic and three real-world data sets.

Keywords: domain generalization, generalization error bounds, Rademacher complexity, kernel methods, universal consistency, kernel approximation

1. Introduction

Domain generalization (DG) is a machine learning problem where the learner has access to labeled training data sets from several related prediction problems, and must generalize to a future prediction problem for which no labeled data are available. In more detail, there are N labeled training data sets $S_i = (X_{ij}, Y_{ij})_{1 \le j \le n_i}$, $i = 1, \ldots, N$, that describe similar but possibly distinct prediction tasks. The objective is to learn a rule that takes as input

a previously unseen *unlabeled* test data set $X_1^T, \ldots, X_{n_T}^T$, and accurately predicts outcomes for these or possibly other unlabeled points drawn from the associated learning task.

DG arises in several applications. One prominent example is precision medicine, where a common objective is to design a patient-specific classifier (e.g., of health status) based on clinical measurements, such as an electrocardiogram or electroencephalogram. In such measurements, patient-to-patient variation is common, arising from biological variations between patients, or technical or environmental factors influencing data acquisition. Because of patient-to-patient variation, a classifier that is trained on data from one patient may not be well matched to another patient. In this context, domain generalization enables the transfer of knowledge from historical patients (for whom labeled data are available) to a new patient without the need to acquire training labels for that patient. A detailed example in the context of flow cytometry is given below.

We view domain generalization as a conventional supervised learning problem where the original feature space is augmented to include the marginal distribution generating the features. We refer to this reframing of DG as "marginal transfer learning," because it reflects the fact that in DG, information about the test task must be drawn from that task's marginal feature distribution. Leveraging this perspective, we formulate two statistical frameworks for analyzing DG. The first framework allows the observations within each data set to have arbitrary dependency structure, and makes connections to the literature on Campbell measures and structured prediction. The second framework is a special case of the first, assuming the data points are drawn i.i.d. within each task, and allows for a more refined risk analysis.

We further develop a distribution-free kernel machine that employs a kernel on the aforementioned augmented feature space. Our methodology is shown to yield a universally consistent learning procedure under both statistical frameworks, meaning that the domain generalization risk tends to the best possible value as the relevant sample sizes tend infinity, with no assumptions on the data generating distributions. Although DG may be viewed as a conventional supervised learning problem on an augmented feature space, the analysis is nontrivial owing to unique aspects of the sampling plans and risks. We offer a computationally efficient and freely available implementation of our algorithm, and present a thorough experimental study validating the proposed approach on one synthetic and three real-world data sets, including comparisons to a simple pooling approach.[1]

To our knowledge, the problem of domain generalization was first proposed and studied by our earlier conference publication (Blanchard et al., 2011) which this work extends in several ways. It adds (1) a new statistical framework, the agnostic generative model described below; (2) generalization error and consistency results for the new statistical model; (3) an extensive literature review; (4) an extension to the regression setting in both theory and experiments; (5) a more general statistical analysis, in particular, we no longer assume a bounded loss, and therefore accommodate common convex losses such as the hinge and logistic losses; (6) extensive experiments (the conference paper considered a single small data set); (7) a scalable implementation based on a novel extension of random Fourier features; and (8) error analysis for the random Fourier features approximation.

1. Code is available at https://github.com/aniketde/DomainGeneralizationMarginal.

2. Motivating Application: Automatic Gating of Flow Cytometry Data

Flow cytometry is a high-throughput measurement platform that is an important clinical tool for the diagnosis of blood-related pathologies. This technology allows for quantitative analysis of individual cells from a given cell population, derived for example from a blood sample from a patient. We may think of a flow cytometry data set as a set of d-dimensional attribute vectors $(X_j)_{1\leq j\leq n}$, where n is the number of cells analyzed, and d is the number of attributes recorded per cell. These attributes pertain to various physical and chemical properties of the cell. Thus, a flow cytometry data set may be viewed as a random sample from a patient-specific distribution.

Now suppose a pathologist needs to analyze a new (test) patient with data $(X_j^T)_{1\leq j\leq n_T}$. Before proceeding, the pathologist first needs the data set to be "purified" so that only cells of a certain type are present. For example, lymphocytes are known to be relevant for the diagnosis of leukemia, whereas non-lymphocytes may potentially confound the analysis. In other words, it is necessary to determine the label $Y_j^T \in \{-1, 1\}$ associated to each cell, where $Y_j^T = 1$ indicates that the j-th cell is of the desired type.

In clinical practice this is accomplished through a manual process known as "gating." The data are visualized through a sequence of two-dimensional scatter plots, where at each stage a line segment or polygon is manually drawn to eliminate a portion of the unwanted cells. Because of the variability in flow cytometry data, this process is difficult to quantify in terms of a small subset of simple rules. Instead, it requires domain-specific knowledge and iterative refinement. Modern clinical laboratories routinely see dozens of cases per day, so it is desirable to automate this process.

Since clinical laboratories maintain historical databases, we can assume access to a number (N) of historical (training) patients that have already been expert-gated. Because of biological and technical variations in flow cytometry data, the distributions $P_{XY}^{(i)}$ of the historical patients will vary. To illustrate the flow cytometry gating problem, we use the NDD data set from the FlowCap-I challenge.[2] For example, Fig. 1 shows exemplary two-dimensional scatter plots for two different patients – see caption for details. Despite differences in the two distributions, there are also general trends that hold for all patients. Virtually every cell type of interest has a known tendency (e.g., high or low) for most measured attributes. Therefore, it is reasonable to assume that there is an underlying distribution (on distributions) governing flow cytometry data sets, that produces roughly similar distributions thereby making possible the automation of the gating process.

3. Formal Setting and General Results

In this section we formally define domain generalization via two possible data generation models together with associated notions of risk. We also provide a basic generalization error bound for the first of these data generation models.

Let $\mathcal{X}$ denote the observation space (assumed to be a Radon space) and $\mathcal{Y} \subseteq \mathbb{R}$ the output space. Let $\mathfrak{P}_{\mathcal{X}}$ and $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ denote the set of probability distributions on $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. The spaces $\mathfrak{P}_{\mathcal{X}}$ and $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ are endowed with the topology of weak convergence and the associated Borel σ-algebras.

2. We will revisit this data set in Section 8.5 where details are given.

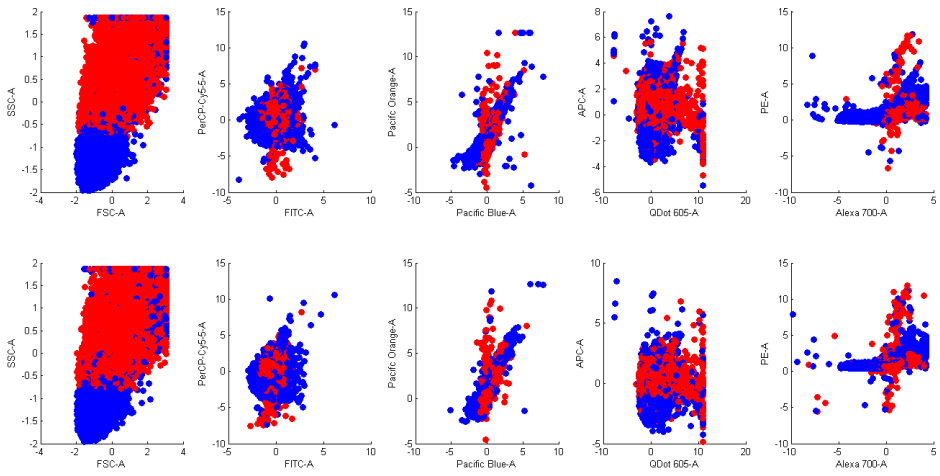

Figure 1: Two-dimensional projections of multi-dimensional flow cytometry data. Each row corresponds to a single patient, and each column to a particular two-dimensional projection. The distribution of cells differs from patient to patient. The colors indicate the results of gating, where a particular type of cell, marked dark (blue), is separated from all other cells, marked bright (red). Labels were manually selected by a domain expert.

The disintegration theorem for joint probability distributions (see for instance Kallenberg, 2002, Theorem 6.4) tells us that (under suitable regularity properties, satisfied if $\mathcal{X}$ is a Radon space) any element $P_{XY} \in \mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ can be written as a Markov semi-direct product $P_{XY} = P_X \bullet P_{Y|X}$, with $P_X \in \mathfrak{P}_{\mathcal{X}}$, $P_{Y|X} \in \mathfrak{P}_{Y|X}$, where $\mathfrak{P}_{Y|X}$ is the space of conditional probability distributions of Y given X, also called Markov transition kernels from $\mathcal{X}$ to $\mathcal{Y}$. This specifically means that

$$\mathbb{E}_{(X,Y)\sim P_{XY}}\left[h(X,Y)\right] = \int \left(\int h(x,y) P_{Y|X}(dy|X=x) \right) P_X(dx), \tag{1}$$

for any integrable function $h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$. Following common terminology in the statistical learning literature, we will also call $P_{Y|X}$ the *posterior* distribution (of Y given X).

We assume that N training samples $S_i = (X_{ij}, Y_{ij})_{1\le j\le n_i}$, $i = 1,\dots,N$, are observed. To allow for possibly unequal sample sizes n_i, it is convenient to formally identify each sample S_i with its associated empirical distribution $\widehat{P}^{(i)}_{XY} = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{(X_{ij},Y_{ij})} \in \mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$. We assume that the ordering of the observations inside a given sample S_i is arbitrary and does not contain any relevant information. We also denote by $\widehat{P}^{(i)}_X = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{ij}} \in \mathfrak{P}_{\mathcal{X}}$ the ith training sample without labels. Similarly, a test sample is denoted by $S^T = (X^T_j, Y^T_j)_{1\le j\le n_T}$, and the empirical distribution of the unlabeled data by $\widehat{P}^T_X$.

3.1 Data Generation Models

We propose two data generation models. The first is more general, and includes the second as a special case.

Assumption 1 (AGM) *There exists a distribution P_S on $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ such that $S_1, \ldots, S_N$ are i.i.d. realizations from P_S.*

We call this the *agnostic generative model*. This is a quite general model in which samples are assumed to be identically distributed and independent of each other, but nothing particular is assumed about the generation mechanism for observations inside a given sample, nor for the (random) sample size.

We also introduce a more specific generative mechanism, where observations (X_{ij}, Y_{ij}) inside the sample S_i are themselves i.i.d. from $P_{XY}^{(i)}$, a latent unobserved random distribution, as follows. The symbol $\otimes$ indicates a product measure.

Assumption 2 (2SGM) *There exists a distribution μ on $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ and a distribution ν on $\mathbb{N}$, such that $(P_{XY}^{(1)}, n_1), \ldots, (P_{XY}^{(N)}, n_N)$ are i.i.d. realizations from $\mu \otimes \nu$, and conditional to $(P_{XY}^{(i)}, n_i)$ the sample S_i is made of n_i i.i.d. realizations of (X, Y) following the distribution $P_{XY}^{(i)}$.*

This model, called the *2-stage generative model*, is a subcase of **(AGM)**: since the $(P_{XY}^{(i)}, n_i)$ are i.i.d., the samples S_i also are. This model was the one studied in our conference paper (Blanchard et al., 2011). It has been considered in the distinct but related context of "learning to learn" (Baxter, 2000; see also a more detailed discussion below, Section 4.2). Many of our results will hold for the agnostic generative model, but the two-stage generative model allows for additional developments.

Since in **(2SGM)** we assume that the latent random distribution of the points in the sample S_i and its size n_i are independent (which is not necessarily the case for **(AGM)**), in this model it becomes a formally well-defined question to ask how the learning problem evolves if we only change the size of the samples. In other words, we may study the setting where the generating distribution μ remains fixed, but their size distribution ν changes. In particular, this work examines the following different situations of interest in which the distribution μ is fixed:

- The samples all have the same fixed size n, i.e. $\nu = \delta_n$;
- The training samples are subsampled (without replacement) to a fixed size n in order to reduce computational complexity; this reduces to the first setting;
- Both the training samples N and their size grow. In this case the size distribution ν_N depends on N (possibly $\nu_N = \delta_{n(N)}$)

We note that when the distribution of the sample sizes n_i is a Poisson or a mixture of Poisson distributions, the **(2SGM)** is (a particular case of) what is known as a Cox model or doubly stochastic Poisson process in the point process literature (see, e.g., Daley and Vere-Jones, 2003, Section 6.2), which is a Poisson process with random (inhomogeneous) intensity.

3.2 Decision Functions and Augmented Feature Space

In domain generalization, the learner's goal is to infer from the training data a general rule that takes an arbitrary, previously unseen, unlabeled data set corresponding to a new prediction task, and produces a classifier for that prediction task that could be applied to any x (possibly outside the unlabeled data set). In other words, the learner should output a mapping $g : \mathfrak{P}_{\mathcal{X}} \to (\mathcal{X} \to \mathbb{R})$. Equivalently, the learner should output a function $f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$, where the two notations are related via $g(P_X)(x) = f(P_X, x)$. In the latter viewpoint, f may be viewed as a standard decision function on the "augmented" or "extended" feature space $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$, which facilitates connections to standard supervised learning. We refer to this view of DG as *marginal transfer learning*, because the information that facilitates generalization to a new task is conveyed entirely through the marginal distribution. In the next two subsections, we present two definitions of the risk of a decision function f, one associated to each of the two data generation models.

3.3 Risk and Generalization Error Bound under the Agnostic Generative Model

Consider a test sample $S^T = (X_j^T, Y_j^T)_{1 \le j \le n_T}$, whose labels are not observed. If $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is a loss function for a single prediction, and predictions of a fixed decision function f on the test sample are given by $\widehat{Y}_j^T = f(\widehat{P}_X^T, X_j^T)$, then the empirical average loss incurred on the test sample is

$$\mathcal{L}(S^T, f) := \frac{1}{n_T} \sum_{j=1}^{n_T} \ell(\widehat{Y}_j^T, Y_j^T) \, .$$

Thus, we define the *risk* of a decision function as the average of the above quantity when test samples are drawn according to the same mechanism as the training samples:

$$\mathcal{E}(f) := \mathbb{E}_{S^T \sim P_S} \left[\mathcal{L}(S^T, f) \right] = \mathbb{E}_{S^T \sim P_S} \left[\frac{1}{n_T} \sum_{j=1}^{n_T} \ell(f(\widehat{P}_X^T, X_j^T), Y_j^T) \right] .$$

In a similar way, we define the *empirical risk* of a decision function as its average prediction error over the training samples:

$$\widehat{\mathcal{E}}(f, N) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(S_i, f) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f(\widehat{P}_X^{(i)}, X_{ij}), Y_{ij}). \tag{2}$$

Remark 3 *It is possible to understand the above setting as a particular instance of a* structured output learning problem *(Tsochantaridis et al., 2005; Bakır et al., 2007), in which the input variable X^* is $\widehat{P}_X^T$, and the "structured output" Y^* is the collection of labels $(Y_i^T)_{1 \le i \le n_T}$ (matched to their respective input points). As is generally the case for structured output learning, the nature of the problem and the "structure" of the outputs is very much encoded in the particular form of the loss function. In our setting the loss function is additive over the labels forming the collection Y^*, and we will exploit this particular form for our method and analysis.*

Remark 4 *The risk $\mathcal{E}(f)$ defined above can be described in the following way: consider the random variable $\xi := (\widehat{P}_{XY}; (X, Y))$ obtained by first drawing $\widehat{P}_{XY}$ according to P_S, then, conditional to this, drawing (X, Y) according to $\widehat{P}_{XY}$. The risk is then the expectation of a certain function of ξ (namely $F_f(\xi) = \ell(f(\widehat{P}_X, X), Y)$). In probability theory literature, the distribution of the variable ξ is known as the* Campbell measure *associated to the distribution P_S over the measure space $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$; this object is in particular of fundamental use in point process theory (see, e.g., Daley and Vere-Jones, 2008, Section 13.1). We will denote it by $\mathcal{C}(P_S)$ here. This intriguing connection suggests that more elaborate tools of point process literature may find their use to analyze DG when various classical point processes are considered for the generating distribution. The Campbell measure will also appear in the Rademacher analysis below.*

The next result establishes an analogue of classical Rademacher analysis under the agnostic generative model.

Theorem 5 (Uniform estimation error control under (AGM)) *Let $\mathcal{F}$ be a class of decision functions $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$. Assume the following boundedness condition holds:*

$$\sup_{f\in\mathcal{F}} \sup_{P_X\in\mathfrak{P}_{\mathcal{X}}} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(f(P_X, x), y) \leq B_\ell. \tag{3}$$

Under **(AGM)**, *if $S_1, \ldots, S_N$ are i.i.d. realizations from P_S, then with probability at least $1 - \delta$ with respect to the draws of the training samples:*

$$\begin{aligned}&\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{E}}(f, N) - \mathcal{E}(f)\right| \\ &\leq \frac{2}{N}\mathbb{E}_{(\widehat{P}^{(i)}_{XY};(X_i,Y_i))\sim\mathcal{C}(P_S)^{\otimes N}}\mathbb{E}_{(\varepsilon_i)_{1\leq i\leq N}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^N \varepsilon_i \ell(f(\widehat{P}^{(i)}_X, X_i), Y_i)\right|\right] + B_\ell\sqrt{\frac{\log(\delta^{-1})}{2N}},\end{aligned} \tag{4}$$

where $(\varepsilon_i)_{1\leq i\leq N}$ are i.i.d. Rademacher variables, independent from $(\widehat{P}^{(i)}_{XY}, (X_i, Y_i))_{1\leq i\leq N}$, and $\mathcal{C}(P_S)$ is the Campbell measure on $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}} \times (\mathcal{X} \times \mathcal{Y})$ associated to P_S (see Remark 4).

Proof Since the $(S_i)_{1\leq i\leq N}$ are i.i.d., $\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{E}}(f, N) - \mathcal{E}(f)\right|$ takes the form of a uniform deviation between average and expected loss over the function class $\mathcal{F}$. We can therefore apply standard analysis (Azuma-McDiarmid inequality followed by Rademacher complexity analysis for a nonnegative bounded loss; see, e.g., Koltchinskii, 2001; Bartlett and Mendelson, 2002, Theorem 8) to obtain that with probability at least $1 - \delta$ with respect to the draw of the training samples $(S_i)_{1\leq i\leq N}$:

$$\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{E}}(f, N) - \mathcal{E}(f)\right| \leq \frac{2}{N}\mathbb{E}_{(S_i)_{1\leq i\leq N}}\mathbb{E}_{(\varepsilon_i)_{1\leq i\leq N}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^N \varepsilon_i \mathcal{L}(S_i, f)\right|\right] + B_\ell\sqrt{\frac{\log(\delta^{-1})}{2N}},$$

where $(\varepsilon_i)_{1\leq i\leq N}$ are i.i.d. Rademacher variables, independent of $(S_i)_{1\leq i\leq N}$.

We may write

$$\mathcal{L}(S_i, f) = \frac{1}{n_i}\sum_{j=1}^{n_i} \ell(f(\widehat{P}^{(i)}_X, X_{ij}), Y_{ij}) = \mathbb{E}_{(X,Y)\sim\widehat{P}^{(i)}_{XY}}\left[\ell(f(\widehat{P}^{(i)}_X, X), Y)\right];$$

thus, we have

$$\begin{aligned}
&\mathbb{E}_{(S_i)_{1\le i\le N}}\mathbb{E}_{(\varepsilon_i)_{1\le i\le N}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{N}\varepsilon_i\mathcal{L}(S_i,f)\right|\right]\\
&\quad=\mathbb{E}_{(\widehat{P}_{XY}^{(i)})_{1\le i\le N}}\mathbb{E}_{(\varepsilon_i)_{1\le i\le N}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{N}\varepsilon_i\mathbb{E}_{(X_i,Y_i)\sim\widehat{P}_{XY}^{(i)}}\left[\ell(f(\widehat{P}_X^{(i)},X_i),Y_i)\right]\right|\right]\\
&\quad\le\mathbb{E}_{(\widehat{P}_{XY}^{(i)})_{1\le i\le N}}\mathbb{E}_{(X_1,Y_1)\sim\widehat{P}_{XY}^{(1)},\ldots,(X_N,Y_N)\sim\widehat{P}_{XY}^{(N)}}\mathbb{E}_{(\varepsilon_i)_{1\le i\le N}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{N}\varepsilon_i\ell(f(\widehat{P}_X^{(i)},X_i),Y_i)\right|\right].
\end{aligned}$$

In the above inequality, the inner expectation on the (X_i, Y_i) is pulled outwards by Jensen's inequality and convexity of the supremum.

To obtain the announced estimate, notice that the above expectation is the same as the expectation with respect to the N-fold Campbell measure $\mathcal{C}(P_S)$. ■

Remark 6 *The main term in the theorem is just the conventional Rademacher complexity for the augmented feature space $\mathfrak{P}_{\mathcal{X}}\times\mathcal{X}$ endowed with the Campbell measure $\mathcal{C}(P_S)$. It could also be thought of as the Rademacher complexity for the meta-distribution P_S.*

3.4 Idealized Risk under the 2-stage Generative Model

The additional structure of **(2SGM)** allows us to define a different notion of risk under this model. Toward this end, let P_{XY}^T denote the testing data distribution, P_X^T the marginal X-distribution of P_{XY}^T, n_T the test sample size, $S^T=(X_i^T,Y_i^T)_{1\le i\le n_T}$ the testing sample, and $\widehat{P}_X^T$ the empirical X-distribution. Parallel to the training data generating mechanism under **(2SGM)**, we assume that P_{XY}^T is drawn according to μ.

We first define the risk of any $f:\mathfrak{P}_{\mathcal{X}}\times\mathcal{X}\to\mathbb{R}$, conditioned on a test sample of size n_T, to be

$$\mathcal{E}(f|n_T):=\mathbb{E}_{P_{XY}^T\sim\mu}\mathbb{E}_{(X_i^T,Y_i^T)_{1\le i\le n_T}\sim(P_{XY}^T)^{\otimes n_T}}\left[\frac{1}{n_T}\sum_{i=1}^{n_T}\ell(f(\widehat{P}_X^T,X_i^T),Y_i^T)\right]. \tag{5}$$

In this definition, the test sample S^T consist of n_T iid draws from and random P_{XY}^T drawn from μ. This conditional risk may be viewed as the previously defined risk for **(AGM)** specialized to **(2SGM)**, where $\nu=\delta_{n_T}$.

We are particularly interested in the idealized situation where the test sample size n_T grows to infinity. By the law of large numbers, as n_T grows, $\widehat{P}_X^T$ converges to P_X^T (in the sense of weak convergence). This motivates the introduction of the following *idealized risk* which assumes access to an infinite test sample, and thus to the true marginal P_X^T:

$$\mathcal{E}^\infty(f):=\mathbb{E}_{P_{XY}^T\sim\mu}\mathbb{E}_{(X^T,Y^T)\sim P_{XY}^T}\left[\ell(f(P_X^T,X^T),Y^T)\right]. \tag{6}$$

Note that both notions of risk in (5) and (6) depends on μ, but not on ν.

The following proposition more precisely motivates viewing $\mathcal{E}^\infty(f)$ as a limiting case of $\mathcal{E}(f|n_T)$.

Proposition 7 *Assume ℓ is a bounded, L-Lipschitz loss function and $f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ is a fixed decision function which is continuous with respect to both its arguments (recalling $\mathfrak{P}_{\mathcal{X}}$ is endowed with the weak convergence topology). Then it holds under* **(2SGM)***:*

$$\lim_{n_T \to \infty} \mathcal{E}(f|n_T) = \mathcal{E}^{\infty}(f).$$

Remark 8 *This result provides one setting where the risk $\mathcal{E}^{\infty}$ is clearly motivated as the goal of asymptotic analysis when $n_T \to \infty$. Although Proposition 7 is not used elsewhere in this work, a more quantitative version of this result is stated below for kernels (see Theorem 15), where convergence holds uniformly and the assumption of a bounded loss is dropped.*

To gain more insight into the idealized risk $\mathcal{E}^{\infty}$, recalling the standard decomposition (1) of P_{XY} into the marginal P_X and the posterior $P_{Y|X}$, we observe that we can apply the disintegration theorem not only to any P_{XY}, but also to μ, and thus decompose it into two parts, μ_X which generates the marginal distribution P_X, and $\mu_{Y|X}$ which, conditioned on P_X, generates the posterior $P_{Y|X}$. (More precise notation might be μ_{P_X} instead of μ_X and $\mu_{P_{Y|X}|P_X}$ instead of $\mu_{Y|X}$, but this is rather cumbersome.) Denote $\widetilde{X} = (P_X, X)$. We then have, using Fubini's theorem,

$$\begin{aligned}
\mathcal{E}^{\infty}(f) &= \mathbb{E}_{P_X \sim \mu_X} \mathbb{E}_{P_{Y|X} \sim \mu_{Y|X}} \mathbb{E}_{X \sim P_X} \mathbb{E}_{Y \sim P_{Y|X}} \left[\ell(f(\widetilde{X}), Y) \right] \\
&= \mathbb{E}_{P_X \sim \mu_X} \mathbb{E}_{X \sim P_X} \mathbb{E}_{P_{Y|X} \sim \mu_{Y|X}} \mathbb{E}_{Y \sim P_{Y|X}} \left[\ell(f(\widetilde{X}), Y) \right] \\
&= \mathbb{E}_{(\widetilde{X}, Y) \sim Q^{\mu}} \left[\ell(f(\widetilde{X}), Y) \right].
\end{aligned}$$

Here Q^{μ} is the distribution that generates $\widetilde{X}$ by first drawing P_X according to μ_X, and then drawing X according to P_X; similarly, Y is generated, conditioned on $\widetilde{X}$, by first drawing $P_{Y|X}$ according to $\mu_{Y|X}$, and then drawing Y from $P_{Y|X}$. (The distribution of $\widetilde{X}$ again takes the form of a Campbell measure, see Remark 4.)

From the previous expression, we see that the risk $\mathcal{E}^{\infty}$ is like a standard supervised learning risk based on $(\widetilde{X}, Y) \sim Q^{\mu}$. Thus, we can deduce properties that are known to hold for supervised learning risks. For example, in the binary classification setting, if the loss is the 0/1 loss, then $f^*(\widetilde{X}) = 2\tilde{\eta}(\widetilde{X}) - 1$ is an optimal predictor, where $\tilde{\eta}(\widetilde{X}) = \mathbb{E}_{Y \sim Q^{\mu}_{Y|\widetilde{X}}} \left[\mathbf{1}_{\{Y=1\}} \right]$, and

$$\mathcal{E}^{\infty}(f) - \mathcal{E}^{\infty}(f^*) = \mathbb{E}_{\widetilde{X} \sim Q^{\mu}_{\widetilde{X}}} \left[\mathbf{1}_{\{\mathrm{sign}(f(\widetilde{X})) \neq \mathrm{sign}(f^*(\widetilde{X}))\}} |2\tilde{\eta}(\widetilde{X}) - 1| \right].$$

Furthermore, consistency in the sense of $\mathcal{E}^{\infty}$ with respect to a general loss ℓ (thought of as a surrogate) will imply consistency for the 0/1 loss, provided ℓ is classification calibrated (Bartlett et al., 2006).

For a given loss ℓ, the optimal or Bayes $\mathcal{E}^{\infty}$-risk in DG is in general larger than the expected Bayes risk under the (random) test sample generating distribution P^T_{XY}, because it is typically not possible to fully determine the Bayes-optimal predictor from only the marginal P^T_X. There is, however, a condition where for μ-almost all test distributions P^T_{XY}, the decision function $f^*(P^T_X, .)$ (where f^* is a global minimizer of Equation 6) coincides with

an optimal Bayes decision function for P_{XY}^T. This condition is simply that the posterior $P_{Y|X}$ is (μ-almost surely) a function of P_X (in other words, with the notation introduced above, $\mu_{Y|X}(P_X)$ is a Dirac measure for μ-almost all P_X). Although we will *not* be assuming this condition throughout the paper under **(2SGM)**, observe that it is implicitly assumed in the motivating application presented in Section 2, where an expert labels the data points by just looking at their marginal distribution.

Lemma 9 *For a fixed distribution P_{XY}, and a decision function $g : \mathcal{X} \to \mathbb{R}$, let us denote $\mathcal{R}(g, P_{XY}) = \mathbb{E}_{(X,Y)\sim P_{XY}}[\ell(g(X), Y)]$ and*

$$\mathcal{R}^*(P_{XY}) := \min_{g:\mathcal{X}\to\mathbb{R}} \mathcal{R}(g, P_{XY}) = \min_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{(X,Y)\sim P_{XY}}[\ell(g(X), Y)]$$

the corresponding optimal (Bayes) risk for the loss function ℓ under data distribution P_{XY}. Then under **(2SGM)***:*

$$\mathcal{E}^\infty(f^*) \geq \mathbb{E}_{P_{XY}\sim\mu}[\mathcal{R}^*(P_{XY})],$$

where $f^* : \mathfrak{P}_\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a minimizer of the idealized DG risk $\mathcal{E}^\infty$ defined in (6)*.*

Furthermore, if μ is a distribution on $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ such that μ-a.s. it holds $P_{Y|X} = F(P_X)$ for some deterministic mapping F, then for μ-almost all P_{XY}:

$$\mathcal{R}(f^*(P_X, .), P_{XY}) = \mathcal{R}^*(P_{XY})$$

and

$$\mathcal{E}^\infty(f^*) = \mathbb{E}_{P_{XY}\sim\mu}[\mathcal{R}^*(P_{XY})].$$

Proof For any $f : \mathfrak{P}_\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, one has for all P_{XY}: $\mathcal{R}(f(P_X, .), P_{XY}) \geq \mathcal{R}^*(P_{XY})$. Taking expectation with respect to P_{XY} establishes the first claim. Now for any fixed $P_X \in \mathfrak{P}_\mathcal{X}$, consider $P_{XY} := P_X \bullet F(P_X)$ and $g^*(P_X)$ a Bayes decision function for this joint distribution. Pose $f(P_X, x) := g^*(P_X)(x)$. Then f coincides for μ-almost all P_{XY} with a Bayes decision function for P_{XY}, achieving equality in the above inequality. The second equality follows by taking expectation over $P_{XY} \sim \mu$. ■

Under **(2SGM)**, we will establish that our proposed learning algorithm is $\mathcal{E}^\infty$-consistent, provided the average sample size grows to infinity as well as the total number of samples. Thus, the above result provides a condition on μ under which it is possible to asymptotically attain the (classical, single-task) Bayes risk on any test distribution although *no labels from this test distribution are observed*.

More generally, and speaking informally, if μ is such that $P_{Y|X}$ is close to being a function of P_X in some sense, we can expect the Bayes $\mathcal{E}^\infty$-risk for domain generalization to be close to the expected (classical single-task) Bayes risk for a random test distribution. We reiterate, however, that we make no assumptions on μ in this work so that the two quantities may be far apart. In the worst case, the posterior may be independent of the marginal, in which case a method for domain generalization will do no better than the naïve pooling strategy. For further discussion, see the comparison of domain adaptation and domain generalization in the next section.

4. Related Work

Since at least the 1990s, machine learning researchers have investigated the possibility of solving one learning problem by leveraging data from one or more related problems. In this section, we provide an overview of such problems and their relation to domain generalization, while also reviewing prior work on DG.

Two critical terms are *domain* and *task*. Use of these terms is not consistent throughout the literature, but at a minimum, the domain of a learning problem describes the input (feature) space $\mathcal{X}$ and marginal distribution of X, while the task describes the output space $\mathcal{Y}$ and the conditional distribution of Y given X (also called posterior). In many settings, however, the sets $\mathcal{X}$ and $\mathcal{Y}$ are the same for all learning problems, and the terms "domain" and "task" are used interchangeably to refer to a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. This is the perspective adopted in this work, as well as in much of the work on multi-task learning, domain adaptation (DA), and domain generalization.

Multi-task learning is similar to DG, except only the training tasks are of interest, and the goal is to leverage the similarity among distributions to improve the learning of individual predictors for each task (Caruana, 1997; Evgeniou et al., 2005; Yang et al., 2009). In contrast, in DG, we are concerned with generalization to a new task.

4.1 Domain Generalization vs. Domain Adaptation

Domain adaptation refers to the setting in which there is a specific target task and one or more source tasks. The goal is to design a predictor for the target task, for which there are typically few to no labeled training examples, by leveraging labeled training data from the source task(s).

Formulations of domain adaptation may take several forms, depending on the number of sources and whether there are any labeled examples from the target to supplement the unlabeled examples. In multi-source, unsupervised domain adaptation, the learner is presented with labeled training data from several source distributions, and unlabeled data from a target marginal distribution (see Zhang et al. (2015) and references therein). Thus, the available data are the same as in domain generalization, and algorithms for one of these problems may be applied to the other.

In all forms of DA, the goal is to attain optimal performance with respect to the joint distribution of the target domain. For example, if the performance measure is a risk, the goal is to attain the Bayes risk for the target domain. To achieve this goal, it is necessary to make assumptions about how the source and target distributions are related (Quionero-Candela et al., 2009). For example, several works adopt the covariate shift assumption, which requires the source and target domains to have the same posterior, allowing the marginals to differ arbitrarily (Zadrozny, 2004; Huang et al., 2007; Cortes et al., 2008; Sugiyama et al., 2008; Bickel et al., 2009; Kanamori et al., 2009; Yu and Szepesvari, 2012; Ben-David and Urner, 2012). Another common assumption is target shift, which stipulates that the source and target have the same class-conditional distributions, allowing the prior class probability to change (Hall, 1981; Titterington, 1983; Latinne et al., 2001; Storkey, 2009; Du Plessis and Sugiyama, 2012; Sanderson and Scott, 2014; Azizzadenesheli et al., 2019). Mansour et al. (2009b); Zhang et al. (2015) assume that the target posterior is a weighted combination of source posteriors, while Zhang et al. (2013); Gong et al. (2016) extend target

shift by also allowing the class-conditional distributions to undergo a location-scale shift, and Tasche (2017) assumes the ratio of class-conditional distributions is unchanged. Work on classification with label noise assumes the source data are obtained from the target distribution but the labels have been corrupted in either a label-dependent (Blanchard et al., 2016; Natarajan et al., 2018; van Rooyen and Williamson, 2018) or feature-dependent (Menon et al., 2018; Cannings et al., 2018; Scott, 2019) way. Finally, there are several works that assume the existence of a predictor that achieves good performance on both source and target domains (Ben-David et al., 2007, 2010; Blitzer et al., 2008; Mansour et al., 2009a; Cortes et al., 2015; Germain et al., 2016).

The key difference between DG and DA may be found in the performance measures optimized. In DG, the goal is to design a single predictor $f(P_X, x)$ that can apply to any future task, and risk is assessed with respect to the draw of both a new task, and (under **2SGM**) a new data point from that task. This is in contrast to DA, where the target distribution is typically considered fixed, and the goal is to design a predictor $f(x)$ where, in assessing the risk, the only randomness is in the draw of a new sample from the target task. This difference in performance measures for DG and DA has an interesting consequence for analysis. As we will show, it is possible to attain optimal risk (asymptotically) in DG without making any distributional assumptions like those described above for DA. Of course, this optimal risk is typically larger than the Bayes risk for any particular target domain (see Lemma 9). An interesting question for future research is whether it is possible to close or eliminate this gap (between DG and expected DA risks) by imposing distributional assumptions like those for DA.

Another difference between DA and DG lies in whether the learning algorithm must be rerun for each new test data set. Most unsupervised DA methods employ the unlabeled target data for training and thus, when a new unlabeled target data set is presented, the learning algorithm must be rerun. In contrast, most existing DG methods do not assume access to the unlabeled test data at learning time, and are capable of making predictions as new unlabeled data sets arrive without any further training.

4.2 Domain Generalization vs. Learning to Learn

In the problem of learning to learn (LTL, Thrun, 1996), which has also been called bias learning, meta-learning, and (typically in an online setting) lifelong learning, there are labeled data sets for several tasks, as in DG. There is also a given family of learning algorithms, and the objective is to design a meta-learner that selects the learning algorithm that will perform best on future tasks. The learning theoretic study of LTL traces to the work of Baxter (2000), who was the first to propose a distribution on tasks, which he calls an "environment," and which coincides with our μ. Given this setting, the performance of the learning algorithm selected by a meta-learner is obtained by drawing a new task at random, drawing a labeled training data set from that task, running the selected algorithm, drawing a test point, and evaluating the expected loss, where the expectation is with respect to all sources of randomness (new task, training data from new task, test point from new task).

Baxter analyzes learning algorithms given by usual empirical risk minimization over a hypothesis (prediction function) class, and the goal of the meta-learner is then to select a hypothesis class from a family of such classes. He shows that it is possible to find a good

trade-off between the complexity of a hypothesis class and its approximation capabilities for tasks sampled from μ, in an average sense. In particular, the information gained by finding a well-adapted hypothesis class can lead to significantly improved sample efficiency when learning a new task. See Maurer (2009) for further discussion of the results of Baxter (2000).

Later work on LTL establishes similar results that quantify the ability of a meta-learner to transfer knowledge to a new task. These meta-learners all optimize a particular structure that defines a learning algorithm, such as a feature representation (Maurer, 2009; Maurer et al., 2016; Denevi et al., 2018b), a prior on predictors in a PAC-Bayesian setting (Pentina and Lampert, 2014), a dictionary (Maurer et al., 2013), the bias of a regularizer (Denevi et al., 2018a), and a pretrained neural network (Finn et al., 2017). It is also worth noting that some algorithms on multi-task learning extract structures that characterize an environment and can be applied to LTL.

Although DG and LTL both involve generalization to a new task, they are clearly different problems because LTL assumes access to labeled data from the new task, whereas DG only sees unlabeled data and requires no additional learning. In LTL, the learner can achieve the Bayes risk for the new task, the only issue is the sample complexity. DG is thus a more challenging problem, but also potentially more useful since in many transfer learning settings, labeled data for the new task are unavailable.

4.3 Prior Work on Domain Generalization

To our knowledge, the first paper to consider domain generalization (as formulated in Section 3.2) was our earlier conference paper (Blanchard et al., 2011). The term "domain generalization" was coined by Muandet et al. (2013), who study the same setting and build upon our work by extracting features that facilitate DG. Yang et al. (2013) study an active learning variant of DG in the realizable setting, and directly learn the task sampling distribution.

Other methods for DG were studied by Khosla et al. (2012); Xu et al. (2014); Grubinger et al. (2015); Ghifary et al. (2015); Gan et al. (2016); Ghifary et al. (2017); Motiian et al. (2017); Li et al. (2017, 2018a,b,c,d); Balaji et al. (2018); Ding and Fu (2018); Shankar et al. (2018); Hu et al. (2019); Dou et al. (2019); Carlucci et al. (2019); Wang et al. (2019); Akuzawa et al. (2019). Many of these methods learn a common feature space for all tasks. Such methods are complementary to the method that we study. Indeed, our kernel-based learning algorithm may be applied after having learned a feature representation by another method, as was done by Muandet et al. (2013). Since our interest is primarily theoretical, we restrict our experimental comparison to another algorithm that also operates directly on the original input space, namely, a simple pooling algorithm that lumps all training tasks into a single data set and trains a single support vector machine.

5. Learning Algorithm

In this section, we introduce a concrete algorithm to tackle the learning problem exposed in Section 3, using an approach based on kernels. The function $k : \Omega \times \Omega \to \mathbb{R}$ is called a *kernel* on Ω if the matrix $(k(x_i, x_j))_{1 \le i,j \le n}$ is symmetric and positive semi-definite for all positive integers n and all $x_1, \ldots, x_n \in \Omega$. It is well known that every kernel k on Ω is associated to

a space of functions $f : \Omega \to \mathbb{R}$ called the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with kernel k. One way to envision $\mathcal{H}_k$ is as follows. Define $\Phi(x) := k(\cdot, x)$, which is called the *canonical feature map* associated with k. Then the span of $\{\Phi(x) : x \in \Omega\}$, endowed with the inner product $\langle \Phi(x), \Phi(x') \rangle = k(x, x')$, is dense in $\mathcal{H}_k$. We also recall the *reproducing property*, which states that $\langle f, \Phi(x) \rangle = f(x)$ for all $f \in \mathcal{H}_k$ and $x \in \Omega$.

For later use, we introduce the notion of a *universal* kernel. A kernel k on a compact metric space Ω is said to be *universal* when its RKHS is dense in $\mathcal{C}(\Omega)$, the set of continuous functions on Ω, with respect to the supremum norm. Universal kernels are important for establishing universal consistency of many learning algorithms. See Steinwart and Christmann (2008) for background on kernels and reproducing kernel Hilbert spaces.

Several well-known learning algorithms, such as support vector machines and kernel ridge regression, may be viewed as minimizers of a norm-regularized empirical risk over the RKHS of a kernel. A similar development has also been made for multi-task learning (Evgeniou et al., 2005). Inspired by this framework, we consider a general kernel-based algorithm as follows.

Consider the loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$. Let $\overline{k}$ be a kernel on $\mathfrak{P}_\mathcal{X} \times \mathcal{X}$, and let $\mathcal{H}_{\overline{k}}$ be the associated RKHS. For the sample S_i, recall that $\widehat{P}_X^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{X_{ij}}$ denotes the corresponding empirical X distribution. Also consider the extended input space $\mathfrak{P}_\mathcal{X} \times \mathcal{X}$ and the extended data $\widetilde{X}_{ij} = (\widehat{P}_X^{(i)}, X_{ij})$. Note that $\widehat{P}_X^{(i)}$ plays a role analogous to the task index in multi-task learning. Now define

$$\widehat{f}_\lambda = \operatorname*{arg\,min}_{f \in \mathcal{H}_{\overline{k}}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f(\widetilde{X}_{ij}), Y_{ij}) + \lambda \|f\|^2 . \tag{7}$$

Algorithms for solving (7) will be discussed in Section 7.

5.1 Specifying the Kernels

In the rest of the paper we will consider a kernel $\overline{k}$ on $\mathfrak{P}_\mathcal{X} \times \mathcal{X}$ of the product form

$$\overline{k}((P_1, x_1), (P_2, x_2)) = k_P(P_1, P_2) k_X(x_1, x_2), \tag{8}$$

where k_P is a kernel on $\mathfrak{P}_\mathcal{X}$ and k_X a kernel on $\mathcal{X}$.

Furthermore, we will consider kernels on $\mathfrak{P}_\mathcal{X}$ of a particular form. Let k'_X denote a kernel on $\mathcal{X}$ (which might be different from k_X) that is measurable and bounded. We define the *kernel mean embedding* $\Psi : \mathfrak{P}_\mathcal{X} \to \mathcal{H}_{k'_X}$:

$$P_X \mapsto \Psi(P_X) := \int_\mathcal{X} k'_X(x, \cdot) dP_X(x). \tag{9}$$

This mapping has been studied in the framework of "characteristic kernels" (Gretton et al., 2007a), and it has been proved that universality of k'_X implies injectivity of Ψ (Gretton et al., 2007b; Sriperumbudur et al., 2010).

Note that the mapping Ψ is linear. Therefore, if we consider the kernel $k_P(P_X, P'_X) = \langle \Psi(P_X), \Psi(P'_X) \rangle$, it is a linear kernel on $\mathfrak{P}_\mathcal{X}$ and cannot be a universal kernel. For this reason, we introduce yet another kernel $\mathfrak{K}$ on $\mathcal{H}_{k'_X}$ and consider the kernel on $\mathfrak{P}_\mathcal{X}$ given by

$$k_P(P_X, P'_X) = \mathfrak{K}\left(\Psi(P_X), \Psi(P'_X)\right) . \tag{10}$$

Note that particular kernels inspired by the finite dimensional case are of the form

$$\mathfrak{K}(v, v') = F(\|v - v'\|), \tag{11}$$

or

$$\mathfrak{K}(v, v') = G(\langle v, v' \rangle), \tag{12}$$

where F, G are real functions of a real variable such that they define a kernel. For example, $F(t) = \exp(-t^2/(2\sigma^2))$ yields a Gaussian-like kernel, while $G(t) = (1 + t)^d$ yields a polynomial-like kernel. Kernels of the above form on the space of probability distributions over a compact space $\mathcal{X}$ have been introduced and studied in Christmann and Steinwart (2010). Below we apply their results to deduce that $\overline{k}$ is a universal kernel for certain choices of k_X, k'_X, and $\mathfrak{K}$.

5.2 Relation to Other Kernel Methods

By choosing $\overline{k}$ differently, one can recover other existing kernel methods. In particular, consider the class of kernels of the same product form as above, but where

$$k_P(P_X, P'_X) = \begin{cases} 1 & P_X = P'_X \\ \tau & P_X \neq P'_X \end{cases}$$

If $\tau = 0$, the algorithm (7) corresponds to training N kernel machines $f(\widehat{P}_X^{(i)}, \cdot)$ using kernel k_X (e.g., support vector machines in the case of the hinge loss) on each training data set, independently of the others (note that this does not offer any generalization ability to a new data set). If $\tau = 1$, we have a "pooling" strategy that, in the case of equal sample sizes n_i, is equivalent to pooling all training data sets together in a single data set, and running a conventional supervised learning algorithm with kernel k_X (*i.e.*, this corresponds to trying to find a single "one-fits-all" prediction function which does not depend on the marginal). In the intermediate case $0 < \tau < 1$, the resulting kernel is a "multi-task kernel," and the algorithm recovers a multitask learning algorithm like that of Evgeniou et al. (2005). We compare to the pooling strategy below in our experiments. We also examined the multi-task kernel with $\tau < 1$, but found that, as far as generalization to a new unlabeled task is concerned, it was always outperformed by pooling, and so those results are not reported. This fits the observation that the choice $\tau = 0$ does not provide any generalization to a new task, while $\tau = 1$ at least offers some form of generalization, if only by fitting the same predictor to all data sets.

In the special case where all labels Y_{ij} are the same value for a given task, and k_X is taken to be the constant kernel, the problem we consider reduces to "distributional" classification or regression, which is essentially standard supervised learning where a distribution (observed through a sample) plays the role of the feature vector. Many of our analysis techniques specialize to this setting.

6. Learning Theoretic Study

This section presents generalization error and consistency analysis for the proposed kernel method under the agnostic and 2-stage generative models. Although the regularized estimation formula (7) defining $\widehat{f}_\lambda$ is standard, the generalization error analysis is not, owing to the particular sampling structures and risks under **(AGM)** and **(2SGM)**.

6.1 Universal Consistency under the Agnostic Generative Model

We will consider the following assumptions on the loss function and kernels:

(LB) The loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is L_ℓ-Lipschitz in its first variable and satisfies $B_0 := \sup_{y \in \mathcal{Y}} \ell(0, y) < \infty$.

(K-Bounded) The kernels k_X, k'_X and $\mathfrak{K}$ are bounded respectively by constants $B_k^2, B_{k'}^2 \geq 1$, and $B_{\mathfrak{K}}^2$.

The condition $B_0 < \infty$ always holds for classification, as well as certain regression settings. The boundedness assumptions are clearly satisfied for Gaussian kernels, and can be enforced by normalizing the kernel (discussed further below).

We begin with a generalization error bound that establishes uniform estimation error control over functions belonging to a ball of $\mathcal{H}_{\overline{k}}$. We then discuss universal kernels, and finally deduce universal consistency of the algorithm.

Let $\mathcal{B}_k(r)$ denote the closed ball of radius r, centered at the origin, in the RKHS of the kernel k. We start with the following simple result allowing us to bound the loss on a RKHS ball.

Lemma 10 *Suppose k is a kernel on a set Ω, bounded by B^2. Let $\ell : \mathbb{R} \times \mathcal{Y} \to [0, \infty)$ be a loss satisfying* **(LB)**. *Then for any $R > 0$ and $f \in B_k(R)$, and any $z \in \Omega$ and $y \in \mathcal{Y}$,*

$$\left|\ell(f(z), y)\right| \leq B_0 + L_\ell R B \tag{13}$$

Proof By the Lipschitz continuity of ℓ, the reproducing property, and Cauchy-Schwarz, we have

$$\begin{aligned}
\left|\ell(f(z), y)\right| &\leq \ell(0, y) + \left|\ell(f(z), y) - \ell(0, y)\right| \\
&\leq B_0 + L_\ell |f(z) - 0| \\
&= B_0 + L_\ell \left|\langle f, k(z, \cdot)\rangle\right| \\
&\leq B_0 + L_\ell \|f\|_{\mathcal{H}_k} B \\
&\leq B_0 + L_\ell R B.
\end{aligned}$$

■

The expression in (13) serves to replace the boundedness assumption (3) in Theorem 5. We now state the following, which is a specialization of Theorem 5 to the kernel setting.

Theorem 11 (Uniform estimation error control over RKHS balls) *Assume* **(LB)** *and* **(K-Bounded)** *hold, and data generation follows* **(AGM)**. *Then for any $R > 0$, with probability at least $1 - \delta$ (with respect to the draws of the samples $S_i, i = 1, \ldots, N$)*

$$\sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left|\widehat{\mathcal{E}}(f, N) - \mathcal{E}(f)\right| \leq (B_0 + L_\ell R B_{\mathfrak{K}} B_k) \frac{(\sqrt{\log \delta^{-1}} + 2)}{\sqrt{N}}. \tag{14}$$

Proof This is a direct consequence of Theorem 5 and of Lemma 10, the kernel $\overline{k}$ on $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$ being bounded by $B_k^2 B_{\mathfrak{K}}^2$. As noted there, the main term in the upper bound (4) is a standard Rademacher complexity on the augmented input space $\mathfrak{P} \times \mathcal{X}$, endowed with the Campbell measure $\mathcal{C}(P_S)$.

In the kernel learning context, we can bound the Rademacher complexity term using a standard bound for the Rademacher complexity of a Lipschitz loss function on the ball of radius R of $\mathcal{H}_{\overline{k}}$ (Koltchinskii, 2001; Bartlett and Mendelson, 2002, e.g., Theorems 8, 12 and Lemma 22 there), using again the bound $B_k^2 B_{\mathfrak{K}}^2$ on the kernel $\overline{k}$, giving the conclusion. ■

Next, we turn our attention to universal kernels (see Section 5 for the definition). A relevant notion for our purposes is that of a normalized kernel. If k is a kernel on Ω, then

$$k^*(x, x') := \frac{k(x, x')}{\sqrt{k(x, x)k(x', x')}}$$

is the associated *normalized* kernel. If a kernel is universal, then so is its associated normalized kernel. For example, the exponential kernel $k(x, x') = \exp(\kappa \langle x, x' \rangle_{\mathbb{R}^d})$, $\kappa > 0$, can be shown to be universal on $\mathbb{R}^d$ through a Taylor series argument. Consequently, the Gaussian kernel

$$k_\sigma(x, x') := \frac{\exp(\frac{1}{\sigma^2}\langle x, x' \rangle)}{\exp(\frac{1}{2\sigma^2}\|x\|^2)\exp(\frac{1}{2\sigma^2}\|x'\|^2)}$$

is universal, being the normalized kernel associated with the exponential kernel with $\kappa = 1/\sigma^2$. See Steinwart and Christmann (2008) for additional details and discussion.

To establish that $\overline{k}$ is universal on $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$, the following lemma is useful.

Lemma 12 *Let Ω, Ω' be two compact spaces and k, k' be kernels on Ω, Ω', respectively. If k, k' are both universal, then the product kernel*

$$\overline{k}((x, x'), (y, y')) := k(x, y)k'(x', y')$$

is universal on $\Omega \times \Omega'$.

Several examples of universal kernels are known on Euclidean space. For our purposes, we also need universal kernels on $\mathfrak{P}_{\mathcal{X}}$. Fortunately, this was studied by Christmann and Steinwart (2010). Some additional assumptions on the kernels and feature space are required:

(K-Univ) k_X, k'_X, $\mathfrak{K}$, and $\mathcal{X}$ satisfy the following:

- $\mathcal{X}$ is a compact metric space
- k_X is universal on $\mathcal{X}$
- k'_X is continuous and universal on $\mathcal{X}$
- $\mathfrak{K}$ is universal on any compact subset of $\mathcal{H}_{k'_X}$.

Adapting the results of Christmann and Steinwart (2010), we have the following.

Theorem 13 (Universal kernel) *Assume condition* **(K-Univ)** *holds. Then, for k_P defined as in* (10)*, the product kernel $\overline{k}$ in* (8) *is universal on $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$.*

Furthermore, the assumption on $\mathfrak{K}$ is fulfilled if $\mathfrak{K}$ is of the form (12)*, where G is an analytical function with positive Taylor series coefficients, or if $\mathfrak{K}$ is the normalized kernel associated to such a kernel.*

Proof By Lemma 12, it suffices to show $\mathfrak{P}_{\mathcal{X}}$ is a compact metric space, and that $k_P(P_X, P'_X)$ is universal on $\mathfrak{P}_{\mathcal{X}}$. The former statement follows from Theorem 6.4 of Parthasarathy (1967), where the metric is the Prohorov metric. We will deduce the latter statement from Theorem 2.2 of Christmann and Steinwart (2010). The statement of Theorem 2.2 there is apparently restricted to kernels of the form (12), but the proof actually only uses that the kernel $\mathfrak{K}$ is universal on any compact set of $\mathcal{H}_{k'_X}$. To apply Theorem 2.2, it remains to show that $\mathcal{H}_{k'_X}$ is a separable Hilbert space, and that Ψ is injective and continuous. Injectivity of Ψ is equivalent to k'_X being a characteristic kernel, and follows from the assumed universality of k'_X (Sriperumbudur et al., 2010). The continuity of k'_X implies separability of $\mathcal{H}_{k'_X}$ (Steinwart and Christmann (2008), Lemma 4.33) as well as continuity of Ψ (Christmann and Steinwart (2010), Lemma 2.3 and preceding discussion). Now Theorem 2.2 of Christmann and Steinwart (2010) may be applied, and the results follows.

The fact that kernels of the form (12), where G is analytic with positive Taylor coefficients, are universal on any compact set of $\mathcal{H}_{k'_X}$ was established in the proof of Theorem 2.2 of the same work (Christmann and Steinwart, 2010). ■

As an example, suppose that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Let k_X and k'_X be Gaussian kernels on $\mathcal{X}$. Taking $G(t) = \exp(t)$, it follows that $\mathfrak{K}(P_X, P'_X) = \exp(\langle \Psi(P_X), \Psi(P'_X) \rangle_{\mathcal{H}_{k'_X}})$ is universal on $\mathfrak{P}_{\mathcal{X}}$. By similar reasoning as in the finite dimensional case, the Gaussian-like kernel $\mathfrak{K}(P_X, P'_X) = \exp(-\frac{1}{2\sigma^2}\|\Psi(P_X) - \Psi(P'_X)\|^2_{\mathcal{H}_{k'_X}})$ is also universal on $\mathfrak{P}_{\mathcal{X}}$. Thus the product kernel is universal on $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$.

From Theorems 11 and 13, we may deduce universal consistency of the learning algorithm.

Corollary 14 (Universal consistency) *Assume that conditions* **(LB)**, **(K-Bounded)** *and* **(K-Univ)** *are satisfied. Let $\lambda = \lambda(N)$ be a sequence such that as $N \to \infty$: $\lambda(N) \to 0$ and $\lambda(N)N/\log N \to \infty$. Then*

$$\mathcal{E}(\widehat{f}_{\lambda(N)}) \to \inf_{f:\mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}} \mathcal{E}(f) \text{ a.s., as } N \to \infty.$$

The proof of the corollary relies on the bound established in Theorem 11, the universality of $\overline{k}$ established in Theorem 13, and otherwise relatively standard arguments.

One notable feature of this result is that we have established consistency where only N is required to diverge. In particular, the training sample sizes n_i may remain bounded. In the next subsection, we consider the role of the n_i under the 2-stage generative model.

6.2 Role of the Individual Sample Sizes under the 2-Stage Generative Model

In this section, we are concerned with the role of the individual sample sizes $(n_i)_{1 \le i \le N}$, more precisely, of their distribution ν under **(2SGM)**, see Section 3.1. A particular motivation

for investigating this point is that in some applications the number of training points per task is large, which can give rise to a high computational burden at the learning stage (and also for storing the learned model in computer memory). A practical way to alleviate this issue is to reduce the number of training points per task by random subsampling, which in effect modifies the sample size distribution ν while keeping the generating distribution μ for the tasks' point distributions unchanged. Observe that under **(AGM)** the sample size and the sample point distribution may be dependent in general, and subsampling would then affect that relationship in an unknown manner. This is why we assume **(2SGM)** in the present section.

We will consider the following additional assumption.

(K-Hölder) The canonical feature map $\Phi_{\mathfrak{K}} : \mathcal{H}_{k'_X} \to \mathcal{H}_{\mathfrak{K}}$ associated to $\mathfrak{K}$ satisfies a Hölder condition of order $\alpha \in (0, 1]$ with constant $L_{\mathfrak{K}}$, on $\mathcal{B}_{k'_X}(B_{k'})$:

$$\forall v, w \in \mathcal{B}_{k'_X}(B_{k'}) : \qquad \|\Phi_{\mathfrak{K}}(v) - \Phi_{\mathfrak{K}}(w)\| \leq L_{\mathfrak{K}} \|v - w\|^{\alpha} . \tag{15}$$

Sufficient conditions for (15) are described in Section A.4. As an example, the condition is shown to hold with $\alpha = 1$ when $\mathfrak{K}$ is the Gaussian-like kernel on $\mathcal{H}_{k'_X}$.

Since we are interested in the influence of the number of training points per task, it is helpful to introduce notations for the **(2SGM)** risks that are conditioned on a fixed task P_{XY}. Thus, we introduce the following notation, in analogy to (5)–(6) introduced in Section 3.4, for risk at sample size n, and risk at infinite sample size, conditional to P_{XY}:

$$\mathcal{E}(f|P_{XY}, n) := \mathbb{E}_{S^T \sim (P_{XY})^{\otimes n}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(f(\widehat{P}_X, X_i), Y_i) \right]; \tag{16}$$

$$\mathcal{E}^{\infty}(f|P_{XY}) := \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\ell(f(P_X, X), Y) \right]. \tag{17}$$

The following proposition gives an upper bound on the discrepancy between these risks. It can be seen as a quantitative version of Proposition 7 in the kernel setting, which is furthermore uniform over an RKHS ball.

Theorem 15 *Assume conditions* **(LB)**, **(K-Bounded)**, *and* **(K-Hölder)** *hold. If the sample $S = (X_j, Y_j)_{1 \leq j \leq n}$ is made of n i.i.d. realizations from P_{XY}, with P_{XY} and n fixed, then for any $R > 0$, with probability at least $1 - \delta$:*

$$\sup_{f \in \mathcal{B}_{\bar{k}}(R)} |\mathcal{L}(S, f) - \mathcal{E}^{\infty}(f|P_{XY})| \;\leq\; (B_0 + 3L_{\ell} R B_k (B_{k'}^{\alpha} L_{\mathfrak{K}} + B_{\mathfrak{K}})) \left(\frac{\log(3\delta^{-1})}{n} \right)^{-\frac{\alpha}{2}} . \tag{18}$$

Averaging over the draw of S, again with P_{XY} and n fixed, it holds for any $R > 0$:

$$\sup_{f \in \mathcal{B}_{\bar{k}}(R)} |\mathcal{E}(f|P_{XY}, n) - \mathcal{E}^{\infty}(f|P_{XY})| \leq 2L_{\ell} R B_k L_{\mathfrak{K}} B_{k'}^{\alpha} n^{-\alpha/2}. \tag{19}$$

As a consequence, for the unconditional risks when (P_{XY}, n) is drawn from $\mu \otimes \nu$ under **(2SGM)**, *for any $R > 0$:*

$$\sup_{f \in \mathcal{B}_{\bar{k}}(R)} |\mathcal{E}(f) - \mathcal{E}^{\infty}(f)| \leq 2RL_{\ell} B_k L_{\mathfrak{K}} B_{k'}^{\alpha} \mathbb{E}_{\nu} \left[n^{-\frac{\alpha}{2}} \right]. \tag{20}$$

The above results are useful in a number of ways. First, under **(2SGM)**, we can consider the goal of asymptotically achieving the idealized optimal risk $\inf_f \mathcal{E}^\infty(f)$, where we recall that $\mathcal{E}^\infty(f)$ is the expected loss of a decision function f over a random test task P_{XY}^T in the case where P_X^T would be perfectly observed (this can be thought of as observing an infinite sample from the marginal). Equation (20) bounds the risk under **(2SGM)** in terms of the risk under **(AGM)**, for which we have already established consistency. Thus, consistency to the idealized risk under **(2SGM)** will be possible if the number of examples n_i per training task also grows together with the number of training tasks N. The following result formalizes this intuition.

Corollary 16 *Assume* **(LB)**, **(K-Bounded)**, *and* **(K-Hölder)**, *and assume* **(2SGM)**. *Then for any $R > 0$, with probability at least $1 - \delta$ with respect to the draws of the training tasks and training samples*

$$\sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left| \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(S_i, f) - \mathcal{E}^\infty(f) \right| \leq (B_0 + L_\ell R B_{\mathfrak{K}} B_k) \frac{(\sqrt{\log \delta^{-1}} + 2)}{\sqrt{N}} + 2RL_\ell B_k L_{\mathfrak{K}} B_{k'}^\alpha \mathbb{E}_\nu \left[n^{-\frac{\alpha}{2}} \right]. \quad (21)$$

Consider an asymptotic setting under **(2SGM)** *in which, as the number of training tasks $N \to \infty$, the distribution μ remains fixed but the sample size distribution ν_N depends on N. Denote $\kappa_N^{-1} := \mathbb{E}_{\nu_N}\left[n^{-\alpha/2}\right]$. Assuming* **(K-Univ)** *is satisfied, and the regularization parameter $\lambda(N)$ is such that $\lambda(N) \to 0$ and $\lambda(N) \min(N, \kappa_N^2) \to \infty$, then*

$$\mathcal{E}(\widehat{f}_{\lambda(N)}) \to \inf_{f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}} \mathcal{E}^\infty(f) \text{ in probability, as } N \to \infty.$$

Proof The setting is that of the **(2SGM)** model. This is a particular case of **(AGM)**, so we can apply Theorem 11 and combine with (20) to get the announced bound. The consistency statement follows the same argument as in the proof of Corollary 14, with $\mathcal{E}(f)$ replaced by $\mathcal{E}^\infty(f)$, and $\varepsilon(N)$ there replaced by the RHS in (21). ■

Remark 17 *The bound* (21) *is non-asymptotic and can be used as such to assess the respective role of number of tasks N and of the sample size distribution ν when the objective is the idealized risk (see below). The result of consistency to that risk on the other hand, is formalized as a "triangular array" type of asymptotics where the distribution of the sizes n_i of the i.i.d. training samples S_i changes with their number N.*

Remark 18 *Our conference paper (Blanchard et al., 2011) also established a generalization error bound and consistency for $\mathcal{E}^\infty$ under* **(2SGM)**. *That bound had a different form for two main reasons. First, it assumed the loss to be bounded, whereas the present analysis avoids that assumption via Lemma 10. Second, that analysis did not leverage a connection to* **(AGM)**, *which led to a $\log N$ in the second term. This required the two sample sizes to be coupled asymptotically to achieve consistency. In the present analysis, the two sample sizes N and n may diverge at arbitrary rates.*

Remark 19 *It is possible to obtain a result similar to Corollary 16 when the training task sample sizes $(n_i)_{1\leq i\leq N}$ are fixed (considered as deterministic), unequal and possibly arbitrary. In this case we would follow a slightly different argument, leveraging* (18) *for each single training task together with a union bound, and applying Theorem 11 to the idealized situation with an infinite number of samples per training task. This way, the term $\mathbb{E}_\nu\left[n^{-\frac{\alpha}{2}}\right]$ is replaced by $\log(N)N^{-1}\sum_{i=1}^N n_i^{-\frac{\alpha}{2}}$, the additional logarithmic factor being the price of the union bound. We eschew an exact statement for brevity.*

We now come back to our initial motivation of possibly reducing computational burden by subsampling and analyze to what extent this affects statistical error. Under **(2SGM)** the effect of subsampling (without replacement) is transparent: it amounts to changing the original individual sample size distribution ν by $\nu' = \delta_{n'}$, while keeping the generating distribution μ for the tasks' point distributions fixed. Here n' is the common fixed size of the training subsamples, and we must assume implicitly that the original sample sizes are a.s. larger than n', i.e. that their distribution ν has support $[n', \infty)$. For simplicity, for the rest of the discussion we only consider the case of equal, deterministic sizes of sample (n) and subsample $(n' < n)$. Using (21) we compare the two settings to a common reference, namely the idealized risk $\mathcal{E}^\infty$. We see that the statistical risk bound in (21) is unchanged up to a small factor if $n' \geq \min(N^\alpha, n)$. Assuming $\alpha = 1$ to simplify, in the case where the original sample sizes n are much larger than the number of training tasks N, this suggests that we can subsample to $n' \approx N$ without taking a significant hit to performance. This applies equally well to subsampling the tasks used for prediction or testing. The most precise statement in this regard is (18), since it bounds the deviations of the observed prediction loss for a fixed task P_{XY} and i.i.d. sample from that task.

The minimal subsampling size n' can be interpreted as an optimal efficiency/accuracy tradeoff, since it reduces computational complexity as much as possible without sacrificing statistical accuracy. Similar considerations appear in the context of distribution regression (Szabó et al., 2016, Remark 6). In that reference, a sharp analysis giving rise to *fast convergence rates* is presented, resulting in a more involved optimal balance between N and n. In the present work, we have focused on *slow rates* based on a uniform control of the estimation error over RKHS balls; we leave for future work sharper convergence bounds (under additional regularity conditions), which would also give rise to more refined balancing conditions between n and N.

7. Implementation

Implementation of the algorithm in (7) relies on techniques that are similar to those used for other kernel methods, but with some variations.[3] The first subsection illustrates how, for the case of hinge loss, the optimization problem corresponds to a certain cost-sensitive support vector machine. The second subsection focuses on more scalable implementations based on approximate feature mappings.

3. Code is available at https://github.com/aniketde/DomainGeneralizationMarginal

7.1 Representer Theorem and Hinge Loss

For a particular loss ℓ, existing algorithms for optimizing an empirical risk based on that loss can be adapted to the setting of marginal transfer learning. We now illustrate this idea for the case of the hinge loss, $\ell(t,y) = \max(0, 1-yt)$. To make the presentation more concise, we will employ the extended feature representation $\widetilde{X}_{ij} = (\widehat{P}_X^{(i)}, X_{ij})$, and we will also "vectorize" the indices (i,j) so as to employ a single index on these variables and on the labels. Thus the training data are $(\widetilde{X}_i, Y_i)_{1 \le i \le M}$, where $M = \sum_{i=1}^N n_i$, and we seek a solution to

$$\min_{f \in \mathcal{H}_{\overline{k}}} \sum_{i=1}^M c_i \max(0, 1 - Y_i f(\widetilde{X}_i)) + \frac{1}{2} \|f\|^2 .$$

Here $c_i = \frac{1}{\lambda N n_m}$, where m is the smallest positive integer such that $i \le n_1 + \cdots + n_m$. By the representer theorem (Steinwart and Christmann, 2008), the solution of (7) has the form

$$\widehat{f}_\lambda = \sum_{i=1}^M r_i \overline{k}(\widetilde{X}_i, \cdot)$$

for real numbers r_i. Plugging this expression into the objective function of (7), and introducing the auxiliary variables ξ_i, we have the quadratic program

$$\begin{aligned} \min_{r,\xi} \ & \frac{1}{2} r^T \overline{K} r + \sum_{i=1}^M c_i \xi_i \\ \text{s.t. } & Y_i \sum_{j=1}^M r_j \overline{k}(\widetilde{X}_i, \widetilde{X}_j) \ge 1 - \xi_i, \ \forall i \\ & \xi_i \ge 0, \ \forall i, \end{aligned}$$

where $\overline{K} := (\overline{k}(\widetilde{X}_i, \widetilde{X}_j))_{1 \le i,j \le M}$. Using Lagrange multiplier theory, the dual quadratic program is

$$\begin{aligned} \max_{\alpha} \ & -\frac{1}{2} \sum_{i,j=1}^M \alpha_i \alpha_j Y_i Y_j \overline{k}(\widetilde{X}_i, \widetilde{X}_j) + \sum_{i=1}^M \alpha_i \\ \text{s.t. } & 0 \le \alpha_i \le c_i \ \forall i, \end{aligned}$$

and the optimal function is

$$\widehat{f}_\lambda = \sum_{i=1}^M \alpha_i Y_i \overline{k}(\widetilde{X}_i, \cdot).$$

This is equivalent to the dual of a cost-sensitive support vector machine, without offset, where the costs are given by c_i. Therefore we can learn the weights α_i using any existing software package for SVMs that accepts example-dependent costs and a user-specified kernel matrix, and allows for no offset. Returning to the original notation, the final predictor given a test X-sample S^T has the form

$$\widehat{f}_\lambda(\widehat{P}_X^T, x) = \sum_{i=1}^N \sum_{j=1}^{n_i} \alpha_{ij} Y_{ij} \overline{k}((\widehat{P}_X^{(i)}, X_{ij}), (\widehat{P}_X^T, x))$$

where the α_{ij} are nonnegative. Like the SVM, the solution is often sparse, meaning most α_{ij} are zero.

Finally, we remark on the computation of $k_P(\widehat{P}_X, \widehat{P}'_X)$. When $\mathfrak{K}$ has the form of (11) or (12), the calculation of k_P may be reduced to computations of the form $\left\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \right\rangle$. If $\widehat{P}_X$ and $\widehat{P}'_X$ are empirical distributions based on the samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_{n'}$, then

$$\begin{aligned}\left\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \right\rangle &= \left\langle \frac{1}{n}\sum_{i=1}^{n} k'_X(X_i, \cdot), \frac{1}{n'}\sum_{j=1}^{n'} k'_X(X'_j, \cdot) \right\rangle \\ &= \frac{1}{nn'}\sum_{i=1}^{n}\sum_{j=1}^{n'} k'_X(X_i, X'_j).\end{aligned}$$

Note that when k'_X is a (normalized) Gaussian kernel, $\Psi(\widehat{P}_X)$ coincides (as a function) with a smoothing kernel density estimate for P_X.

7.2 Approximate Feature Mapping for Scalable Implementation

Assuming $n_i = n$, for all i, the computational complexity of a nonlinear SVM solver (in our context) is between $O(N^2n^2)$ and $O(N^3n^3)$ (Joachims, 1999; Chang and Lin, 2011). Thus, standard nonlinear SVM solvers may be insufficient when N or n are very large.

One approach to scaling up kernel methods is to employ approximate feature mappings together with linear solvers. This is based on the idea that kernel methods are solving for a linear predictor after first nonlinearly transforming the data. Since this nonlinear transformation can have an extremely high- or even infinite-dimensional output, classical kernel methods avoid computing it explicitly. However, if the feature mapping can be approximated by a finite dimensional transformation with a relatively low-dimensional output, one can directly solve for the linear predictor, which can be accomplished in $O(Nn)$ time (Hsieh et al., 2008).

In particular, given a kernel $\overline{k}$, the goal is to find an approximate feature mapping $\overline{z}(\tilde{x})$ such that $\overline{k}(\tilde{x}, \tilde{x}') \approx \overline{z}(\tilde{x})^T\overline{z}(\tilde{x}')$. Given such a mapping $\overline{z}$, one then applies an efficient linear solver, such as Liblinear (Fan et al., 2008), to the training data $(\overline{z}(\tilde{X}_{ij}), Y_{ij})_{ij}$ to obtain a weight vector w. The final prediction on a test point $\tilde{x}$ is then $\text{sign}\{w^T\overline{z}(\tilde{x})\}$. As described in the previous subsection, the linear solver may need to be tweaked, as in the case of unequal sample sizes n_i, but this is usually straightforward.

Recently, such low-dimensional approximate future mappings $z(x)$ have been developed for several kernels. We examine two such techniques in the context of marginal transfer learning, the Nyström approximation (Williams and Seeger, 2001; Drineas and Mahoney, 2005) and random Fourier features. The Nyström approximation applies to any kernel method, and therefore extends to the marginal transfer setting without additional work. On the other hand, we give a novel extension of random Fourier features to the marginal transfer learning setting (for the case of all Gaussian kernels), together with performance analysis. Our approach is similar to the one in Jitkrittum et al. (2015) which proposes a two-stage approximation for the mean embedding. Note that Jitkrittum et al. (2015) does not give an error bound. We describe our novel extension of random Fourier features to

the marginal transfer learning setting, with error bounds, in the appendix, where we also review the Nyström method.

The Nyström approximation holds for any positive definite kernel, but random Fourier features can be used only for shift invariant kernels. On the other hand, random Fourier features are very easy to implement and the Nyström method has additional time complexity due to an eigenvalue decomposition. Moreover, the Nyström method is useful only when the kernel matrix has low rank. For additional comparison of various kernel approximation approaches we refer the reader to Le et al. (2013). In our experiments, we use random Fourier features when all kernels are Gaussian and the Nyström method otherwise.

8. Experiments

This section empirically compares our marginal transfer learning method with pooling.[4] One implementation of the pooling algorithm was mentioned in Section 5.2, where k_P is taken to be a constant kernel. Another implementation is to put all the training data sets together and train a single conventional kernel method. The only difference between the two implementations is that in the former, weights of $1/n_i$ are used for examples from training task i. In almost all of our experiments below, the various training tasks have the same sample sizes, in which case the two implementations coincide. The only exception is the fourth experiment when we use all training data, in which case we use the second of the two implementations mentioned above.

We consider three classification problems ($\mathcal{Y} = \{-1, 1\}$), for which the hinge loss is employed, and one regression problem ($\mathcal{Y} \subset \mathbb{R}$), where the ϵ-insensitive loss is employed. Thus, the algorithms implemented are natural extensions of support vector classification and regression to domain generalization. Performance of a learning strategy is assessed by holding out several data sets $S_1^T, \ldots, S_{N_T}^T$, learning a decision function $\widehat{f}$ on the remaining data sets, and reporting the average empirical risk $\frac{1}{N_T}\sum_{i=1}^{N_T} \mathcal{L}(S_i^T, \widehat{f})$. In some cases, this value is again averaged over several randomized versions of the experiment.

8.1 Model Selection

The various experiments use different combinations of kernels. In all experiments, linear kernels $k(x_1, x_2) = x_1^T x_2$ and Gaussian kernels $k_\sigma(x_1, x_2) = \exp\big(-\frac{||x_1 - x_2||^2}{2\sigma^2}\big)$ were used.

The bandwidth σ of each Gaussian kernel and the regularization parameter λ of the machines were selected by grid search. For model selection, five-fold cross-validation was used. In order to stabilize the cross-validation procedure, it was repeated 5 times over independent random splits into folds (Kohavi, 1995). Thus, candidate parameter values were evaluated on the 5×5 validation sets and the configuration yielding the best average performance was selected. If any of the chosen hyper-parameters was at the grid boundary, the grid was extended accordingly, i.e., the same grid size has been used, however, the center of grid has been assigned to the previously selected point. The grid used for kernels was $\sigma \in \big(10^{-2}, 10^{4}\big)$ with logarithmic spacing, and the grid used for the regularization parameter was $\lambda \in \big(10^{-1}, 10^{1}\big)$ with logarithmic spacing.

4. Code available at https://github.com/aniketde/DomainGeneralizationMarginal

8.2 Synthetic Data Experiment

To illustrate the proposed method, a synthetic problem was constructed. The synthetic data generation algorithm is given in Algorithm 1. In brief, for each classification task, the data are uniformly supported on an ellipse, with the major axis determining the labels, and the rotation of the major axis randomly generated in a 90 degree range for each task. One random realization of this synthetic data is shown in Figure 2. This synthetic data set is an ideal candidate for marginal transfer learning, because the Bayes classifier for a task is uniquely determined by the marginal distribution of the features, i.e. Lemma 9 applies (and the optimal error $\inf_f \mathcal{E}^\infty(f)$ is zero). On the other hand, observe that the expectation of each X distribution is the same regardless of the task and thus does not provide any relevant information, so that taking into account at least second order information is needed to perform domain generalization.

To analyse the effects of number of examples per task (n) and number of tasks (N), we constructed 12 synthetic data sets by taking combinations $N \times n$ where $N \in \{16, 64, 256\}$ and $n \in \{8, 16, 32, 256\}$. For each synthetic data set, the test set contains 10 tasks and each task contains one million data points. All kernels are taken to be Gaussian, and the random Fourier features speedup is used. The results are shown in Figure 3 and Tables 1 and 2 (see appendix). The marginal transfer learning (MTL) method significantly outperforms the baseline pooling method. Furthermore, the performance of MTL improves as N and n increase, as expected. The pooling method, however, does no better than random guessing regardless of N and n.

In the remaining experiments, the marginal distribution does not perfectly characterize the optimal decision function, but still provides some information to offer improvements over pooling.

Algorithm 1: Synthetic Data Generation

input : N: Number of tasks, n: Number of training examples per task
output: Realization of synthetic data set for N tasks
for $i = 1$ *to* N **do**

- sample rotation α_i uniformly in $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$;
- Take an ellipse whose major axis is aligned with the horizontal axis, and rotate it by an angle of α_i about its center;
- Sample n points X_{ij}, $j = 1, \ldots, n$ uniformly at random from the rotated ellipse;
- Label the points according to their position with respect to the major axis, i.e. the points that are on the right of the major axis are considered as class 1 and the points on the left of the major axis are considered as class -1.

end

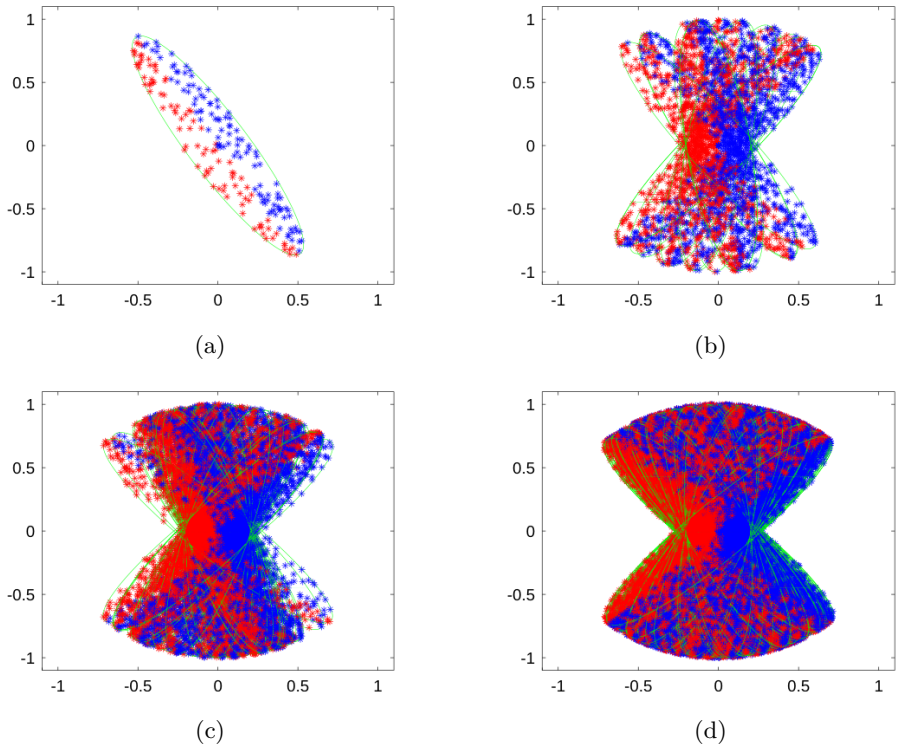

Figure 2: Plots of synthetic data sets (red and blue points represent negative and positive classes) for different settings: (a) Random realization of a single task with 256 training examples per task. Plots (b), (c) and(d) are random realizations of synthetic data with 256 training examples for 16, 64 and 256 tasks.

8.3 Parkinson's Disease Telemonitoring

We test our method in the regression setting using the Parkinson's disease telemonitoring data set, which is composed of a range of biomedical voice measurements using a telemonitoring device from 42 people with early-stage Parkinson's. The recordings were automatically captured in the patients' homes. The aim is to predict the clinician's Parkinson's disease symptom score for each recording on the unified Parkinson's disease rating scale (UPDRS) (Tsanas et al., 2010). Thus we are in a regression setting, and employ the ϵ-insensitive loss from support vector regression. All kernels are taken to be Gaussian, and the random Fourier features speedup is used.

There are around 200 recordings per patient. We randomly select 7 test users and then vary the number of training users N from 10 to 35 in steps of 5, and we also vary the number of training examples n per user from 20 to 100. We repeat this process several times to get the average errors which are shown in Fig 4 and Tables 3 and 4 (see appendix). The marginal transfer learning method clearly outperforms pooling, especially as N and n increase.

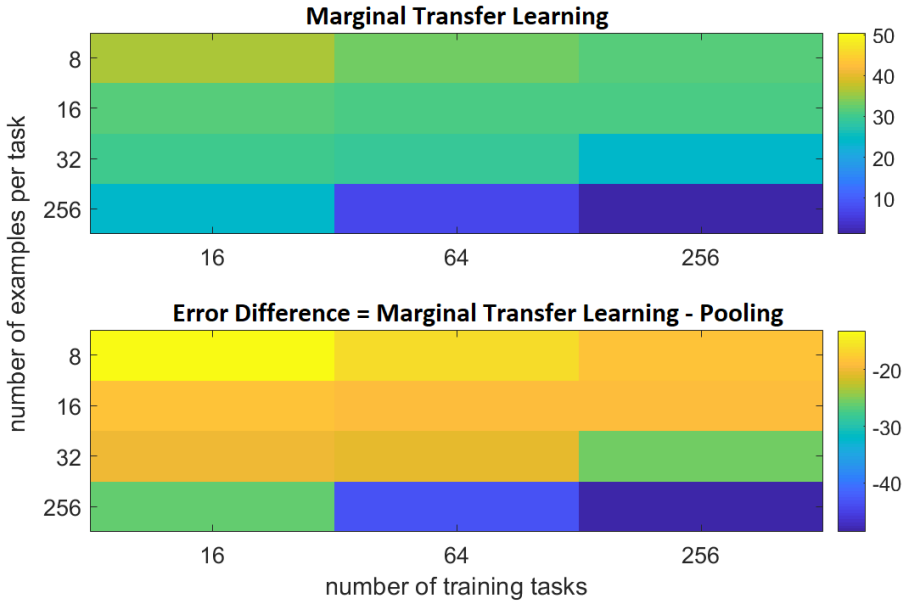

Figure 3: Synthetic data set: Classification error rates for proposed method and difference with baseline for different experimental settings, i.e., number of examples per task and number of tasks.

8.4 Satellite Classification

Microsatellites are increasingly deployed in space missions for a variety of scientific and technological purposes. Because of randomness in the launch process, the orbit of a microsatellite is random, and must be determined after the launch. One recently proposed approach is to estimate the orbit of a satellite based on radiofrequency (RF) signals as measured in a ground sensor network. However, microsatellites are often launched in bunches, and for this approach to be successful, it is necessary to associate each RF measurement vector with a particular satellite. Furthermore, the ground antennae are not able to decode unique identifier signals transmitted by the microsatellites, because (a) of constraints on the satellite/ground antennae links, including transmission power, atmospheric attenuation, scattering, and thermal noise, and (b) ground antennae must have low gain and low directional specificity owing to uncertainty in satellite position and dynamics. To address this problem, recent work has proposed to apply our marginal transfer learning methodology (Sharma and Cutler, 2015).

As a concrete instance of this problem, suppose two microsatellites are launched together. Each launch is a random phenomenon and may be viewed as a task in our framework. For each launch i, training data (X_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, are generated using a highly realistic simulation model, where X_{ij} is a feature vector of RF measurements across a particular sensor network and at a particular time, and Y_{ij} is a binary label identifying which of the two microsatellites produced a given measurement. By applying our methodology, we can classify unlabeled measurements X_j^T from a new launch with high accuracy. Given

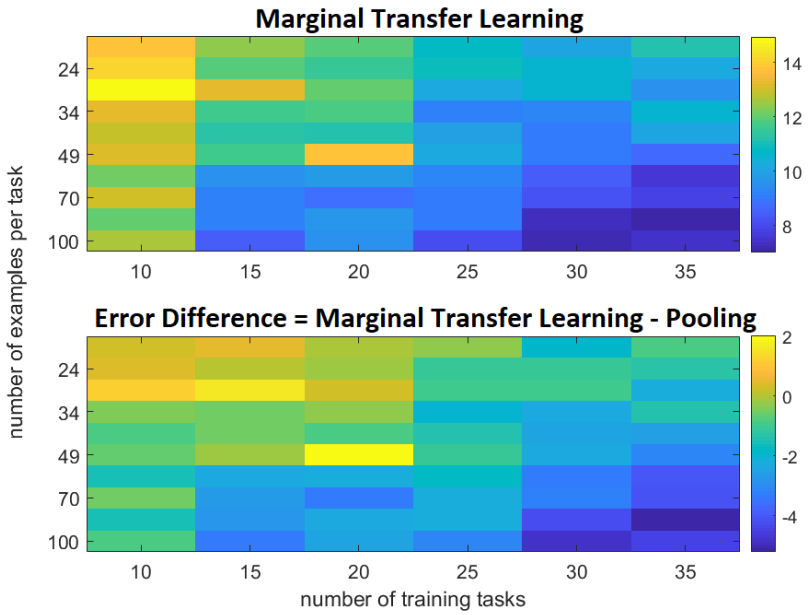

Figure 4: Parkinson's disease telemonitoring data set: Root mean square error rates for proposed method and difference with baseline for different experimental settings, i.e., number of examples per task and number of tasks.

these labels, orbits can subsequently be estimated using the observed RF measurements. We thank Srinagesh Sharma and James Cutler for providing us with their simulated data, and refer the reader to their paper for more details on the application (Sharma and Cutler, 2015).

To demonstrate this idea, we analyzed the data from Sharma and Cutler (2015) for $T = 50$ launches, viewing up to 40 as training data and 10 as testing. We use Gaussian kernels and the RFF kernel approximation technique to speed up the algorithm. Results are shown in Fig 5 (tables given in the appendix). As expected, the error for the proposed method is much lower than for pooling, especially as N and n increase.

8.5 Flow Cytometry Experiments

We demonstrate the proposed methodology for the flow cytometry auto-gating problem, described in Sec. 2. The pooling approach has been previously investigated in this context by Toedling et al. (2006). We used a data set that is a part of the FlowCAP Challenges where the ground truth labels have been supplied by human experts (Aghaeepour et al., 2013). We used the so-called "Normal Donors" data set. The data set contains 8 different classes and 30 subjects. Only two classes (0 and 2) have consistent class ratios, so we have restricted our attention to these two.

The corresponding flow cytometry data sets have sample sizes ranging from 18,641 to 59,411, and the proportion of class 0 in each data set ranges from 25.59 to 38.44%. We randomly selected 10 tasks to serve as the test tasks. These tasks were removed from the

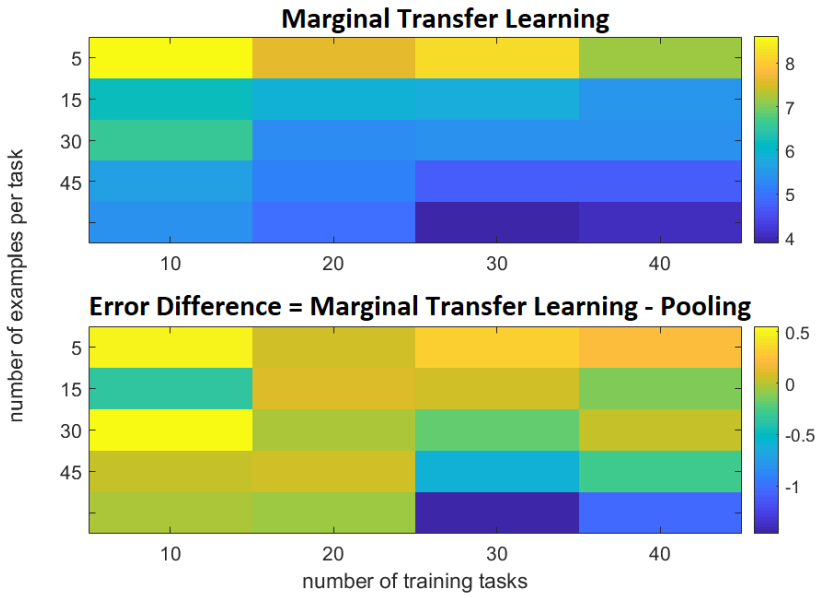

Figure 5: Satellite data set: Classification error rates for proposed method and difference with baseline for different experimental settings, i.e., number of examples per task and number of tasks.

pool of eligible training tasks. We varied the number of training tasks from 5 to 20 with an additive step size of 5, and the number of training examples per task from 1024 to 16384 with a multiplicative step size of 2. We repeated this process 10 times to get the average classification errors which are shown in Fig. 6 and Tables 7 and 8 (see appendix). The kernel k_P was Gaussian, and the other two were linear. The Nyström approximation was used to achieve an efficient implementation.

For nearly all settings the proposed method has a smaller error rate than the baseline. Furthermore, for the marginal transfer learning method, when one fixes the number of training examples and increases the number of tasks then the classification error rate drops.

On the other hand, we observe on Table 7 that the number n of training points per task hardly affects the final performance when $n \geq 10^3$. This is in contrast with the previous experimental examples (synthetic, Parkinson's disease telemonitoring, and satellite classification), for which increasing n led to better performance, but where the values of n remained somewhat modest ($n \leq 256$). This is qualitatively in line with the theoretical results under **(2SGM)** in Section 6.2 (see in particular the concluding discussion there), suggesting that the influence of increasing n on the performance should eventually taper off, in particular if $n \gg N$.

9. Discussion

Our approach to domain generalization relies on the extended input pattern $\widetilde{X} = (P_X, X)$. Thus, we study the natural algorithm of minimizing a regularized empirical loss over a

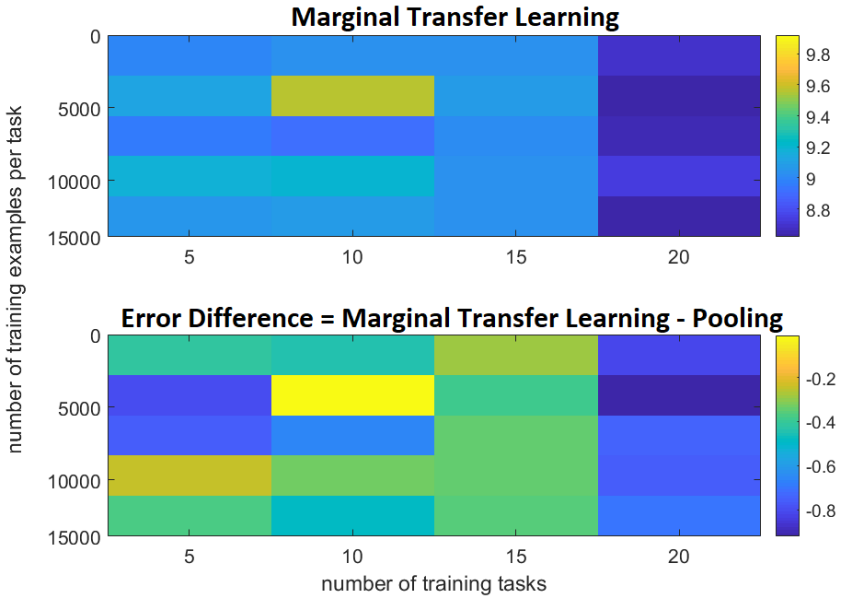

Figure 6: Flow Cytometry Data set: Percentage Classification error rates for proposed method and difference with baseline for different experimental settings, i.e., number of examples per task and number of tasks.

reproducing kernel Hilbert space associated with the extended input domain $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$. We also establish universal consistency under two sampling plans. To achieve this, we present novel generalization error analyses, and construct a universal kernel on $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$. A detailed implementation based on novel approximate feature mappings is also presented.

On one synthetic and three real-world data sets, the marginal transfer learning approach consistently outperforms a pooling baseline. On some data sets, however, the difference between the two methods is small. This is because the utility of transfer learning varies from one DG problem to another. As an extreme example, if all of the task are the same, then pooling should do just as well as our method.

Several future directions exist. From an application perspective, the need for adaptive classifiers arises in many applications, especially in biomedical applications involving biological and/or technical variation in patient data. Examples include brain computer interfaces and patient monitoring. For example, when electrocardiograms are used to continuously monitor cardiac patients, it is desirable to classify each heartbeat as irregular or not. Given the extraordinary amount of data involved, automation of this process is essential. However, irregularities in a test patient's heartbeat will differ from irregularities of historical patients, hence the need to adapt to the test distribution (Wiens, 2010).

From a theoretical and methodological perspective, several questions are of interest. We would like to specify conditions on the meta-distributions P_S or μ under which the DG risk is close to the expected Bayes risk of the test distribution (beyond the simple condition discussed in Lemma 9). We would also like to develop fast learning rates under suitable dis-

tributional assumptions. Furthermore, given the close connections with supervised learning, many common variants of supervised learning can also be investigated in the DG context, including multiclass classification, class probability estimation, and robustness to various forms of noise.

We can also ask how the methodology and analysis can be extended to the context where a small number of labels are available for the test distribution (additionally to a larger number of unlabeled data from the same distribution); this situation appears to be common in practice, and can be seen as intermediary between the DG and learning to learn (LTL, see Section 4.2) settings (one could dub it "semi-supervised domain generalization"). In this setting, two approaches appear promising to take advantage of the labeled data. The simplest one is to use the same optimization problem (7), where we include additionally the labeled examples of the test distribution. However, if several test samples are to be treated in succession, and we want to avoid a full, resource-consuming re-training using all the training samples each time, an interesting alternative is the following: learn once a function $f_0(P_X, x)$ using the available training samples via (7); then, given a partially labeled test sample, learn a decision function on this sample only via the usual kernel (k_X) norm regularized empirical loss minimization method, but replace the usual regularizer term $\|f\|^2_{\mathcal{H}}$ by $\left\|f - f_0(\widehat{P}^T_X, .)\right\|^2_{\mathcal{H}}$ (note that $f_0(\widehat{P}^T_X, .) \in \mathcal{H}_{k_X}$). In this sense, the marginal-adaptive decision function learned from the training samples would serve as a "prior" or "informed guess" for learning on the test data. This can be also interpreted as learning an adequate complexity penalty to improve learning on new samples, thus connecting to the general principles of LTL (see Section 4.2). An interesting difference with underlying existing LTL approaches is that those tend to adapt the hypothesis class or the"shape" of the regularization penalty to the problem at hand, while the approach delineated above would modify the "origin" of the penalty, using the marginal distribution information. These two principles could also be combined.

Acknowledgments

C. Scott and A. Deshmukh were supported in part by NSF Grants No. 1422157, 1217880, 1047871, 1838179, and 2008074. G. Blanchard was supported in part by the Deutsche Forschungsgemeinschaft (DFG) via Research Unit FG1735/2 *Structural Inference in Statistics*, and via SFB1294/1 - 318763901DFG *Data Assimilation*; and by the Agence Nationale de la Recherche (ANR) via the IA Chair *BiSCottE*. G. Lee was supported by the National Research Foundation of Korea (NRF-2014R1A1A1003458).

Appendix A. Proofs, Technical Details, and Experimental Details

This appendix contains the remaining proofs, as well as additional technical and experimental details.

A.1 Proof of Proposition 7

Let P_{XY}^T be a fixed probability distribution on $\mathcal{X} \times \mathbb{R}$, and $\varepsilon > 0$ a fixed number. Since $\mathcal{X}$ is a Radon space, by definition any Borel probability measure on it, in particular P_X^T (the X-marginal of P_{XY}^T), is inner regular, so that there exists a compact set $K \subset \mathcal{X}$ such that $P_X^T(K^c) \leq \varepsilon$.

For all $x \in K$, by the assumed continuity of the decision function f at point (P_X^T, x) there exists an open neighborhood $U_x \times V_x \subset \mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$ of this point such that $\left|f(u,v) - f(P_X^T, x)\right| \leq \varepsilon$ for all $(u,v) \in U_x \times V_x$. Since the family $(V_x)_{x \in K}$ is an open covering of the compact K, there exists a finite subfamily $(V_{x_i})_{i \in I}$ covering K. Denoting $U_0 := \bigcap_{i \in I} U_{x_i}$ which is an open neighborhood of P_X^T in $\mathfrak{P}_{\mathcal{X}}$, it therefore holds for any $P \in U_0$ and uniformly over $x \in K$ that $\left|f(P,x) - f(P_X^T, x)\right| \leq |f(P,x) - f(P, x_{i_0})| + \left|f(P_X^T, x) - f(P, x_{i_0})\right| \leq 2\varepsilon$, where $i_0 \in I$ is such that $x \in V_{x_{i_0}}$.

Denote $S^T = (X_i^T, Y_i^T)_{1 \leq i \leq n_T}$ a sample of size n_T drawn i.i.d. from P_{XY}^T, and A the event $\left\{\widehat{P}_X^T \in U_0\right\}$. By the law of large numbers, $\widehat{P}_X^T$ weakly converges to P_X^T in probability, so that $\mathbb{P}\left[A^c\right] \leq \varepsilon$ holds for n_T large enough. We have (denoting B a bound on the loss function):

$$\begin{aligned}
\mathbb{E}_{S_T}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\right] &\leq B\varepsilon + \mathbb{E}_{S_T}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\mathbf{1}_{\{X_i^T \in K\}}\right] \\
&\leq B(\varepsilon + \mathbb{P}\left[A^c\right]) + \mathbb{E}_{S_T}\left[\mathbf{1}_{\{\widehat{P}_X^T \in U_0\}}\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\mathbf{1}_{\{X_i^T \in K\}}\right] \\
&\leq 2B\varepsilon + 2L\varepsilon + \mathbb{E}_{S_T}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(P_X^T, X_i^T), Y_i^T)\right] \\
&\leq 2(B+L)\varepsilon + \mathbb{E}_{(X^T,Y^T)\sim P_{XY}^T}\left[\ell(f(P_X^T, X^T), Y^T)\right].
\end{aligned}$$

Conversely,

$$\begin{aligned}
\mathbb{E}_{S_T}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\right] &\geq \mathbb{E}_{S_T}\left[\mathbf{1}_{\{\widehat{P}_X^T \in U_0\}}\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\mathbf{1}_{\{X_i^T \in K\}}\right] \\
&\geq \mathbb{E}_{S_T}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(P_X^T, X_i^T), Y_i^T)\right] - 2B\varepsilon - 2L\varepsilon \\
&\geq \mathbb{E}_{(X^T,Y^T)\sim P_{XY}^T}\left[\ell(f(P_X^T, X^T), Y^T)\right] - 2(B+L)\varepsilon.
\end{aligned}$$

Since the above inequalities hold for any $\varepsilon > 0$ provided n_T is large enough, this yields that for any fixed P_{XY}^T, we have

$$\lim_{n_T \to \infty} \mathbb{E}_{S_T \sim (P_{XY}^T)^{\otimes n_T}}\left[\frac{1}{n_T}\sum_{i=1}^{n_T} \ell(f(\widehat{P}_X^T, X_i^T), Y_i^T)\right] = \mathbb{E}_{(X^T,Y^T)\sim P_{XY}^T}\left[\ell(f(P_X^T, X^T), Y^T)\right].$$

Finally, since the above right-hand side is bounded by B, applying dominated convergence to integrate over $P_{XY}^T \sim \mu$ yields the desired conclusion.

A.2 Proof of Corollary 14

Proof Denote $\mathcal{E}^* = \inf_{f:\mathfrak{P}_X\times\mathcal{X}\to\mathbb{R}} \mathcal{E}(f)$. Let $\varepsilon > 0$. Since $\overline{k}$ is a universal kernel on $\mathfrak{P}_\mathcal{X}\times\mathcal{X}$ and ℓ is Lipschitz, there exists $f_0 \in \mathcal{H}_{\overline{k}}$ such that $\mathcal{E}(f_0) \leq \mathcal{E}^* + \frac{\varepsilon}{2}$ (Steinwart and Christmann, 2008).

By comparing the objective function in (7) at the minimizer $\widehat{f}_\lambda$ and at the null function, using assumption **(LB)** we deduce that we must have $\|\widehat{f}_\lambda\| \leq \sqrt{B_0/\lambda}$. Applying Theorem 11 for $R = R_\lambda = \sqrt{B_0/\lambda}$, and $\delta = 1/N^2$, gives that with probability at least $1 - 1/N^2$,

$$\sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\widehat{\mathcal{E}}(f,N) - \mathcal{E}(f)\right| \leq \varepsilon(N) := (B_0 + L_\ell B_\Re B_k\sqrt{B_0/\lambda})\frac{(\sqrt{\log N}+2)}{\sqrt{N}}.$$

Let N be large enough so that $\|f_0\| \leq R_\lambda$. We can now deduce that with probability at least $1 - 1/N^2$,

$$\begin{aligned}
\mathcal{E}(\widehat{f}_\lambda) &\leq \widehat{\mathcal{E}}(\widehat{f}_\lambda, N) + \varepsilon(N) \\
&= \widehat{\mathcal{E}}(\widehat{f}_\lambda, N) + \lambda\|\widehat{f}_\lambda\|^2 - \lambda\|\widehat{f}_\lambda\|^2 + \varepsilon(N) \\
&\leq \widehat{\mathcal{E}}(f_0, N) + \lambda\|f_0\|^2 - \lambda\|\widehat{f}_\lambda\|^2 + \varepsilon(N) \\
&\leq \widehat{\mathcal{E}}(f_0, N) + \lambda\|f_0\|^2 + \varepsilon(N) \\
&\leq \mathcal{E}(f_0) + \lambda\|f_0\|^2 + 2\varepsilon(N) \\
&\leq \mathcal{E}^* + \frac{\varepsilon}{2} + \lambda\|f_0\|^2 + 2\varepsilon(N).
\end{aligned}$$

The last two terms become less than $\frac{\varepsilon}{2}$ for N sufficiently large by the assumptions on the growth of $\lambda = \lambda(N)$. This establishes that for any $\varepsilon > 0$, there exists N_0 such that

$$\sum_{N\geq N_0} \Pr(\mathcal{E}(\widehat{f}_\lambda) \geq \mathcal{E}^* + \varepsilon) \leq \sum_{N\geq N_0} \frac{1}{N^2} < \infty,$$

and so the result follows by the Borel-Cantelli lemma. ■

A.3 Proof of Theorem 15

We control the difference between the training loss and the conditional risk at infinite sample size via the following decomposition:

$$\begin{aligned}
\sup_{f\in\mathcal{B}_{\overline{k}}(R)} |\mathcal{L}(S,f) - \mathcal{E}^\infty(f|P_{XY})| &= \sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\frac{1}{n}\sum_{i=1}^n \ell(f(\widehat{P}_X, X_i), Y_i) - \mathcal{E}^\infty(f|P_{XY})\right| \\
&\leq \sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\frac{1}{n}\sum_{i=1}^n \Big(\ell(f(\widehat{P}_X, X_i), Y_i) - \ell(f(P_X, X_i), Y_i)\Big)\right| \\
&\quad + \sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\frac{1}{n}\sum_{i=1}^n \ell(f(P_X, X_i), Y_i) - \mathcal{E}^\infty(f|P_{XY})\right| \\
&=: (I) + (II). \qquad (22)
\end{aligned}$$

A.3.1 Control of Term (I)

Using the assumption that the loss ℓ is L_ℓ-Lipschitz in its first coordinate, we can bound the first term as follows:

$$(I) \leq L_\ell \sup_{f \in \mathcal{B}_{\overline{k}}(R)} \frac{1}{n} \sum_{i=1}^{n} \left| f(\widehat{P}_X, X_i) - f(P_X, X_i) \right| \leq L_\ell \sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left\| f(\widehat{P}_X, \cdot) - f(P_X, \cdot) \right\|_\infty . \quad (23)$$

This can now be controlled using the first part of the following result:

Lemma 20 *Assume* **(K-Bounded)** *holds. Let P_X be an arbitrary distribution on $\mathcal{X}$ and $\widehat{P}_X$ denote an empirical distribution on $\mathcal{X}$ based on an iid sample of size n from P_X. Then with probability at least $1-\delta$ over the draw of this sample, it holds that*

$$\sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left\| f(\widehat{P}_X, \cdot) - f(P_X, \cdot) \right\|_\infty \leq 3RB_k L_{\mathfrak{K}} B_{k'}^\alpha \left(\frac{\log 2\delta^{-1}}{n} \right)^{\frac{\alpha}{2}} . \quad (24)$$

In expectation, it holds

$$\mathbb{E}\left[\sup_{f \in \mathcal{B}_{\overline{k}}(R)} \left\| f(\widehat{P}_X, \cdot) - f(P_X, \cdot) \right\|_\infty \right] \leq 2RB_k L_{\mathfrak{K}} B_{k'}^\alpha n^{-\alpha/2}. \quad (25)$$

Proof Let $X_1, \ldots, X_n$ denote the n-sample from P_X. Let us denote by Φ'_X the canonical feature mapping $x \mapsto k'_X(x, \cdot)$ from $\mathcal{X}$ into $\mathcal{H}_{k'_\mathcal{X}}$. We have for all $x \in \mathcal{X}$, $\|\Phi'_X(x)\| \leq B_{k'}$, and so, as a consequence of Hoeffding's inequality in a Hilbert space (see, e.g., Pinelis and Sakhanenko, 1985), it holds with probability at least $1-\delta$:

$$\left\| \Psi(P_X) - \Psi(\widehat{P}_X) \right\| = \left\| \frac{1}{n} \sum_{i=1}^{n} \Phi'_X(X_i) - \mathbb{E}_{X \sim P_X} \left[\Phi'_X(X) \right] \right\| \leq 3B_{k'} \sqrt{\frac{\log 2\delta^{-1}}{n}}. \quad (26)$$

Furthermore, using the reproducing property of the kernel $\overline{k}$, we have for any $x \in \mathcal{X}$ and $f \in \mathcal{B}_{\overline{k}}(R)$:

$$\begin{aligned}
|f(\widehat{P}_X, x) - f(P_X, x)| &= \left| \left\langle \overline{k}((\widehat{P}_X, x), \cdot) - \overline{k}((P_X, x), \cdot), f \right\rangle \right| \\
&\leq \|f\| \left\| \overline{k}((\widehat{P}_X, x), \cdot) - \overline{k}((P_X, x), \cdot) \right\| \\
&\leq R k_X(x,x)^{\frac{1}{2}} \Big(\mathfrak{K}(\Psi(P_X), \Psi(P_X)) \\
&\qquad + \mathfrak{K}(\Psi(\widehat{P}_X), \Psi(\widehat{P}_X)) - 2\mathfrak{K}(\Psi(P_X), \Psi(\widehat{P}_X)) \Big)^{\frac{1}{2}} \\
&\leq RB_k \left\| \Phi_{\mathfrak{K}}(\Psi(P_X)) - \Phi_{\mathfrak{K}}(\Psi(\widehat{P}_X)) \right\| \\
&\leq RB_k L_{\mathfrak{K}} \left\| \Psi(P_X) - \Psi(\widehat{P}_X) \right\|^\alpha ,
\end{aligned}$$

where in the last step we have used property **(K-Hölder)** together with the fact that for all $P \in \mathfrak{P}_\mathcal{X}$, $\|\Psi(P)\| \leq \int_\mathcal{X} \|k'_X(x, \cdot)\| \, dP_X(x) \leq B_{k'}$, so that $\Psi(P) \in \mathcal{B}_{k'_X}(B_{k'})$. Combining with (26) gives (24).

For the bound in expectation, we use the inequality above, and can bound further (using Jensen's inequality, since $\alpha \leq 1$)

$$\begin{aligned}
\mathbb{E}\left[\left\|\Psi(P_X) - \Psi(\widehat{P}_X)\right\|^{\alpha}\right] &\leq \mathbb{E}\left[\left\|\Psi(P_X) - \Psi(\widehat{P}_X)\right\|^{2}\right]^{\alpha/2} \\
&= \left(\frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{E}\left[\left\langle \Phi'_X(X_i) - \mathbb{E}\left[\Phi'_X(X)\right], \Phi'_X(X_j) - \mathbb{E}\left[\Phi'_X(X)\right]\right\rangle\right]\right)^{\alpha/2} \\
&= \left(\frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left[\left\|\Phi'_X(X_i) - \mathbb{E}\left[\Phi'_X(X)\right]\right\|^{2}\right]\right)^{\alpha/2} \\
&\leq \left(\frac{4B_{k'}^2}{n}\right)^{\frac{\alpha}{2}},
\end{aligned}$$

which yields (25) in combination with the above. ■

A.3.2 Control of Term (II)

Term (II) takes the form of a uniform deviation over a RKHS ball of an empirical loss for the data $(\widetilde{X}_i, Y_i)$, where $\widetilde{X}_i := (P_X, X_i)$. Since P_X is fixed (in contrast with term (I) where $\widehat{P}_X$ depended on the whole sample), these data are i.i.d. Similar to the proofs of Theorems 5 and 11, we can therefore apply again standard Rademacher analysis, this time at the level of one specific task (Azuma-McDiarmid inequality followed by Rademacher complexity analysis for a Lipschitz, bounded loss over a RKHS ball; see Koltchinskii, 2001; Bartlett and Mendelson, 2002, Theorems 8, 12 and Lemma 22 there). The kernel $\overline{k}$ is bounded by $B_k^2 B_{\mathfrak{K}}^2$ by assumption **(K-Bounded)**; by Lemma 10 and assumption **(LB)**, the loss is bounded by $B_0 + L_\ell R B_k B_{\mathfrak{K}}$, and is L_ℓ-Lipschitz. Therefore, with probability at least $1-\delta$ we get

$$\begin{aligned}
(II) &\leq (B_0 + 3L_\ell R B_k B_{\mathfrak{K}}) \min\left(\sqrt{\frac{\log(\delta^{-1})}{2n}}, 1\right) \\
&\leq (B_0 + 3L_\ell R B_k B_{\mathfrak{K}}) \min\left(\left(\frac{\log(\delta^{-1})}{2n}\right)^{\frac{\alpha}{2}}, 1\right). \qquad (27)
\end{aligned}$$

Observe that we can cap the second factor at 1 since (II) is upper bounded by the bound on the loss in all cases; the second inequality then uses $\alpha \leq 1$. Combining with a union bound the probabilistic controls (23), (24) of term (I) and (27) of (II) yields (18).

To establish the bound (19) we use a similar argument. We use the decomposition

$$\begin{aligned}
&\sup_{f\in\mathcal{B}_{\overline{k}}(R)} |\mathcal{E}(f|P_{XY},n)-\mathcal{E}^{\infty}(f|P_{XY})| \\
&\quad\leq \sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\mathbb{E}_{S_n\sim(P_{XY})^{\otimes n}}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(\widehat{P}_X,X_i^T),Y_i)-\ell(f(P_X,X_i),Y_i)\Big)\right]\right| \\
&\quad + \sup_{f\in\mathcal{B}_{\overline{k}}(R)} \left|\mathbb{E}_{S\sim(P_{XY})^{\otimes n}}\left[\frac{1}{n}\sum_{i=1}^{n}\big(\ell(f(P_X,X_i^T),Y_i)\big)\right]-\mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell(f(P_X,X),Y)\right]\right| \\
&\quad =: (I')+(II').
\end{aligned}$$

It is easily seen that the second term vanishes: for any fixed $f\in\mathcal{B}_{\overline{k}}(R)$ and P_{XY}, the difference of the expectations is zero. For the first term, using Lipschitzness of the loss, then (25), we obtain

$$(I')\leq L_\ell\mathbb{E}\left[\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\left\|f(\widehat{P}_X,.)-f(P_X,.)\right\|_\infty\right]\leq 2L_\ell R B_k L_{\mathfrak{K}} B_{k'}^{\alpha} n^{-\alpha/2},$$

yielding (19). The bound (20) is obtained as a direct consequence by taking expectation over $P_{XY}\sim\mu$ and using Jensen's inequality to pull out the absolute value.

A.4 Regularity Conditions for the Kernel on Distributions

We investigate sufficient conditions on the kernel $\mathfrak{K}$ to ensure the regularity condition **(K-Hölder)** (15). Roughly speaking, the regularity of the feature mapping of a reproducing kernel is "one half" of the regularity of the kernel in each of its variables. The next result considers the situation where $\mathfrak{K}$ is itself simply a Hölder continuous function of its variables.

Lemma 21 *Let $\alpha\in(0,\frac{1}{2}]$. Assume that the kernel $\mathfrak{K}$ is Hölder continuous of order 2α and constant $L_{\mathfrak{K}}^2/2$ in each of its two variables on $\mathcal{B}_{k'_X}(B_{k'})$. Then* **(K-Hölder)** *is satisfied.*

Proof For any $v,w\in\mathcal{B}_{k'_X}(B_{k'})$:

$$\|\Phi_{\mathfrak{K}}(v)-\Phi_{\mathfrak{K}}(w)\|=(\mathfrak{K}(v,v)+\mathfrak{K}(w,w)-2\mathfrak{K}(v,w))^{\frac{1}{2}}\leq L_{\mathfrak{K}}\|v-w\|^{\alpha}.$$

■

The above type of regularity only leads to a Hölder feature mapping of order at most $\frac{1}{2}$ (when the kernel function is Lipschitz continuous in each variable). Since this order plays an important role in the rate of convergence of the upper bound in the main error control theorem, it is desirable to study conditions ensuring more regularity, in particular a feature mapping which has at least Lipschitz continuity. For this, we consider the following stronger condition, namely that the kernel function is twice differentiable in a specific sense:

Lemma 22 *Assume that, for any $u,v\in\mathcal{B}_{k'_X}(B_{k'})$ and unit norm vector e of $\mathcal{H}_{k'_X}$, the function $h_{u,v,e}:(\lambda,\mu)\in\mathbb{R}^2\mapsto\mathfrak{K}(u+\lambda e,v+\mu e)$ admits a mixed partial derivative $\partial_1\partial_2 h_{u,v,e}$ at the point $(\lambda,\mu)=(0,0)$ which is bounded in absolute value by a constant $C_{\mathfrak{K}}^2$ independent of (u,v,e). Then (15) is satisfied with $\alpha=1$ and $L_{\mathfrak{K}}=C_{\mathfrak{K}}$, that is, the canonical feature mapping of $\mathfrak{K}$ is Lipschitz continuous on $\mathcal{B}_{k'_X}(B_{k'})$.*

Proof The argument is along the same lines as Steinwart and Christmann (2008), Lemma 4.34. Observe that, since $h_{u,v,e}(\lambda+\lambda',\mu+\mu') = h_{u+\lambda e,v+\mu e,e}(\lambda',\mu')$, the function $h_{u,v,e}$ actually admits a uniformly bounded mixed partial derivative in any point $(\lambda,\mu)\in\mathbb{R}^2$ such that $(u+\lambda e, v+\mu e)\in\mathcal{B}_{k'_X}(B_{k'})$. Let us denote $\Delta_1 h_{u,v,e}(\lambda,\mu) := h_{u,v,e}(\lambda,\mu) - h_{u,v,e}(0,\mu)$. For any $u,v\in\mathcal{B}_{k'_X}(B_{k'})$, $u\neq v$, let us set $\lambda := \|v-u\|$ and the unit vector $e := \lambda^{-1}(v-u)$; we have

$$\begin{aligned}\|\Phi_{\mathfrak{K}}(u)-\Phi_{\mathfrak{K}}(v)\|^2 &= \mathfrak{K}(u,u)+\mathfrak{K}(u+\lambda e,u+\lambda e)-\mathfrak{K}(u,u+\lambda e)-\mathfrak{K}(u+\lambda e,u)\\ &= \Delta_1 h_{u,v,e}(\lambda,\lambda)-\Delta_1 h_{u,v,e}(\lambda,0)\\ &= \lambda\partial_2\Delta_1 h_{u,v,e}(\lambda,\lambda'),\end{aligned}$$

where we have used the mean value theorem, yielding existence of $\lambda'\in[0,\lambda]$ such that the last equality holds. Furthermore,

$$\begin{aligned}\partial_2\Delta_1 h_{u,v,e}(\lambda,\lambda') &= \partial_2 h_{u,v,e}(\lambda,\lambda')-\partial_2 h_{u,v,e}(0,\lambda')\\ &= \lambda\partial_1\partial_2 h_{u,v,e}(\lambda'',\lambda'),\end{aligned}$$

using again the mean value theorem, yielding existence of $\lambda''\in[0,\lambda]$ in the last equality. Finally, we get

$$\|\Phi_{\mathfrak{K}}(u)-\Phi_{\mathfrak{K}}(v)\|^2 = \lambda^2\partial_1\partial_2 h_{u,v,e}(\lambda',\lambda'') \le C_{\mathfrak{K}}^2\|v-u\|^2 .$$

■

Lemma 23 *Assume that the kernel $\mathfrak{K}$ takes the form of either (a) $\mathfrak{K}(u,v) = g(\|u-v\|^2)$ or (b) $\mathfrak{K}(u,v) = g(\langle u,v\rangle)$, where g is a twice differentiable real function of a real variable defined on $[0,4B_{k'}^2]$ in case (a), and on $[-B_{k'}^2,B_{k'}^2]$ in case (b). Assume $\|g'\|_\infty\le C_1$ and $\|g''\|_\infty\le C_2$. Then $\mathfrak{K}$ satisfies the assumption of Lemma 22 with $C_{\mathfrak{K}} := 2C_1+16C_2B_{k'}^2$ in case (a), and $C_{\mathfrak{K}} := C_1+C_2B_{k'}^2$ for case (b).*

Proof In case (a), we have $h_{u,v,e}(\lambda,\mu) = g(\|u-v+(\lambda-\mu)e\|^2)$. It follows

$$\begin{aligned}|\partial_1\partial_2 h_{u,v,e}(0,0)| &= \left|-2g'(\|u-v\|^2)\|e\|^2-4g''(\|u-v\|^2)\langle u-v,e\rangle^2\right|\\ &\le 2C_1+16C_2B_{k'}^2 .\end{aligned}$$

In case (b), we have $h_{u,v,e}(\lambda,\mu) = g(\langle u+\lambda e, v+\mu e\rangle)$. It follows

$$\begin{aligned}|\partial_1\partial_2 h_{u,v,e}(0,0)| &= \left|g'(\langle u,v\rangle)\|e\|^2+g''(\langle u,v\rangle)\langle u,e\rangle\langle v,e\rangle\right|\\ &\le C_1+C_2B_{k'}^2 .\end{aligned}$$

■

A.5 Proof of Lemma 12

Proof Let $\mathcal{H}, \mathcal{H}'$ the RKHS associated to k, k' with the associated feature mappings Φ, Φ'. Then it can be checked that $(x, x') \in \mathcal{X} \times \mathcal{X}' \mapsto \Phi(x) \otimes \Phi'(x')$ is a feature mapping for $\overline{k}$ into the Hilbert space $\mathcal{H} \otimes \mathcal{H}'$. Using (Steinwart and Christmann, 2008), Th. 4.21, we deduce that the RKHS $\overline{H}$ of $\overline{k}$ contains precisely all functions of the form $(x, x') \in \mathcal{X} \times \mathcal{X}' \mapsto F_w(x, x') = \langle w, \Phi(x) \otimes \Phi(x') \rangle$, where w ranges over $\mathcal{H} \otimes \mathcal{H}'$. Taking w of the form $w = g \otimes g'$, $g \in \mathcal{H}, g \in \mathcal{H}'$, we deduce that $\overline{H}$ contains in particular all functions of the form $f(x, x') = g(x)g(x')$, and further

$$\widetilde{\mathcal{H}} := \operatorname{span}\left\{(x, x') \in \mathcal{X} \times \mathcal{X}' \mapsto g(x)g(x'); g \in \mathcal{H}, g' \in \mathcal{H}'\right\} \subset \overline{H}.$$

Denote $\mathcal{C}(\mathcal{X}), \mathcal{C}(\mathcal{X}'), \mathcal{C}(\mathcal{X} \times \mathcal{X}')$ the set of real-valued continuous functions on the respective spaces. Let

$$\mathcal{C}(\mathcal{X}) \otimes \mathcal{C}(\mathcal{X}') := \operatorname{span}\left\{(x, x') \in \mathcal{X} \times \mathcal{X}' \mapsto f(x)f'(x'); f \in \mathcal{C}(\mathcal{X}), f' \in \mathcal{C}(\mathcal{X}')\right\}.$$

Let $G(x, x')$ be an arbitrary element of $\mathcal{C}(\mathcal{X}) \otimes \mathcal{C}(\mathcal{X}')$, $G(x, x') = \sum_{i=1}^{k} \lambda_i g_i(x) g_i'(x')$ with $g_i \in \mathcal{C}(\mathcal{X}), g_i' \in \mathcal{C}(\mathcal{X}')$ for $i = 1, \ldots, k$. For $\varepsilon > 0$, by universality of k and k', there exist $f_i \in \mathcal{H}, f_i' \in \mathcal{H}'$ so that $\|f_i - g_i\|_\infty \leq \varepsilon$, $\|f_i' - g_i'\|_\infty \leq \varepsilon$ for $i = 1, \ldots, k$. Let $F(x, x') := \sum_{i=1}^{k} \lambda_i f_i(x) f_i'(x') \in \widetilde{\mathcal{H}}$. We have

$$\begin{aligned}
\left\|F(x, x') - G(x, x')\right\|_\infty &\leq \left\| \sum_{i=1}^{k} \lambda_i (g_i(x)g_i'(x) - f_i(x)f_i'(x)) \right\|_\infty \\
&= \Bigg\| \sum_{i=1}^{k} \lambda_i \Big[(f_i(x) - g_i(x))(g_i'(x') - f_i'(x')) \\
&\qquad + g_i(x)(g_i'(x) - f_i'(x')) + (g_i(x) - f_i(x))g_i'(x') \Big] \Bigg\|_\infty \\
&\leq \varepsilon \sum_{i=1}^{k} |\lambda_i| \left(\varepsilon + \|g_i\|_\infty + \left\|g_i'\right\|_\infty\right).
\end{aligned}$$

This establishes that $\widetilde{\mathcal{H}}$ is dense in $\mathcal{C}(\mathcal{X}) \otimes \mathcal{C}(\mathcal{X}')$ for the supremum norm. It can be easily checked that $\mathcal{C}(\mathcal{X}) \otimes \mathcal{C}(\mathcal{X}')$ is an algebra of functions which does not vanish and separates points on $\mathcal{X} \times \mathcal{X}'$. By the Stone-Weierstrass theorem, it is therefore dense in $\mathcal{C}(\mathcal{X} \times \mathcal{X}')$ for the supremum norm. We deduce that $\widetilde{\mathcal{H}}$ (and thus also $\overline{\mathcal{H}}$) is dense in $\mathcal{C}(\mathcal{X} \times \mathcal{X}')$, so that $\overline{k}$ is universal. ■

A.6 Approximate Feature Mapping for Scalable Implementation

We first treat random Fourier features and then the Nyström method.

A.6.1 Random Fourier Features

The approximation of Rahimi and Recht (2007) is based on Bochner's theorem, which characterizes shift invariant kernels.

Theorem 24 *A continuous kernel $k(x,y) = k(x-y)$ on $\mathbb{R}^d$ is positive definite iff $k(x-y)$ is the Fourier transform of a finite positive measure $p(w)$, i.e.,*

$$k(x-y) = \int_{\mathbb{R}^d} p(w) e^{jw^T(x-y)} dw \,. \tag{28}$$

If a shift invariant kernel $k(x-y)$ is properly scaled then Theorem 24 guarantees that $p(w)$ in (28) is a proper probability distribution.

Random Fourier features (RFFs) approximate the integral in (28) using samples drawn from $p(w)$. If $w_1, w_2, ..., w_L$ are i.i.d. draws from $p(w)$,

$$\begin{aligned}
k(x-y) &= \int_{\mathbb{R}^d} p(w) e^{jw^T(x-y)} dw \\
&= \int_{\mathbb{R}^d} p(w) \cos(w^T x - w^T y) dw \\
&\approx \frac{1}{L} \sum_{i=1}^{L} \cos(w_i^T x - w_i^T y) \\
&= \frac{1}{L} \sum_{i=1}^{L} \cos(w_i^T x)\cos(w_i^T y) + \sin(w_i^T x)\sin(w_i^T y) \\
&= \frac{1}{L} \sum_{i=1}^{L} [\cos(w_i^T x), \sin(w_i^T x)]^T [\cos(w_i^T y), \sin(w_i^T y)] \\
&= z_w(x)^T z_w(y) \,,
\end{aligned} \tag{29}$$

where $z_w(x) = \frac{1}{\sqrt{L}}[\cos(w_1^T x), \sin(w_1^T x), ..., \cos(w_L^T x), \sin(w_L^T x)] \in \mathbb{R}^{2L}$ is an approximate nonlinear feature mapping of dimensionality $2L$. In the following, we extend the RFF methodology to the kernel $\bar{k}$ on the extended feature space $\mathfrak{P}_{\mathcal{X}} \times \mathcal{X}$. Let $X_1, \ldots, X_{n_1}$ and $X'_1, \ldots, X'_{n_2}$ be i.i.d. realizations of P_X and P'_X respectively, and let $\widehat{P}_X$ and $\widehat{P}'_X$ denote the corresponding empirical distributions. Given $x, x' \in \mathcal{X}$, denote $\tilde{x} = (\widehat{P}_X, x)$ and $\tilde{x}' = (\widehat{P}'_X, x')$. The goal is to find an approximate feature mapping $\bar{z}(\tilde{x})$ such that $\bar{k}(\tilde{x}, \tilde{x}') \approx \bar{z}(\tilde{x})^T \bar{z}(\tilde{x}')$. Recall that

$$\bar{k}(\tilde{x}, \tilde{x}') = k_P(\widehat{P}_X, \widehat{P}'_X) k_X(x, x');$$

specifically, we consider k_X and k'_X to be Gaussian kernels and the kernel on distributions k_P to have the Gaussian-like form

$$k_P(\widehat{P}_X, \widehat{P}'_X) = \exp\left\{ \frac{1}{2\sigma_P^2} \|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2_{H_{k'_X}} \right\}.$$

As noted earlier in this section, the calculation of $k_P(\widehat{P}_X, \widehat{P}'_X)$ reduces to the computation of

$$\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \rangle = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k'_X(X_i, X'_j). \tag{30}$$

We use Theorem 24 to approximate k'_X and thus $\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \rangle$. Let $w_1, w_2, ..., w_L$ be i.i.d. draws from $p'(w)$, the inverse Fourier transform of k'_X. Then we have:

$$\begin{aligned}
\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \rangle &= \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k'_X(X_i, X'_j) \\
&\approx \frac{1}{L n_1 n_2} \sum_{l=1}^{L} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \cos(w_l^T X_i - w_l^T X'_j) \\
&= \frac{1}{L n_1 n_2} \sum_{l=1}^{L} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} [\cos(w_l^T X_i)\cos(w_l^T X'_j) + \sin(w_l^T X_i)\sin(w_l^T X'_j)] \\
&= \frac{1}{L n_1 n_2} \sum_{l=1}^{L} \{ \sum_{i=1}^{n_1} [\cos(w_l^T X_i), \sin(w_l^T X_i)]^T \sum_{j=1}^{n_2} [\cos(w_l^T X'_j), \sin(w_l^T X'_j)] \} \\
&= Z_P(\widehat{P}_X)^T Z_P(\widehat{P}'_X),
\end{aligned}$$

where

$$Z_P(\widehat{P}_X) = \frac{1}{n_1 \sqrt{L}} \sum_{i=1}^{n_1} \Big[\cos(w_1^T X_i), \sin(w_1^T X_i), ..., \cos(w_L^T X_i), \sin(w_L^T X_i) \Big], \tag{31}$$

and $Z_P(\widehat{P}'_X)$ is defined analogously with n_1 replaced by n_2. For the proof of Theorem 25, let z'_X denote the approximate feature map corresponding to k'_X, which satisfies $Z_P(\widehat{P}_X) = \frac{1}{n_1} \sum_{i=1}^{n_1} z'_X(X_i)$.

Note that the lengths of the vectors $Z_P(\widehat{P}_X)$ and $Z_P(\widehat{P}'_X)$ are $2L$. To approximate $\bar{k}$ we may write

$$\begin{aligned}
\bar{k}(\tilde{x}, \tilde{x}') &\approx \exp \frac{-\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2_{\mathbb{R}^{2L}}}{2\sigma_P^2} \cdot \exp \frac{-\|x - x'\|^2_{\mathbb{R}^d}}{2\sigma_X^2} \\
&= \exp \frac{-(\sigma_X^2 \|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2_{\mathbb{R}^{2L}} + \sigma_P^2 \|x - x'\|^2_{\mathbb{R}^d})}{2\sigma_P^2 \sigma_X^2} \\
&= \exp \frac{-(\|\sigma_X Z_P(\widehat{P}_X) - \sigma_X Z_P(\widehat{P}'_X)\|^2_{\mathbb{R}^{2L}} + \|\sigma_P x - \sigma_P x'\|^2_{\mathbb{R}^d})}{2\sigma_P^2 \sigma_X^2} \\
&= \exp \frac{-\|(\sigma_X Z_P(\widehat{P}_X), \sigma_P x) - (\sigma_X Z_P(\widehat{P}'_X), \sigma_P x')\|^2_{\mathbb{R}^{2L+d}}}{2\sigma_P^2 \sigma_X^2}.
\end{aligned} \tag{32}$$

This is also a Gaussian kernel, now on $\mathbb{R}^{2L+d}$. Again by applying Theorem 24, we have

$$\bar{k}(\widehat{P}_X, X), (\widehat{P}'_X, X')) \approx \int_{\mathbb{R}^{2L+d}} p(v) e^{jv^T((\sigma_X Z_P(P_X), \sigma_P X) - (\sigma_X Z_P(P'_X), \sigma_P X'))} dv.$$

Let $v_1, v_2, ..., v_q$ be drawn i.i.d. from $p(v)$, the inverse Fourier transform of the Gaussian kernel with bandwidth $\sigma_P \sigma_X$. Let $u = (\sigma_X Z_P(\widehat{P}_X), \sigma_P x)$ and $u' = (\sigma_X Z_P(\widehat{P}'_X), \sigma_P x')$.

Then

$$\begin{aligned}\bar{k}(\tilde{x},\tilde{x}') &\approx \frac{1}{Q}\sum_{q=1}^{Q}\cos(v_q^T(u-u')) \\ &= \bar{z}(\tilde{x})^T\bar{z}(\tilde{x}'),\end{aligned}$$

where

$$\bar{z}(\tilde{x}) = \frac{1}{\sqrt{Q}}[\cos(v_1^T u),\sin(v_1^T u),...,\cos(v_Q^T u),\sin(v_Q^T u)] \in \mathbb{R}^{2Q} \tag{33}$$

and $\bar{z}(\tilde{x}')$ is defined similarly.

This completes the construction of the approximate feature map. The following result, which uses Hoeffding's inequality and generalizes a result of Rahimi and Recht (2007), says that the approximation achieves any desired approximation error with very high probability as $L, Q \to \infty$.

Theorem 25 *Let L be the number of random features to approximate the kernel on distributions and Q be the number of features to approximate the final product kernel. For any $\epsilon_l > 0$, $\epsilon_q > 0$, $\tilde{x} = (\widehat{P}_X, x)$, $\tilde{x}' = (\widehat{P}'_X, x')$,*

$$P(|\bar{k}(\tilde{x},\tilde{x}') - \bar{z}(\tilde{x})^T\bar{z}(\tilde{x}')| \geq \epsilon_l + \epsilon_q) \leq 2\exp\Big(-\frac{Q\epsilon_q^2}{2}\Big) + 6n_1n_2\exp\Big(-\frac{L\epsilon^2}{2}\Big), \tag{34}$$

where $\epsilon = \frac{\sigma_P^2}{2}\log(1+\epsilon_l)$, σ_P is the bandwidth parameter of the Gaussian-like kernel k_P, and n_1 and n_2 are the sizes of the empirical distributions $\widehat{P}_X$ and $\widehat{P}'_X$, respectively.

Proof Observe:

$$\bar{k}(\tilde{x},\tilde{x}') = \exp\left\{\frac{-1}{2\sigma_P^2}\|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\right\}\exp\left\{\frac{-1}{2\sigma_X^2}\|x-x'\|^2\right\},$$

and denote:

$$\tilde{k}(\tilde{x},\tilde{x}') = \exp\left\{\frac{-1}{2\sigma_P^2}\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2\right\}\exp\left\{\frac{-1}{2\sigma_X^2}\|x-x'\|^2\right\},$$

We omit the arguments of $\bar{k}, \tilde{k}$ for brevity. Let k_q be the final approximation ($k_q = \bar{z}(\tilde{x})^T\bar{z}(\tilde{x}')$) and then we have

$$|\bar{k} - k_q| = |\bar{k} - \tilde{k} + \tilde{k} - k_q| \leq |\bar{k} - \tilde{k}| + |\tilde{k} - k_q|. \tag{35}$$

From Eqn. (35) it follows that,

$$P(|\bar{k} - k_q| \geq \epsilon_l + \epsilon_q) \leq P(|\bar{k} - \tilde{k}| \geq \epsilon_l) + P(|\tilde{k} - k_q| \geq \epsilon_q). \tag{36}$$

By a direct application of Hoeffding's inequality,

$$P(|\tilde{k} - k_q| \geq \epsilon_q) \leq 2\exp(-\frac{Q\epsilon_q^2}{2}). \tag{37}$$

Recall that $\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X)\rangle = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} k'_X(X_i, X'_j)$. For a pair X_i, X'_j, we have again by Hoeffding

$$P(|z'_X(X_i)^T z'_X(X'_j) - k'_X(X_i, X'_j)| \geq \epsilon) \leq 2\exp(-\frac{L\epsilon^2}{2}).$$

Let Ω_{ij} be the event $|z'_X(X_i)^T z'_X(X'_j) - k'_X(X_i, X'_j)| \geq \epsilon$, for particular i, j. Using the union bound we have

$$P(\Omega_{11} \cup \Omega_{12} \cup \ldots \cup \Omega_{n_1 n_2}) \leq 2n_1 n_2 \exp(-\frac{L\epsilon^2}{2})$$

This implies

$$P(|Z_P(\widehat{P}_X)^T Z_P(\widehat{P}'_X) - \langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X)\rangle| \geq \epsilon) \leq 2n_1 n_2 \exp(-\frac{L\epsilon^2}{2}). \tag{38}$$

Therefore,

$$\begin{aligned}
\left|\bar{k} - \tilde{k}\right| &= \left| \exp\left\{\frac{-1}{2\sigma_X^2}\|x - x'\|^2\right\}\left[\exp\left\{\frac{-1}{2\sigma_P^2}\|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\right\}\right.\right. \\
&\quad \left.\left. - \exp\left\{\frac{-1}{2\sigma_P^2}\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2\right\}\right]\right| \\
&\leq \left|\left[\exp\left\{\frac{-1}{2\sigma_P^2}\|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\right\} - \exp\left\{\frac{-1}{2\sigma_P^2}\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2\right\}\right]\right| \\
&= \left|\exp\left\{\frac{-1}{2\sigma_P^2}\|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\right\}\left[1 - \exp\left\{\frac{-1}{2\sigma_P^2}\Big(\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2\right.\right.\right. \\
&\quad \left.\left.\left. - \|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\Big)\right\}\right]\right| \\
&\leq \left|\left[1 - \exp\left\{\frac{-1}{2\sigma_P^2}\Big(\|Z_P(\widehat{P}_X) - Z_P(\widehat{P}'_X)\|^2 - \|\Psi(\widehat{P}_X) - \Psi(\widehat{P}'_X)\|^2\Big)\right\}\right]\right| \\
&= \left|1 - \exp\left\{\frac{-1}{2\sigma_P^2}\Big(Z_P(\widehat{P}_X)^T Z_P(\widehat{P}_X) - \langle \Psi(\widehat{P}_X), \Psi(\widehat{P}_X)\rangle + Z_P(\widehat{P}'_X)^T Z_P(\widehat{P}'_X)\right.\right. \\
&\quad \left.\left. - \langle \Psi(\widehat{P}'_X), \Psi(\widehat{P}'_X)\rangle - 2\big(Z_P(\widehat{P}_X)^T Z_P(\widehat{P}'_X) - \langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X)\rangle\big)\Big)\right\}\right| \\
&\leq \left|1 - \exp\left\{\frac{1}{2\sigma_P^2}\Big(|Z_P(\widehat{P}_X)^T Z_P(\widehat{P}_X) - \langle \Psi(\widehat{P}_X), \Psi(\widehat{P}_X)\rangle| + |Z_P(\widehat{P}'_X)^T Z_P(\widehat{P}'_X)\right.\right. \\
&\quad \left.\left. - \langle \Psi(\widehat{P}'_X), \Psi(\widehat{P}'_X)\rangle| + 2|\big(Z_P(\widehat{P}_X)^T Z_P(\widehat{P}'_X) - \langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X)\rangle\big)|\Big)\right\}\right|
\end{aligned}$$

The result now follows by applying the bound of Eqn. (38) to each of the three terms in the exponent of the preceding expression, together with the stated formula for ϵ in terms of ϵ_ℓ.

■

The above results holds for fixed $\tilde{x}$ and $\tilde{x}'$. Following again Rahimi and Recht (2007), one can use an ϵ-net argument to prove a stronger statement for every pair of points in the input space simultaneously. They show

Lemma 26 *Let $\mathcal{M}$ be a compact subset of $\mathbb{R}^d$ with diameter $r = \mathrm{diam}(\mathcal{M})$ and let D be the number of random Fourier features used. Then for the mapping defined in* (29)*, we have*

$$P\Big(\sup_{x,y\in\mathcal{M}} |z_w(x)^T z_w(y) - k(x-y)| \geq \epsilon\Big) \leq 2^8\Big(\frac{\sigma r}{\epsilon}\Big)^2 \exp\Big(\frac{-D\epsilon^2}{2(d+2)}\Big),$$

where $\sigma = \mathbb{E}[w^T w]$ is the second moment of the Fourier transform of k.

Our RFF approximation of $\bar{k}$ is grounded on Gaussian RFF approximations on Euclidean spaces, and thus, the following result holds by invoking Lemma 26, and otherwise following the argument of Theorem 25.

Theorem 27 *Using the same notations as in Theorem 25 and Lemma 26,*

$$P\Big(\sup_{x,x'\in\mathcal{M}} |\bar{k}(\tilde{x},\tilde{x}') - \bar{z}(\tilde{x})^T\bar{z}(\tilde{x}')| \geq \epsilon_l + \epsilon_q\Big)$$
$$\leq 2^8\Big(\frac{\sigma'_X r}{\epsilon_q}\Big)^2 \exp\Big(\frac{-Q\epsilon_q^2}{2(d+2)}\Big) + 2^9 3 n_1 n_2\Big(\frac{\sigma_P\sigma_X r}{\epsilon_l}\Big)^2 \exp\Big(\frac{-L\epsilon_l^2}{2(d+2)}\Big) \quad (39)$$

where σ'_X is the width of kernel k'_X in Eqn. (30) and σ_P and σ_X are the widths of kernels k_P and k_X respectively.

Proof The proof is very similar to the proof of Theorem 25. We use Lemma 26 to replace bound (37) with:

$$P\Big(\sup_{x,x'\in\mathcal{M}} |\tilde{k} - k_q| \geq \epsilon_q\Big) \leq 2^8\Big(\frac{\sigma'_X r}{\epsilon_q}\Big)^2 \exp\Big(\frac{-Q\epsilon_q^2}{2(d+2)}\Big). \quad (40)$$

Similarly, Eqn. (38) is replaced by

$$P\Big(\sup_{x,x'\in\mathcal{M}} \big|Z_P(\widehat{P}_X)^T Z_P(\widehat{P}'_X) - \big\langle\Psi(\widehat{P}_X), \Psi(\widehat{P}'_X)\big\rangle\big| \geq \epsilon\Big)$$
$$\leq 2^9 n_1 n_2\Big(\frac{\sigma_P\sigma_X r}{\epsilon_l}\Big)^2 \exp\Big(\frac{-L\epsilon_l^2}{2(d+2)}\Big). \quad (41)$$

The remainder of the proof now proceeds as in the previous proof.

■

There are recent developments that give faster rates for approximation quality of random Fourier features and could potentially be combined with our analysis (Sriperumbudur and

Szabó, 2015; Sutherland and Schneider, 2015). For example, approximation quality for the kernel mean map is discussed in Sutherland and Schneider (2015), and these ideas could be extended to Theorem 27 by combining with the two-stage approach presented in this paper. We also note that our analysis of random Fourier features is separate from our analysis of the kernel learning algorithm. We have not presented a generalization error bound for the learning algorithm using random Fourier features (Rudi and Rosasco, 2017).

A.6.2 Nyström Approximation

Like random Fourier features, the Nyström approximation is a technique to approximate kernel matrices (Williams and Seeger, 2001; Drineas and Mahoney, 2005). Unlike random Fourier features, for the Nyström approximation, the feature maps are data-dependent. Also, in the last subsection, all kernels were assumed to be shift invariant. With the Nyström approximation there is no such assumption.

For a general kernel k, the goal is to find a feature mapping $z : \mathbb{R}^d \to \mathbb{R}^L$, where $L > d$, such that $k(x, x') \approx z(x)^T z(x')$. Let r be the target rank of the final approximated kernel matrix, and m be the number of selected columns of the original kernel matrix. In general $r \leq m \ll n$.

Given data points $x_1, \ldots, x_n$, the Nyström method approximates the kernel matrix by first sampling m data points $x'_1, x'_2, ..., x'_m$ without replacement from the original sample, and then constructing a low rank matrix by $\widehat{K}_r = K_b \widehat{K}^{-1} K_b^T$, where $K_b = [k(x_i, x'_j)]_{n \times m}$, and $\widehat{K} = [k(x'_i, x'_j)]_{m \times m}$. Hence, the final approximate feature mapping is

$$z_n(x) = \widehat{D}^{-\frac{1}{2}} \widehat{V}^T [k(x, x'_1), ..., k(x, x'_m)], \tag{42}$$

where $\widehat{D}$ is the eigenvalue matrix of $\widehat{K}$ and $\widehat{V}$ is the corresponding eigenvector matrix.

A.7 Results in Tabular Format

Examples per Task \ Tasks	16	64	256
8	36.01	33.08	31.69
16	31.55	31.03	30.96
32	30.44	29.31	23.87
256	23.78	7.22	1.27

Table 1: Average Classification Error of Marginal Transfer Learning on Synthetic Data set

Examples per Task \ Tasks	16	64	256
8	49.14	49.11	50.04
16	49.89	50.04	49.68
32	50.32	50.21	49.61
256	50.01	50.43	49.93

Table 2: Average Classification Error of Pooling on Synthetic Data set

Examples per Task \ Tasks	10	15	20	25	30	35
20	13.78	12.37	11.93	10.74	10.08	11.17
24	14.18	11.89	11.51	10.90	10.55	10.18
28	14.95	13.29	12.00	10.21	10.59	9.52
34	13.27	11.66	11.79	9.16	9.34	10.50
41	12.89	11.27	11.17	9.91	9.10	10.05
49	13.15	11.70	13.81	10.12	9.01	8.69
58	12.16	9.59	9.85	9.28	8.44	7.62
70	13.03	9.16	8.80	9.03	8.16	7.88
84	11.98	9.18	9.74	9.03	7.30	7.01
100	12.69	8.48	9.52	8.01	7.14	7.5

Table 3: RMSE of Marginal Transfer Learning on Parkinson's Disease Data set

Examples per Task \ Tasks	10	15	20	25	30	35
20	13.64	11.93	11.95	11.06	11.91	12.08
24	13.80	11.83	11.70	11.98	11.68	11.48
28	13.78	11.70	11.72	11.18	11.58	11.73
34	13.71	12.20	12.04	11.17	11.67	11.92
41	13.69	11.73	12.08	11.28	11.55	12.59
49	13.75	11.85	11.79	11.17	11.34	11.82
58	13.70	11.89	12.06	11.06	11.82	11.65
70	13.54	11.86	12.14	11.21	11.40	11.96
84	13.55	11.98	12.03	11.25	11.54	12.22
100	13.53	11.85	11.92	11.12	11.96	11.84

Table 4: RMSE of Pooling on Parkinson's Disease Data set

Examples per Task \ Tasks	10	20	30	40
5	8.62	7.61	8.25	7.17
15	6.21	5.90	5.85	5.43
30	6.61	5.33	5.37	5.35
45	5.61	5.19	4.71	4.70
all training data	5.36	4.91	3.86	4.08

Table 5: Average Classification Error of Marginal Transfer Learning on Satellite Data set

Examples per Task \ Tasks	10	20	30	40
5	8.13	7.54	7.94	6.96
15	6.55	5.81	5.79	5.57
30	6.06	5.36	5.56	5.31
45	5.58	5.12	5.30	4.99
all training data	5.37	4.98	5.32	5.14

Table 6: Average Classification Error of Pooling on Satellite Data set

Examples per Task \ Tasks	5	10	15	20
1024	9	9.03	9.03	8.70
2048	9.12	9.56	9.07	8.62
4096	8.96	8.91	9.01	8.66
8192	9.18	9.20	9.04	8.74
16384	9.05	9.08	9.04	8.63

Table 7: Average Classification Error of Marginal Transfer Learning on Flow Cytometry Data set

Examples per Task \ Tasks	5	10	15	20
1024	9.41	9.48	9.32	9.52
2048	9.92	9.57	9.45	9.54
4096	9.72	9.56	9.36	9.40
8192	9.43	9.53	9.38	9.50
16384	9.42	9.56	9.40	9.33

Table 8: Average Classification Error of Pooling on Flow Cytometry Data set

References

Nima Aghaeepour, Greg Finak, The FlowCAP Consortium, The DREAM Consortium, Holger Hoos, Tim R. Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H. Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238, 2013.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2019.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl0r3R9KX`.

Gökhan Bakır, Thomas Hofmann, Bernhard Schölkopf, Alexander J Smola, and Ben Taskar. *Predicting Structured Data*. MIT Press, 2007.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: Towards domain generalization using meta-regularization. In S. Bengio, H. Wallach, H. Larochelle,

K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 998–1008. Curran Associates, Inc., 2018.

Peter Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 139–153, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2178–2186. 2011.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10:2780–2824, 2016.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 129–136. 2008.

Timothy I. Cannings, Yingying Fan, and Richard J. Samworth. Classification with imperfect training labels. Technical Report arXiv:1805.11505, 2018.

Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2224–2233, 2019.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 406–414, 2010.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Algorithmic Learning Theory*, pages 38–53, 2008.

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*, KDD '15, pages 169–178, 2015.

Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes, volume I: Elementary Theory and Methods*. Springer, 2003.

Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes, volume II: General Theory and Structure*. Springer, 2008.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10169–10179. 2018a.

Giulia Denevi, Carlo Ciliberto, Dimitros Stamos, and Massimiliano Pontil. Incremental learning-to-learn with statistical guarantees. In *Proc. Uncertainty in Artificial Intelligence*, 2018b.

Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27:304–313, 2018.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6450–6461. 2019.

Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In J. Langford and J. Pineau, editors, *Proc. 29th Int. Conf. on Machine Learning*, pages 823–830, 2012.

Theodoros Evgeniou, Charles A. Michelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, 2017.

Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *Internation Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 859–868, 2016.

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *IEEE International Conference on Computer Vision*, page 2551–2559, 2015.

Muhammad Ghifary, David Balduzzi, W. Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1411–1430, 2017.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848, 2016.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel approach to comparing distributions. In R. Holte and A. Howe, editors, *22nd AAAI Conference on Artificial Intelligence*, pages 1637–1641, 2007a.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520, 2007b.

Thomas Grubinger, Adriana Birlutiu, Holger Schöner, Thomas Natschläger, and Tom Heskes. Domain generalization based on transfer component analysis. In I. Rojas, G. Joya, and A. Catala, editors, *Advances in Computational Intelligence, International Work-Conference on Artificial Neural Networks*, volume 9094 of *Lecture Notes in Computer Science*, pages 325–334. Springer International Publishing, 2015.

Peter Hall. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 43(2):147–156, 1981.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *International Conference on Machine Learning*, pages 408–415. ACM, 2008.

Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In Amir Globerson and Ricardo Silva, editors, *Uncertainty in Artificial Intelligence*, 2019.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Scholkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608, 2007.

Wittawat Jitkrittum, Arthur Gretton, Nicolas Heess, SM Eslami, Balaji Lakshminarayanan, Dino Sejdinovic, and Zoltán Szabó. Kernel-based just-in-time learning for passing expectation propagation messages. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 405–414. AUAI Press, 2015.

Thorsten Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *12th European Conference on Computer Vision - Volume Part I*, page 158–171, 2012.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145, 1995.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902 – 1914, 2001.

Patrice Latinne, Marco Saerens, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. In C. Sammut and A. H. Hoffmann, editors, *International Conference on Machine Learning*, pages 298–305, 2001.

Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: approximating kernel expansions in loglinear time. In *International Conference on International Conference on Machine Learning-Volume 28*, pages III–244, 2013.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018b.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI Conference on Artificial Intelligence*, 2018c.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018d.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, pages 1041–1048, 2009b.

Andreas Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75(3): 327–350, 2009.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In Sanjoy Dasgupta and David McAllester, editors, *International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 343–351, 2013.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107:1561–1595, 2018.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision*, pages 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages I–10–I–18, 2013.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018. URL `http://jmlr.org/papers/v18/15-226.html`.

Kalyanapuram Rangachari Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In Eric P. Xing and Tony Jebara, editors, *International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 991–999, 2014.

Iosif F. Pinelis and Aleksandr Ivanovich Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

Tyler Sanderson and Clayton Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Conferencen on Artificial Intelligence and Statistics*, 2014.

Clayton Scott. A generalized Neyman-Pearson criterion for optimal domain adaptation. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 738–761, 2019.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Dx7fbCW`.

Srinagesh Sharma and James W. Cutler. Robust orbit determination and classification: A learning theoretic approach. *Interplanetary Network Progress Report*, 203:1, 2015.

Bharath Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152, 2015.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Amos J Storkey. When training and test sets are different: characterising learning transfer. In *Dataset Shift in Machine Learning*, pages 3–28. MIT Press, 2009.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Danica J. Sutherland and Jeff Schneider. On the error of random Fourier features. In *Uncertainty in Artificial Intelligence*, pages 862–871, 2015.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1): 5272–5311, 2016.

Dirk Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18:1–32, 2017.

Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in Neural Information Processing Systems*, pages 640–646, 1996.

D. Michael Titterington. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 45(1):37–46, 1983.

Joern Toedling, Peter Rhein, Richard Ratei, Leonid Karawajew, and Rainer Spang. Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. *BMC Bioinformatics*, 7:282, 2006.

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484, 2005.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJEjjoR9K7`.

Jenna Wiens. *Machine Learning for Patient-Adaptive Ectopic Beat Classication*. Masters Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2010.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

Liu Yang, Steve Hanneke, and Jamie Carbonell. A theory of transfer learning with applications to active learning. *Machine Learning*, 90(2):161–189, 2013.

Xiaolin Yang, Seyoung Kim, and Eric P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing Systems*, pages 2151–2159, 2009.

Yao-Liang Yu and Csaba Szepesvari. Analysis of kernel mean matching under covariate shift. In *International Conference on Machine Learning*, pages 607–614, 2012.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *International Conference on Machine Learning*, 2004.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

Kun Zhang, Mingming Gong, and Bernhard Scholkopf. Multi-source domain adaptation: A causal view. In *AAAI Conference on Artificial Intelligence*, pages 3150–3157. AAAI Press, 2015.