Kernel Mean Embedding of Distributions: A Review and Beyonds

Krikamol Muandet
Mahidol University and MPI for Intelligent Systems
272 Rama VI Road, Ratchathewi District, Bangkok 10400, Thailand
Spemannstraße 38, Tübingen 72076, Germany
krikamol.mua@mahidol.ac.th

Kenji Fukumizu
Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa, Tokyo 190-8562 Japan
fukumizu@ism.ac.jp

Bharath Sriperumbudur
Department of Statistics, Pennsylvania State University
University Park, PA 16802, USA
bks18@psu.edu

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Spemannstraße 38, Tübingen 72076, Germany
bs@tuebingen.mpg.de

Contents

1 Introduction **2**
- 1.1 Purpose and Scope 7
- 1.2 Outline of the Survey 8

2 Background **10**
- 2.1 Learning with Kernels 10
- 2.2 Reproducing Kernel Hilbert Spaces 17
- 2.3 Hilbert-Schmidt Operators 20

3 Hilbert Space Embedding of Marginal Distributions **22**
- 3.1 From Data Point to Probability Measure 22
- 3.2 Covariance Operators 30
- 3.3 Properties of the Mean Embedding 33
- 3.4 Kernel Mean Estimation and Approximation 39
- 3.5 Maximum Mean Discrepancy 44
- 3.6 Kernel Dependency Measures 51
- 3.7 Learning on Distributional Data 54
- 3.8 Recovering Information from Mean Embeddings 64

4 Hilbert Space Embedding of Conditional Distributions **70**
- 4.1 From Marginal to Conditional Distribution 71
- 4.2 Regression Interpretation 76

4.3 Basic Operations: Sum, Product, and Bayes' Rules 78
4.4 Graphical Models and Probabilistic Inference 84
4.5 Markov Decision Processes and Reinforcement Learning . . 89
4.6 Conditional Dependency Measures 92
4.7 Causal Discovery 94

5 Relationships between KME and Other Methods **97**

6 Future Directions **103**

7 Conclusions **107**

References **109**

Abstract

A Hilbert space embedding of distributions—in short, kernel mean embedding—has recently emerged as a powerful machinery for probabilistic modeling, statistical inference, machine learning, and causal discovery. The basic idea behind this framework is to map distributions into a reproducing kernel Hilbert space (RKHS) in which the whole arsenal of kernel methods can be extended to probability measures. It gave rise to a great deal of research and novel applications of positive definite kernels. The goal of this survey is to give a comprehensive review of existing works and recent advances in this research area, and to discuss some of the most challenging issues and open problems that could potentially lead to new research directions. The survey begins with a brief introduction to the RKHS and positive definite kernels which forms the backbone of this survey, followed by a thorough discussion of the Hilbert space embedding of marginal distributions, theoretical guarantees, and review of its applications. The embedding of distributions enables us to apply RKHS methods to probability measures which prompts a wide range of applications such as kernel two-sample testing, independent testing, group anomaly detection, and learning on distributional data. Next, we discuss the Hilbert space embedding for conditional distributions, give theoretical insights, and review some applications. The conditional mean embedding enables us to perform sum, product, and Bayes' rules—which are ubiquitous in graphical model, probabilistic inference, and reinforcement learning—in a non-parametric way using the new representation of distributions in RKHS. We then discuss relationships between this framework and other related areas. Lastly, we give some suggestions on future research directions. The targeted audience includes graduate students and researchers in machine learning and statistics who are interested in the theory and applications of kernel mean embedding.

1

Introduction

Kernel mean embedding of distributions has recently emerged as a powerful tool for statistical inference and machine learning. This work aims to give a comprehensive review of the most recent advances in this area and, in the course of doing so, to discuss some challenging issues that could potentially lead to new research directions. To the best of our knowledge, there is no comparable review in this area so far; however, the review paper of Song et al. (2013) on Hilbert space embedding of conditional distributions and its applications in nonparametric inference in graphical models may also be of interest to some readers.

As the name suggests, the kernel mean embedding owes its success to the concept of a positive definite function commonly known as *kernel*. The kernel function has been popular in the machine learning community for more than 20 years. Initially, it arises as an effortless way to perform an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ in a high-dimensional feature space $\mathcal{F}$. The positive definiteness of the kernel function guarantees the existence of a dot product space $\mathcal{F}$ and a mapping $\phi : \mathcal{X} \to \mathcal{F}$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}}$ (Aronszajn 1950). But, ϕ need not be computed explicitly (Boser et al. 1992, Cortes and Vapnik 1995, Vapnik 2000, Schölkopf and Smola 2001) and it can be applied to any learn-

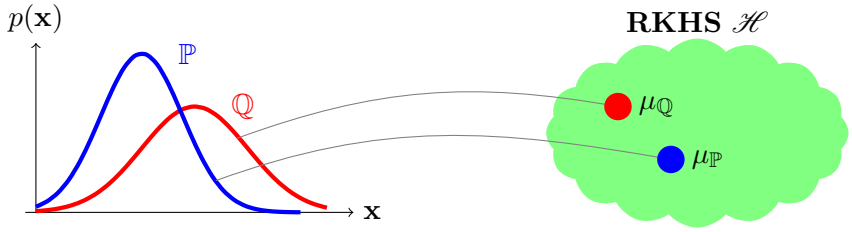


Figure 1.1: Embedding of marginal distributions: Each distribution is mapped into an RKHS via an expectation operation.

ing algorithms as long as they can be expressed entirely in terms of a dot product $\langle \mathbf{x}, \mathbf{y} \rangle$. This trick is commonly known as the *kernel trick* (see Section 2 for a more detailed account). Many kernel functions have been proposed for various kinds of data structures including non-vectorial data such as graph, text documents, semi-groups, and probability distributions (Schölkopf and Smola 2001, Gärtner 2003). Many well-known learning algorithms have already been *kernelized* and have proven successful in many scientific disciplines such as bioinformatics, natural language processing, computer vision, robotics, and causal inference, among others.

Figure 1.1 and 1.2 depict schematic illustrations of the kernel mean embedding framework. In words, the idea of *kernel mean embedding* is to extend the feature map ϕ to the space of probability distributions by representing each distribution $\mathbb{P}$ as a mean function

$$\phi(\mathbb{P}) = \mu_{\mathbb{P}} := \int_{\mathcal{X}} k(x, \cdot) \, \mathrm{d}\mathbb{P}(x), \tag{1.1}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive definite kernel function (Berlinet and Thomas-Agnan 2004, Smola et al. 2007). The integral in (1.1) should be interpreted as a Bochner integral (see, *e.g.*, Diestel and Uhl (1977; Chapter 2) and Dinculeanu (2000; Chapter 1) for a definition of the Bochner integral). Conditions ensuring the existence of such an integral will be discussed later in Section 3, but in this case we essentially transform the distribution $\mathbb{P}$ to an element in the feature space $\mathcal{F}$, which is nothing but a reproducing kernel Hilbert space (RKHS) endowed with the kernel k. Through (1.1), most RKHS methods can therefore be extended to probability measures. There are

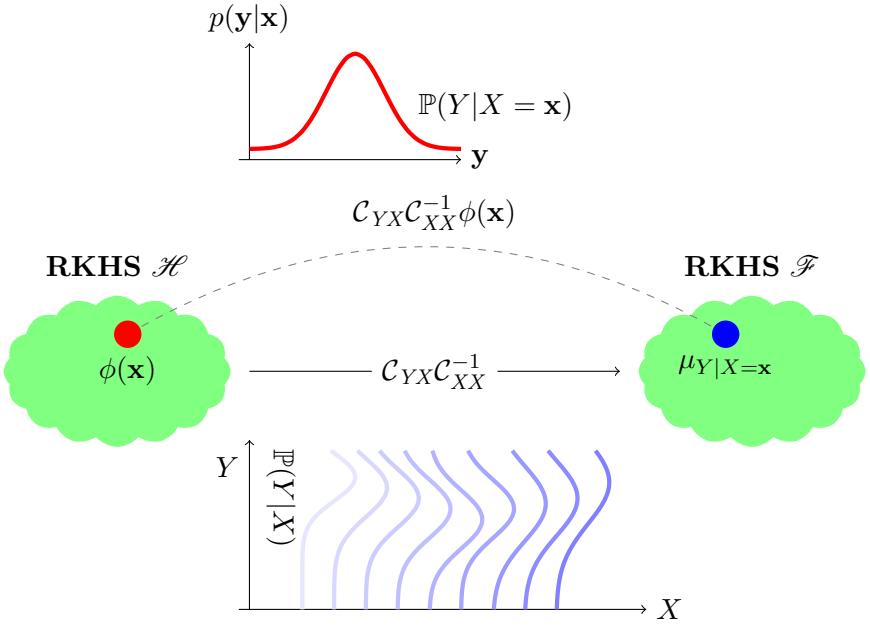


Figure 1.2: From marginal distribution to conditional distribution: Unlike the embeddings shown in Figure 1.1, the embedding of conditional distribution $\mathbb{P}(Y|X)$ is not a single element in the RKHS. Instead, it may be viewed as a family of Hilbert space embeddings of conditional distributions $\mathbb{P}(Y|X=\mathbf{x})$ indexed by the conditioning variable X. In other words, the conditional mean embedding can be viewed as an operator mapping from $\mathscr{H}$ to $\mathscr{F}$ (cf. §4.2).

several reasons why this representation may be beneficial.

First of all, for a class of kernel functions known as *characteristic kernels*, the kernel mean representation captures all necessary information about the distribution $\mathbb{P}$ (Fukumizu et al. 2004, Sriperumbudur et al. 2008; 2010). In other words, the mean map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective which means that $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{F}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. Consequently, the kernel mean representation can be used to define a metric over the space of probability distributions (Sriperumbudur et al. 2010). Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ makes it suitable for applications that require a unique characterization of distributions such as two-sample homogeneity tests (Gretton et al. 2012a, Fukumizu et al. 2008, Zhang et al. 2011, Doran et al. 2014). Moreover, using the kernel mean

representation, most learning algorithms can be extended to the space of probability distributions with minimal assumptions on the underlying data generating process (Gómez-Chova et al. 2010, Muandet et al. 2012, Guevara et al. 2014, Lopez-Paz et al. 2015). See §3.3 for details.

Secondly, several elementary operations on distributions (and associated random variables) can be performed directly by means of this representation. For example, by the reproducing property of $\mathcal{F}$, we have

$$\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{F}} \tag{1.2}$$

for all $f \in \mathcal{F}$. Likewise, $\mathbb{E}_{Y|\mathbf{x}}[g(Y) \,|\, X = \mathbf{x}] = \langle g, \mathcal{U}_{Y|\mathbf{x}} \rangle_{\mathcal{F}_Y}$ for all $g \in \mathcal{F}_Y$ where $\mathcal{U}_{Y|\mathbf{x}}$ denotes the embedding of the conditional distribution $\mathbb{P}(Y|X = \mathbf{x})$ (see Section 4 for further detail). The kernel mean representation allows us to implement these operations in non-parametric probabilistic inference, *e.g.*, filtering for dynamical systems (Song et al. 2009), kernel belief propagation (Song et al. 2011a), kernel Monte Carlo filter (Kanagawa et al. 2013), kernel Bayes' rule (Fukumizu et al. 2011), often with strong theoretical guarantees. Moreover, it can be used to perform functional operations $f(X, Y)$ on independent random variables X and Y (Schölkopf et al. 2015).

In some applications such as testing for homogeneity from finite samples, representing the distribution $\mathbb{P}$ by $\mu_{\mathbb{P}}$ bypasses an intermediate density estimation, which is known to be difficult in the high-dimensional setting (Wasserman 2006; Section 6.5). Moreover, we can extend the applications of kernel mean embedding straightforwardly to non-vectorial data such as graphs, strings, and semi-groups, thanks to the kernel function. As a result, statistical inference—such as two-sample testing and independence testing—can be adapted directly to distributions over complex objects (Gretton et al. 2012a).

Under additional assumptions, we can generalize the principle underlying (1.1) to conditional distributions $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|X = \mathbf{x})$. Essentially, the latter two objects are represented as an operator mapping the feature space $\mathcal{F}_X$ to $\mathcal{F}_Y$, and as an object in the feature space $\mathcal{F}_Y$, respectively, where $\mathcal{F}_X$ and $\mathcal{F}_Y$ denote the feature space for X and Y (see Figure 1.2). These representations allow us to develop a powerful language for algebraic manipulation of probability distributions in an analogous way to the sum rule, product rule, and Bayes' rule—

which are ubiquitous in graphical models and probabilistic inference—without making assumption on parametric forms of the underlying distributions. The details of conditional mean embeddings will be given in Section 4.

Table 1.1 provides an overview comparison between kernel mean embedding and classical methods such as the divergence method, kernel density estimation, probabilistic models, and empirical characteristic functions. Notably, the KME and the ECF exhiibit similar properties. See Example 3.3 and Section 5 for a connection between these two approaches.

Table 1.1: The comparisons between kernel mean embedding and classical methods. The columns marked 'DM', 'KDE', 'PM', 'ECF', and 'KME' correspond to divergence method, kernel density estimation, probabilistic model, empirical characteristic function approach, and kernel mean embedding.

Characteristic	DM	KDE	PM	ECF	KME
❶ parametric assumption	☹	☹	☹	☺	☺
❷ curse of dimensionality	☹	☹	☹	☺	☺
❸ fast convergence	☹	☹	☹	☺	☺
❹ interpretability	☺	☹	☺	☹	☹

A Synopsis. As a result of the aforementioned advantages, the kernel mean embedding has made widespread contributions in various directions. Firstly, most tasks in machine learning and statistics involve estimation of the data-generating process whose success depends critically on the accuracy and the reliability of this estimation. It is known that estimating the kernel mean embedding is easier than estimating the distribution itself, which help improve many statistical inference methods. These include, for example, two-sample testing (Gretton et al. 2012a), independence and conditional independence tests (Fukumizu et al. 2008, Zhang et al. 2011, Doran et al. 2014), causal inference (Sgouritsa et al. 2013, Chen et al. 2014), adaptive MCMC (Sejdinovic et al. 2014), and approximate Bayesian computation (Fukumizu et al. 2013).

Secondly, several attempts have been made in using kernel mean embedding as a representation in the predictive learning on distributions (Muandet et al. 2012, Szabó et al. 2015, Muandet and Schölkopf 2013, Guevara et al. 2014, Lopez-Paz et al. 2015). As opposed to the classical setting where training and test examples are data points, many applications call for a learning framework in which training and test examples are probability distributions. This is ubiquitous in, for example, multiple-instance learning (Doran 2013), learning with noisy and uncertain input, learning from missing data, group anomaly detection (Muandet and Schölkopf 2013, Guevara et al. 2014), dataset squishing, bag-of-words data (Yoshikawa et al. 2014; 2015), etc. The kernel mean representation equipped with the RKHS methods enables classification, regression, and anomaly detection to be performed on distributions.

Finally, the kernel mean embedding also allows one to perform complex approximate inference without making strong parametric assumption on the form of underlying distribution. The idea is to represent all relevant probabilistic quantities as a kernel mean embedding. Then, basic operations such as *sum rule* and *product rule* can be formulated in terms of expectation and inner product in feature space. Examples of algorithms in this class include kernel belief propagation (KBP), kernel embedding of latent tree model, kernel Bayes rule, predictive-state representation, etc (Song et al. 2010b; 2009; 2011a; 2013, Fukumizu et al. 2013). Recently, the kernel mean representation has become one of the prominent tools in causal inference and discovery (Lopez-Paz et al. 2015, Sgouritsa et al. 2013, Chen et al. 2014, Schölkopf et al. 2015).

The aforementioned examples represent only a handful of successful applications of kernel mean embedding. More examples and details will be provided throughout the survey.

1.1 Purpose and Scope

The purpose of this survey is to give a comprehensive review of kernel mean embedding of distributions, to present important theoretical results and practical applications, and to draw connections to related areas. We restrict the scope of this survey to key theoretical results and

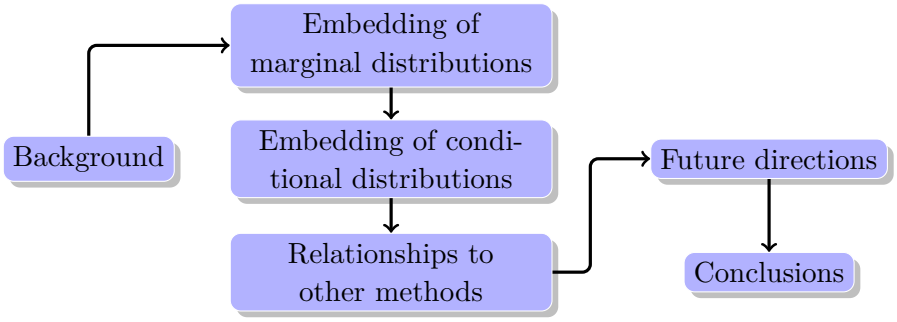


Figure 1.3: Schematic outline of this survey.

new applications of kernel mean embedding with references to other related works. We focus primarily on basic intuitions and sketches for proofs, leaving the full proofs to the papers cited.

All materials presented in this paper should be accessible to a wide range of audiences. In particular, we hope that this survey will be most useful to readers who are not at all familiar with the idea of kernel mean embedding, but already have some background knowledge in machine learning. To ease the reading, we suggest non-expert readers to also consult basic machine learning textbooks such as Bishop (2006), Schölkopf and Smola (2001), Mohri et al. (2012), Murphy (2012). Experienced machine learners who are interested in applying the idea of kernel mean embedding to their works are also encouraged to read this survey. Lastly, we will also provide some practical considerations that could be useful to practitioners who are interested in implementing the idea in real-world applications.

1.2 Outline of the Survey

The schematic outline of this survey is depicted in Figure 1.3 and can be summarized as follows.

Section 2 introduces notations and basic idea of a positive definite kernel and reproducing kernel Hilbert space (RKHS) (§2.1 and §2.2). It also presents general theoretical results such as reproducing property (Prop 2.1), Riesz representation theorem (Thm 2.4), Mercer's theorem

(Thm 2.1), Bochner's theorem (Thm 2.2), and Schoenberg's characterization (Thm 2.3). In addition, it contains a brief discussion about Hilbert-Schmidt operators on RKHS (§2.3).

Section 3 conveys the idea of Hilbert space embedding of marginal distributions (§3.1) as well as covariance operators (§3.2), presents essential properties of mean embedding (§3.3), discusses its estimation and approximation procedures (§3.4), and reviews important applications, notably maximum mean discrepancy (MMD) (§3.5), kernel dependence measure (§3.6), learning on distributional data (§3.7), and how to recover information from the embedding of distributions (§3.8).

Section 4 generalizes the idea of kernel mean embedding to the space of conditional distributions, called *conditional mean embedding* (§4.1), presents regression perspective (§4.2), and describes basic operations—namely sum rule, product rule, and Bayes' rule—in terms of marginal and conditional mean embeddings (§4.3). We review applications in graphical models, probabilistic inference (§4.4), reinforcement learning (§4.5), conditional dependence measures (§4.6), and causal discovery (§4.7). Estimating the conditional mean embedding is challenging both theoretically and empirically. We discuss some of the key challenges as well as some applications.

Section 5 draws connections between kernel mean embedding framework and other methods including kernel density estimation, empirical characteristic function, divergence methods and probabilistic modeling.

Section 6 provides suggestions for future research.

2

Background

This section introduces the kernel methods and the concept of reproducing kernel Hilbert space (RKHS) which form the backbone of the survey. Readers who are familiar with probability theory and the concept of RKHS may skip this section entirely and return to it later. More detailed account on this topic can be found, in Schölkopf and Smola (2001), Berlinet and Thomas-Agnan (2004), and Hofmann et al. (2008), for example. Readers who are interested particularly in the typical applications of kernels in machine learning, *e.g.*, support vector machines (SVMs) and Gaussian processes (GPs), are encouraged to read Cucker and Smale (2002), Burges (1998), Rasmussen and Williams (2005) and references therein.

2.1 Learning with Kernels

Many classical learning algorithms—such as the perceptron (Rosenblatt 1958), support vector machine (SVM) (Cortes and Vapnik 1995), and principle component analysis (PCA) (Pearson 1901, Hotelling 1933)—employ data instances only through an inner product $\langle \mathbf{x}, \mathbf{x}' \rangle$, which basically is a similarity measure between $\mathbf{x}$ and $\mathbf{x}'$.

However, a class of linear functions induced by this inner product may be too restrictive for many real-world problems. Kernel methods aim to build more flexible and powerful learning algorithms by replacing $\langle \mathbf{x}, \mathbf{x}' \rangle$ with some other, possibly non-linear, similarity measures.

The most natural extension of $\langle \mathbf{x}, \mathbf{x}' \rangle$ is to explicitly apply a non-linear transformation:

$$\begin{array}{rcl} \phi : \mathcal{X} & \longrightarrow & \mathcal{F} \\ \mathbf{x} & \longmapsto & \phi(\mathbf{x}) \end{array} \tag{2.1}$$

into a high-dimensional *feature space* $\mathcal{F}$ and subsequently evaluate the inner product in $\mathcal{F}$, *i.e.*,

$$k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}, \tag{2.2}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denotes the inner product of $\mathcal{F}$. We will refer to ϕ and k as a *feature map* and a *kernel function*, respectively. Likewise, we can interpret $k(\mathbf{x}, \mathbf{x}')$ as a non-linear similarity measure between $\mathbf{x}$ and $\mathbf{x}'$. Since most algorithms depend on the data set only through the inner product $\langle \mathbf{x}, \mathbf{x}' \rangle$, we can obtain a non-linear extension of these algorithms by simply substituting $\langle \mathbf{x}, \mathbf{x}' \rangle$ with $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$. Note that the learning algorithm remains the same: we only change the space in which these algorithms operate. As (2.1) is non-linear, a linear algorithm in $\mathcal{F}$ corresponds to the non-linear counterpart in the original space $\mathcal{X}$.

For example, let consider a polynomial feature map $\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ when $\mathbf{x} \in \mathbb{R}^2$. Then, we have

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} = x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1x_2x_1'x_2' = \langle \mathbf{x}, \mathbf{x}' \rangle^2. \tag{2.3}$$

Put differently, the new similarity measure is simply the square of the inner product in $\mathcal{X}$. The equality (2.3) holds more generally for a d-degree polynomial, *i.e.*, ϕ maps $\mathbf{x} \in \mathbb{R}^N$ to the vector $\phi(\mathbf{x})$ whose entries are all possible dth degree ordered products of the entries of $\mathbf{x}$. In that case, we have $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} = \langle \mathbf{x}, \mathbf{x}' \rangle^d$. Thus, the complexity of the learning algorithm is controlled by the complexity of ϕ and by increasing the degree d, one would expect the resulting algorithm to become more complex. Additional examples of how to

construct an explicit feature map can be found in Schölkopf and Smola (2001; Chapter 2).

Unfortunately, evaluating $k(\mathbf{x}, \mathbf{x}')$ as above requires a two-step procedure: i) one construct the feature maps $\phi(\mathbf{x})$ and $\phi(\mathbf{x}')$ explicitly, and ii) then evaluate $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$. These two steps can be computationally expensive if $\phi(\mathbf{x})$ lives in a high-dimensional feature space, *e.g.*, when the degree d of the polynomial is large. Fortunately, (2.3) implies that there is an alternative way to evaluate $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$ without resorting to constructing $\phi(\mathbf{x})$ explicitly if all we need is an inner product $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$. That is, it is sufficient to consider $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^2$ directly. This is an essential aspect of kernel methods, often referred to as the *kernel trick* in the machine learning community.

It turns out that—if k is *positive definite* (cf. Definition 2.1)—there always exists some $\phi : \mathcal{X} \to \mathcal{F}$ for which $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$. Since $\langle \cdot, \cdot \rangle$ is positive definite, it follows that k defined as in (2.2) is positive definite for any choice of explicit feature map ϕ.

Definition 2.1 (positive definite kernel)**.** A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetric, *i.e.*, $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})$, and the Gram matrix is positive definite:

$$\sum_{i,j=1}^{n} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{2.4}$$

for any $n \in \mathbb{N}$, any choice of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and any $c_1, \ldots, c_n \in \mathbb{R}$. It is said to be *strictly* positive definite if the equality in (2.4) implies $c_1 = c_2 = \cdots = c_n = 0$.

Moreover, a positive definite kernel in the sense of Definition 2.1 defines a space of functions from $\mathcal{X}$ to $\mathbb{R}$ called a *reproducing kernel Hilbert space* (RKHS) $\mathscr{H}$, hence also called *reproducing kernel* (Aronszajn 1950). We defer the details of this link to Section 2.2. An RKHS has two important properties: (i) for any $\mathbf{x} \in \mathcal{X}$, the function $k(\mathbf{x}, \cdot) : \mathbf{y} \mapsto k(\mathbf{x}, \mathbf{y})$ is an element of $\mathscr{H}$. That is, whenever we use the kernel k, we often think of a *canonical feature map*

$$k : \mathcal{X} \quad \to \quad \mathscr{H} \subset \mathbb{R}^{\mathcal{X}}$$

$$\mathbf{x} \quad \mapsto \quad k(\mathbf{x}, \cdot) \tag{2.5}$$

where $\mathbb{R}^{\mathcal{X}}$ denotes the vector space of functions from $\mathcal{X}$ to $\mathbb{R}$; (ii) an inner product in $\mathscr{H}$ satisfies the *reproducing property*, *i.e.*, for all $f \in \mathscr{H}$ and $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathscr{H}}. \tag{2.6}$$

Although we do not need to know $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ explicitly, it can be derived directly from the kernel k (see, *e.g.*, Schölkopf and Smola (2001) for concrete examples). Further details of RKHS will be given in Section 2.2.

The capability of kernel trick not only results in powerful learning algorithms, but also allows domain experts to easily invent domain-specific kernel functions which are suitable for particular applications. This leads to a number of kernel functions in various application domains (Genton 2002). In machine learning, commonly used kernels include the Gaussian and Laplace kernels, *i.e.*,

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right), \qquad k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right), \tag{2.7}$$

where $\sigma > 0$ is a bandwidth parameter. These kernels belong to a class of kernel functions called *radial basis functions* (RBF). Both these kernels are *translation invariant* on $\mathbb{R}^d$ which form an important class of kernel functions with essential properties (cf. Theorem 2.2).[1]

Another essential property of kernel trick is that it can be applied not only to Euclidean data, but also to non-Euclidean structured data, functional data, and other domains on which positive definite kernels may be defined (Gärtner 2003). A review of several classes of kernel functions can be found in Genton (2002). A review of kernels for vector-valued functions can be found in Álvarez et al. (2012). Hofmann et al. (2008) also provides a general review of kernel methods in machine learning.

Next, we give an important characterization of continuous positive definite kernel k on a compact set, known as the Mercer's theorem (Mercer 1909).

[1] A kernel k is said to be translation invariant on $\mathbb{R}^d$ if $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ for some positive definite function φ.

Theorem 2.1 (Mercer's theorem)**.** Let $\mathcal{X}$ be a compact Hausdorff space and μ be a finite Borel measure with support $\mathcal{X}$. Suppose k is a continuous positive definite kernel on $\mathcal{X}$, and define the integral operator $\mathcal{T}_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ by

$$(\mathcal{T}_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}), \tag{2.8}$$

which is positive definite, *i.e.*, $\forall f \in L_2(\mathcal{X}, \mu)$,

$$\int_{\mathcal{X}} k(\mathbf{u}, \mathbf{v}) f(\mathbf{u}) f(\mathbf{v}) \, \mathrm{d}\mathbf{u} \, \mathrm{d}\mathbf{v} \geq 0. \tag{2.9}$$

Then, there is an orthonormal basis $\{\psi_i\}$ of $L_2(\mathcal{X}, \mu)$ consisting of eigenfunctions of $\mathcal{T}_k$ such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ are non-negative. The eigenfunctions corresponding to non-zero eigenvalues can be taken as continuous functions on $\mathcal{X}$ and $k(\mathbf{u}, \mathbf{v})$ has the representation

$$k(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{u}) \psi_i(\mathbf{v}) \tag{2.10}$$

where the convergence is absolute and uniform.

The condition (2.9) is known as the *Mercer's condition* and the kernel functions that satisfy this condition is often referred to as *Mercer kernels*. One should note that Mercer's theorem characterizes a richer class of kernel functions than the notion of positive definiteness considered previously. That is, while all Mercer's kernels satisfy (2.4), the converse is not necessarily true. Steinwart and Scovel (2012) studied the Mercer's theorem in general domains by relaxing the compactness assumption on $\mathcal{X}$. Since we are interested in the feature map ϕ, throughout this survey, we consider positive definite kernels that satisfy (2.2). Moreover, there is an intrinsic connection between the integral operator $\mathcal{T}_k$, covariance operator $\mathcal{C}_{XX}$, and Gram matrix $\mathbf{K}$ (Rosasco et al. 2010). A fundamental connection between Mercer's theorem in functional analysis and Karhunen-Loève theorem in the theory of stochastic processes (Rogers and Williams 2000a;b) is also fruitful in stochastic processes

Table 2.1: Well-known kernel functions and their corresponding spectral densities. More examples can be found in Rasmussen and Williams (2005), Rahimi and Recht (2007), Fukumizu et al. (2009b), Kar and Karnick (2012), Pham and Pagh (2013). K_ν is a modified Bessel function of the second kind of order ν and Γ is the Gamma function. We also define $h(\nu, d\sigma) = \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)(2\nu)^\nu}{\Gamma(\nu)\sigma^{2\nu}}$.

Kernel	$k(\mathbf{x}, \mathbf{x}')$	$\Lambda(\boldsymbol{\omega})$
Gaussian	$\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right),\ \sigma > 0$	$\frac{1}{(2\pi/\sigma^2)^{d/2}} \exp\left(-\frac{-\sigma^2\|\boldsymbol{\omega}\|_2^2}{2}\right)$
Laplace	$\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_1}{\sigma}\right),\ \sigma > 0$	$\prod_{i=1}^d \frac{\sigma}{\pi(1+\omega_i^2)}$
Cauchy	$\prod_{i=1}^d \frac{\sigma}{1+(x_i-x_i')},\ \sigma > 0$	$\exp\left(-\frac{\|\boldsymbol{\omega}\|_1}{\sigma}\right)$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right) K_\nu\left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right)$ $\sigma > 0,\ \nu > 0$	$h(\nu, d\sigma)\left(\frac{2\nu}{\sigma^2} + 4\pi^2\|\boldsymbol{\omega}\|_2^2\right)^{\nu+d/2}$

literature such as Gaussian processes (Rasmussen and Williams 2005).

When the kernel k is translation invariant on $\mathbb{R}^d$, *i.e.*, $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the kernel can be characterized by Bochner's theorem (Bochner 1933), which states that any bounded continuous kernel k is the inverse Fourier transform of some finite non-negative Borel measure.

Theorem 2.2 (Bochner's theorem)**.** A complex-valued bounded continuous kernel $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$ on $\mathbb{R}^d$ is positive definite if and only if there exists a finite non-negative Borel measure Λ on $\mathbb{R}^d$ such that

$$\varphi(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} e^{\sqrt{-1}\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{x}')}\, \mathrm{d}\Lambda(\boldsymbol{\omega}). \tag{2.11}$$

Put differently, the continuous positive definite functions form a convex cone with the generators given by the Fourier kernels $\{e^{\sqrt{-1}\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{x}')} \,|\, \boldsymbol{\omega} \in \mathbb{R}^d\}$.

By virtue of Theorem 2.2, we may interpret the similarity measure $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$ in the Fourier domain. That is, the measure Λ determines which frequency component occurs in the kernel by putting non-negative power on each frequency ω. Note that we may normalize k such that $\varphi(\mathbf{0}) = 1$, in which case Λ will be a probability measure and

k corresponds to its characteristic function. For example, the measure Λ that corresponds to the Gaussian kernel $k(\mathbf{x}-\mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|_2^2/(2\sigma^2)}$ is a Gaussian distribution of the form $(2\pi/\sigma^2)^{-d/2}e^{-\sigma^2\|\boldsymbol{\omega}\|_2^2/2}\,\mathrm{d}\boldsymbol{\omega}$. For Laplacian kernel $k(\mathbf{x},\mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|_1/\sigma}$, the corresponding measure is a Cauchy distribution, *i.e.*, $\Lambda(\boldsymbol{\omega}) = \prod_{i=1}^{d}\frac{\sigma}{\pi(1+\omega_i^2)}$. Table 2.1 summarizes the Fourier transform of some well-known kernel functions.

As we will see later, Bochner's theorem also plays a central role in providing a powerful characterization of the kernel mean embedding. Similarly, the measure $\Lambda(\boldsymbol{\omega})$ determines which frequency component $\boldsymbol{\omega}$ of the characteristic function of $\mathbb{P}$ occurs in the embedding $\mu_{\mathbb{P}}$. Hence, the uniqueness of the characteristic function implies that if the support of Λ is the entire $\mathbb{R}^d$, $\mu_{\mathbb{P}}$ will uniquely determine $\mathbb{P}$ (Sriperumbudur et al. 2008; 2010; 2011a). In the context of this survey, we may think of Λ as a filter that selects certain properties when computing the similarity measure between probability distributions $\langle\mu_{\mathbb{P}},\mu_{\mathbb{Q}}\rangle_{\mathscr{H}}$ w.r.t. a certain class of distributions M_+^1 (more below).

Another promising application of Bochner's theorem is the *approximation of kernel function*, which is useful in speeding up kernel methods. The feature map ϕ of many kernel functions such as the Gaussian kernel is infinite dimensional. To avoid the need to construct the Gram matrix $\mathbf{K}_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$—which scales at least $O(n^2)$ in terms of both computation and memory usage—Rahimi and Recht (2007) proposes to preserve $k(\mathbf{x}_i,\mathbf{x}_j)$ by approximating the integral in (2.11) based on a Monte Carlo sample $\boldsymbol{\omega}\sim\Lambda$. See also Kar and Karnick (2012), Le et al. (2013), Pham and Pagh (2013) and references therein for a generalization of this idea. Theoretical results on the preservation of kernel evaluation can be found in, *e.g.*, Sutherland and Schneider (2015) and Sriperumbudur and Szabo (2015). Other common approaches to approximating the Gram matrix $\mathbf{K}$ are low-rank approximation via incomplete Cholesky decomposition (Fine and Scheinberg 2001) and Nyström method (Williams and Seeger 2001, Bach 2013).

Finally, we briefly mention *Schoenberg's characterization* (Schoenberg 1938) for *radial* kernels $k(\mathbf{x},\mathbf{y}) = \varphi(\|\mathbf{x}-\mathbf{y}\|^2)$ for some positive definite function $\varphi:[0,\infty)\to\mathbb{R}$.[2]

[2]The kernel k is said to be *radial* if $k(\mathbf{x},\mathbf{x}') = \varphi(\|\mathbf{x}-\mathbf{x}'\|)$ for some positive

Theorem 2.3 (Schoenberg's theorem)**.** A continuous function $\varphi : [0, \infty) \to \mathbb{R}$ is positive definite and radial on $\mathbb{R}^d$ for all d if and only if it is of the form

$$\varphi(r) = \int_0^\infty e^{-r^2 t^2} \, \mathrm{d}\mu(t) \tag{2.12}$$

where μ is a finite non-negative Borel measure on $[0, \infty)$.

In other words, this class of kernels can be expressed as a scaled Gaussian mixture. Examples of well-known radial kernels include Gaussian RBF kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2)$, mixture-of-Gaussians kernel $k(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^K \beta_j \exp(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma_j^2)$ where $\boldsymbol{\beta} \geq \mathbf{0}$ and $\sum_{j=1}^K \beta_j = 1$, inverse multiquadratic kernel $k(\mathbf{x}, \mathbf{y}) = (c^2 + \|\mathbf{x}-\mathbf{y}\|^2)^{-\gamma}$ where $c, \gamma > 0$, and Matérn kernel

$$k(\mathbf{x}, \mathbf{y}) = \frac{c^{2r-d}}{\Gamma(r - d/2) 2^{r-1-d/2}} \left(\frac{\|\mathbf{x} - \mathbf{y}\|_2}{c} \right)^{r-d/2} B_{d/2-r}(c\|\mathbf{x} - \mathbf{y}\|_2) \tag{2.13}$$

where $r > d/2$, $c > 0$, B_a is the modified Bessel function of the third kind of order a, and Γ is the Gamma function. For a detailed exposition, see, *e.g.*, Bavaud (2011), Wendland (2005; Ch. 7), and Rasmussen and Williams (2005).

2.2 Reproducing Kernel Hilbert Spaces

A reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ is a Hilbert space where all evaluation functionals in $\mathscr{H}$ are bounded. First, we give a formal definition of Hilbert space.

Definition 2.2. A Hilbert space is a complete normed space where the norm is induced by an inner product.

definite function $\varphi : [0, \infty) \to \mathbb{R}$. That is, it is a translation-invariant kernel that depends only on the distance between two instances.

Examples of Hilbert spaces include the standard Euclidean space $\mathbb{R}^d$ with $\langle \mathbf{x}, \mathbf{y} \rangle$ being the dot product of $\mathbf{x}$ and $\mathbf{y}$, the space of square summable sequences ℓ^2 of $\mathbf{x} = (x_1, x_2, \ldots)$ with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$, and the space of square-integrable functions $L_2[a, b]$ with inner product $\langle f, g \rangle = \int_a^b f(x) g(x) \, \mathrm{d}x$, after identifying functions having different values only on measure zero sets. Hilbert spaces with their norm given by the inner product are examples of *Banach spaces* (Megginson 1998, Folland 1999). A Hilbert space is always a Banach space, but the converse need not hold because a Banach space may have a norm that is not given by an inner product, *e.g.*, the supremum norm. This survey will deal mostly with the embedding of distributions in a Hilbert space, while extensions to Banach spaces are investigated in Zhang et al. (2009) and Sriperumbudur et al. (2011b).

We are now in a position to give the definition of a reproducing kernel Hilbert space.

Definition 2.3. A Hilbert space $\mathscr{H}$ of functions is *a reproducing kernel Hilbert space (RKHS)* if the evaluation functionals are bounded, *i.e.*, if for all $\mathbf{x} \in \mathcal{X}$ there exists some $C > 0$ such that

$$|\mathbf{F}_{\mathbf{x}}[f]| = |f(\mathbf{x})| \leq C \|f\|_{\mathscr{H}}, \qquad \forall f \in \mathscr{H}. \tag{2.14}$$

Intuitively speaking, functions in the RKHS are *smooth* in the sense of (2.14). This smoothness property ensures that the solution in RKHS obtained from learning algorithms will be well-behaved, since regularization on the RKHS norm leads to regularization on function values. For example, in classification and regression problems, it is ensured that by minimizing the empirical risk on the training data w.r.t. the functions in RKHS, we obtain a solution $\hat{f}$ that is *close* to the true solution f and also generalize well to unseen test data. This does not necessarily hold for functions in arbitrary Hilbert spaces. The space of square-integrable functions $L_2[a, b]$ does not have this property. It is very easy to find a function in $L_2[a, b]$ that attains zero risk on the training data, but performs poorly on unseen data, *i.e.*, *overfitting*.

The next theorem provides a characterization of a bounded linear operator in $\mathscr{H}$.

Theorem 2.4 (Riesz representation). If $\mathbf{A} : \mathscr{H} \to \mathbb{R}$ is a bounded linear operator in a Hilbert space $\mathscr{H}$, there exists some $g_{\mathbf{A}} \in \mathscr{H}$ such that

$$\mathbf{A}f = \langle f, g_{\mathbf{A}} \rangle_{\mathscr{H}}, \qquad \forall f \in \mathscr{H}. \tag{2.15}$$

The Riesz representation theorem will be used to prove a sufficient condition for the existence of the kernel mean embedding in an RKHS (see Lemma 3.1). By the definition of RKHS, the evaluation functional $\mathbf{F}_{\mathbf{x}}[f] = f(\mathbf{x})$ is a bounded linear operator in $\mathscr{H}$. Therefore, Riesz representation theorem ensures that for any $\mathbf{x} \in \mathcal{X}$ we can find an element in $\mathscr{H}$ that is a *representer* of the evaluation $f(\mathbf{x})$. Proposition 2.1 states this result, which is called the *reproducing property*.

Proposition 2.1 (reproducing property). For each $\mathbf{x} \in \mathcal{X}$, there exists a function $k_{\mathbf{x}} \in \mathscr{H}$ such that

$$\mathbf{F}_{\mathbf{x}}[f] = \langle k_{\mathbf{x}}, f \rangle_{\mathscr{H}} = f(\mathbf{x}). \tag{2.16}$$

The function $k_{\mathbf{x}}$ is called the reproducing kernel for the point $\mathbf{x}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a two-variable function defined by $k(\mathbf{x}, \mathbf{y}) := k_{\mathbf{y}}(\mathbf{x})$. Then, it follows from the reproducing property that

$$k(\mathbf{x}, \mathbf{y}) = k_{\mathbf{y}}(\mathbf{x}) = \langle k_{\mathbf{x}}, k_{\mathbf{y}} \rangle_{\mathscr{H}} = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathscr{H}}, \tag{2.17}$$

where $\phi(\mathbf{x}) := k_{\mathbf{x}}$ is the feature map $\phi : \mathcal{X} \to \mathscr{H}$. As mentioned earlier, we call ϕ a *canonical feature map* associated with $\mathscr{H}$ essentially because when we apply the function $k(\mathbf{x}, \mathbf{y})$ in the learning algorithms, the data points are implicitly represented by a function $k_{\mathbf{x}}$ in the feature space. As we will see later in Section 3, the kernel mean embedding is defined by means of $k_{\mathbf{x}}$ and can itself be viewed as a canonical feature map of the probability distribution.

The RKHS $\mathscr{H}$ is fully characterized by the reproducing kernel k. In fact, the RKHS uniquely determines k, and vice versa, as stated in the following theorem which is due to Aronszajn (1950):

Theorem 2.5. For every positive definite function $k(\cdot,\cdot)$ on $\mathcal{X}\times\mathcal{X}$ there exits a unique RKHS with k as its reproducing kernel. Conversely, the reproducing kernel of an RKHS is unique and positive definite.

Let $\mathscr{H}$ denote the RKHS endowed with the reproducing kernel k. Then, it can be shown that the kernel k generates the RKHS, *i.e.*,

$$\mathscr{H} = \overline{\operatorname{span}\{k(\mathbf{x},\cdot)\mid \mathbf{x}\in\mathcal{X}\}}$$

where the closure is taken w.r.t. the RKHS norm (Berlinet and Thomas-Agnan 2004; Chapter 1, Theorem 3). For example, the RKHS of translation invariant kernel $k(\mathbf{x},\mathbf{y}) = \varphi(\mathbf{x}-\mathbf{y})$ on $\mathbb{R}^d$ consists of functions whose smoothness is determined by the decay rate of the Fourier transform of φ. Matérn kernel (2.13) generates the Sobolev space $H_2^r(\mathbb{R}^d)$ of r-times differentiable functions for $r > \frac{d}{2}$. Detailed exposition on RKHS can be found in Schölkopf and Smola (2001), Berlinet and Thomas-Agnan (2004), for example.

2.3 Hilbert-Schmidt Operators

We conclude this part by describing the notion of *Hilbert-Schmidt operator* which is ubiquitous in many applications of RKHS. For example, one can show that a covariance operator—defined later in Section 3.2—is Hilbert-Schmidt. Let $\mathscr{H}$ and $\mathscr{F}$ be separable Hilbert spaces and $(h_i)_{i\in I}$ and $(f_j)_{j\in J}$ are orthonormal basis for $\mathscr{H}$ and $\mathscr{F}$, respectively.[3] A Hilbert-Schmidt operator is a bounded operator $\mathcal{A}:\mathscr{F}\to\mathscr{H}$ whose Hilbert-Schmidt norm

$$\|\mathcal{A}\|_{\mathrm{HS}}^2 = \sum_{j\in J}\|\mathcal{A}f_j\|_{\mathscr{H}}^2 = \sum_{i\in I}\sum_{j\in J}|\langle \mathcal{A}f_j, h_i\rangle_{\mathscr{H}}|^2 \tag{2.18}$$

is finite. It can be easily shown that the right hand side does not depend on the choice of orthonormal bases. The Hilbert-Schmidt operators mapping from $\mathscr{F}$ to $\mathscr{H}$ form a Hilbert space $\mathrm{HS}(\mathscr{F},\mathscr{H})$ with inner

[3] A Hilbert space is said to be separable if it has a countable basis.

product $\langle \mathcal{A}, \mathcal{B} \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle \mathcal{A} f_j, \mathcal{B} f_j \rangle_{\mathscr{H}}$. The Hilbert space of Hilbert-Schmidt operators is beyond the scope of this survey; see, *e.g.*, Rudin (1991; Chapter 12) or Reed and Simon (1981; Chapter 6) for further detail.

Recently, the Hilbert-Schmidt operators have received much attention in machine learning community and forms a backbone of many modern applications of kernel methods. For instance, Gretton et al. (2005a) uses a Hilbert-Schmidt norm of the cross-covariance operator as a measure of statistical dependence between two random variables; see also Chwialkowski and Gretton (2014) and Chwialkowski et al. (2014) for an extension to random processes. Likewise, these notions have also been applied in sufficient dimension reduction (Fukumizu et al. 2004), kernel CCA (Fukumizu et al. 2007), and kernel PCA (Zwald et al. 2004). Recently, Quang et al. (2014) proposes a *Log-Hilbert-Schmidt metric* between positive definite operators on a Hilbert space—which genearlizes the Log-Euclidean metric between positive semi-definite matrices to the infinite dimensional setting. It is applied in particular to compute distance between covariance operators in RKHS with applications in multi-category image classification.

3

Hilbert Space Embedding of Marginal Distributions

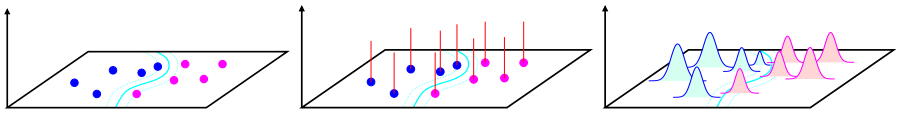
This section introduces the idea of Hilbert-space embedding of distributions by generalizing the standard viewpoint of the kernel feature map of random sample to Dirac measures, and then to more general probability measures. The presentation here is similar to that of Berlinet and Thomas-Agnan (2004; Chapter 4) with some modifications. We summarize this generalization in Figure 3.1.

3.1 From Data Point to Probability Measure

Let $\mathcal{X}$ be a fixed non-empty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a real-valued positive definite kernel function endowed with the Hilbert space $\mathscr{H}$. For all functions $f \in \mathscr{H}$, it follows from the reproducing property that $\langle k(\mathbf{x}, \cdot), f \rangle_{\mathscr{H}} = f(\mathbf{x})$. In particular, $k(\mathbf{x}, \mathbf{y}) = \langle k(\mathbf{x}, \cdot), k(\mathbf{y}, \cdot) \rangle_{\mathscr{H}}$. By virtue of this property, we may view the kernel evaluation as an inner product in $\mathscr{H}$ induced by a map from $\mathcal{X}$ into $\mathscr{H}$

$$\mathbf{x} \longmapsto k(\mathbf{x}, \cdot). \tag{3.1}$$

In other words, $k(\mathbf{x}, \cdot)$ is a high-dimensional *representer* of $\mathbf{x}$. Moreover, by the reproducing property it also acts as a *representer of evaluation* of any function in $\mathscr{H}$ on the data point $\mathbf{x}$. These properties of kernels



(a) $\mathbf{x} \mapsto k(\mathbf{x}, \cdot)$ **(b)** $\delta_{\mathbf{x}} \mapsto \int k(\mathbf{y}, \cdot)\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{y})$ **(c)** $\mathbb{P} \mapsto \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x})$

Figure 3.1: From data points to probability measures: (a) An illustration of typical application of kernel as a high-dimensional feature map of individual data point. (b) A measure-theoretic view of high-dimensional feature map. An embedding of data point into a high-dimensional feature space can be equivalently viewed as an embedding of a Dirac measure assigning the mass 1 to each data point. (c) Generalizing the Dirac measure point of view, we can generally extend the concept of high-dimensional feature map to the class of probability measures.

are imperative in practical applications because if an algorithm can be formulated in terms of an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$, one can construct an alternative algorithm by replacing the inner product by a positive definite kernel $k(\mathbf{x}, \mathbf{y})$ without building the mapping of $\mathbf{x}$ and $\mathbf{y}$ explicitly (a.k.a. the *kernel trick*). Well-known examples of kernelizable learning algorithms include *support vector machine (SVM)* (Cortes and Vapnik 1995) and *principle component analysis (PCA)* (Hotelling 1933).

We can generalize the concept of high-dimensional feature map of data points $\mathbf{x} \in \mathcal{X}$ to measures on a measurable space $(\mathcal{X}, \mathcal{A})$ where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$. The simplest example of measures is the Dirac measure $\delta_{\mathbf{x}}$ defined for $\mathbf{x}$ in $\mathcal{X}$ by

$$\delta_{\mathbf{x}}(A) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{if } \mathbf{x} \notin A, \end{cases} \tag{3.2}$$

where $A \in \mathcal{A}$. Since any measurable function f on $\mathcal{X}$ is integrable w.r.t. $\delta_{\mathbf{x}}$, we have

$$\int f(\mathbf{t})\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) = f(\mathbf{x}). \tag{3.3}$$

When f belongs to the Hilbert space $\mathscr{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, we can rewrite (3.3) using the reproducing property of $\mathscr{H}$ as

$$\begin{aligned} \int f(\mathbf{t})\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) &= \int \langle f, k(\mathbf{t}, \cdot) \rangle_{\mathscr{H}}\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) \\ &= \left\langle f, \int k(\mathbf{t}, \cdot)\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) \right\rangle_{\mathscr{H}} = \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathscr{H}}. \end{aligned} \tag{3.4}$$

Like in the case of input space $\mathcal{X}$, the function $\int k(\mathbf{t}, \cdot)\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t})$ acts as a representer of the measure $\delta_{\mathbf{x}}$ in the Hilbert space. Also, it may be viewed as a representer of evaluation of the following functional:

$$f \longmapsto \int f(\mathbf{t})\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}), \tag{3.5}$$

namely, the expectation of f w.r.t. the Dirac measure $\delta_{\mathbf{x}}$. Although integrating f w.r.t. $\delta_{\mathbf{x}}$ or evaluating $\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathscr{H}}$ gives the same result $f(\mathbf{x})$, *i.e.*, the value of f at point $\mathbf{x}$, the former gives a measure-theoretic point of view of the latter (see also Figure 3.1). Consequently, we can define a feature map from the space of Dirac measures to $\mathscr{H}$ as

$$\delta_{\mathbf{x}} \longmapsto \int_{\mathcal{X}} k(\mathbf{y}, \cdot)\, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{y}). \tag{3.6}$$

Intuitively, the Dirac measure $\delta_{\mathbf{x}}$ is a probability measure on $(\mathcal{X}, \mathcal{A})$ assigning the mass 1 to the set $\{\mathbf{x}\}$. This implies that one can immediately extend any learning algorithm that operates on a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to a set of probability measures $\delta_{\mathbf{x}_1}, \ldots, \delta_{\mathbf{x}_n}$ (Muandet et al. 2012). However, as we can see in (3.4) this extension is not quite useful in practice because both algorithms are in fact equivalent. In what follows, we will consider more interesting cases of non-trivial probability measures.

More generally, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n distinct points in $\mathcal{X}$ and $a_1, \ldots, a_n$ are n non-zero real numbers, we consider a linear combination

$$\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \tag{3.7}$$

of Dirac measures putting the mass a_i at the point $\mathbf{x}_i$. This is a *signed* measure which constitutes a class of measures with finite support. A measure of the form (3.7) is ubiquitous in machine learning community, especially in Bayesian probabilistic inference (Adams 2009). For example, if $a_i = 1/n$ for all i, we obtain an *empirical measure* associated with a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Donsker measure is obtained when a_i is also a random variable (Berlinet and Thomas-Agnan 2004). Lastly, if $a_i = 1$ for all i, the measure of the form (3.7) represents an instance of a *point process* on $\mathcal{X}$ which has numerous applications in Bayesian nonparametric inference and neural coding (Dayan and Abbott 2005).

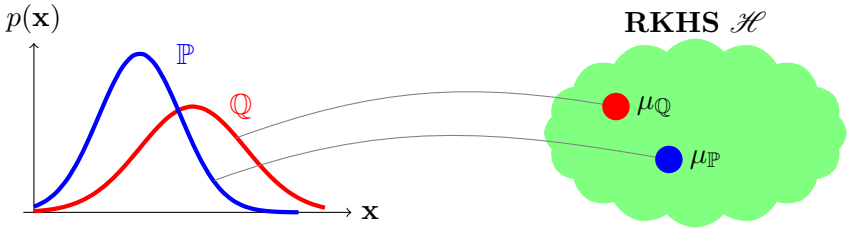


Figure 3.2: Embedding of marginal distributions: Each distribution is mapped into a RKHS via an expectation operation.

A *determinantal point process (DPP)*—*i.e.*, a point process with repulsive property —has recently gained popularity in machine learning community (Kulesza and Taskar 2012).

Likewise, for any measurable function f we have

$$\int f \,\mathrm{d}\left(\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}\right) = \sum_{i=1}^{n} a_i \int f \,\mathrm{d}\delta_{\mathbf{x}_i} = \sum_{i=1}^{n} a_i f(\mathbf{x}_i). \tag{3.8}$$

This extends previous remark on Dirac measures to measures with finite support, and if f belongs to $\mathscr{H}$, we obtain similar results as in the case of Dirac measure. That is, the mapping

$$\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \longmapsto \sum_{i=1}^{n} a_i k(\mathbf{x}_i, \cdot) \tag{3.9}$$

gives a representer in $\mathscr{H}$ of a measure with finite support. Furthermore, it is a representer of expectation w.r.t. the measure, *i.e.*, if $\mu := \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$, we have for any f in $\mathscr{H}$

$$\left\langle f, \sum_{i=1}^{n} a_i k(\mathbf{x}_i, \cdot) \right\rangle_{\mathscr{H}} = \sum_{i=1}^{n} a_i f(\mathbf{x}_i) = \int f \,\mathrm{d}\mu. \tag{3.10}$$

In particular, for any Hilbert space $\mathscr{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, a linear combination $\sum_{i=1}^{n} a_i k(\mathbf{x}_i, \cdot)$ forms a dense subset of $\mathscr{H}$. Here some readers may have concern regarding the measurability of f. This is easily seen, however, since it is known that point-wise convergence of measurable functions gives a measurable function.

In what follows, we use $M_+^1(\mathcal{X})$ to denote the space of probability measures over a measurable space $\mathcal{X}$. Then, we can define the repre-

senter in $\mathscr{H}$ of any probability measure $\mathbb{P}$ through the mapping

$$\mu : M_+^1(\mathcal{X}) \longrightarrow \mathscr{H}, \quad \mathbb{P} \longmapsto \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}) \tag{3.11}$$

which will be denoted by $\mu_{\mathbb{P}}$. Here the integral should be interpreted as Bochner integral, which is in general an integral for Banach space valued functions. See Diestel and Uhl (1977; Chapter 2) and Dinculeanu (2000; Chapter 1) for details. The above mapping is essentially the kernel mean embedding we consider throughout the survey.

Definition 3.1 (Berlinet and Thomas-Agnan (2004), Smola et al. (2007))**.** Suppose that the space $M_+^1(\mathcal{X})$ consists of all probability measures $\mathbb{P}$ on a measurable space $(\mathcal{X}, \Sigma)$. The kernel mean embedding of probability measures in $M_+^1(\mathcal{X})$ into an RKHS $\mathscr{H}$ endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by a mapping

$$\mu : M_+^1(\mathcal{X}) \longrightarrow \mathscr{H}, \quad \mathbb{P} \longmapsto \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}).$$

Next, we provide the conditions under which the embedding $\mu_{\mathbb{P}}$ exists and belongs to $\mathscr{H}$.

Lemma 3.1 (Smola et al. (2007))**.** If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{k(X,X)}] < \infty$, the $\mu_{\mathbb{P}} \in \mathscr{H}$ and $\mathbb{E}_{\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}$.

Proof. Let $\mathbf{L}_{\mathbb{P}}$ be a linear operator defined as $\mathbf{L}_{\mathbb{P}} f := \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$. Under the assumption, $\mathbf{L}_{\mathbb{P}}$ is bounded for all $f \in \mathscr{H}$, *i.e.*,

$$\begin{aligned} |\mathbf{L}_{\mathbb{P}} f| = |\mathbb{E}_{X\sim\mathbb{P}}[f(X)]| &\overset{(*)}{\leq} \mathbb{E}_{X\sim\mathbb{P}}[|f(X)|] \\ &= \mathbb{E}_{X\sim\mathbb{P}}[|\langle f, k(X,\cdot)\rangle_{\mathscr{H}}|] \\ &\leq \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{k(X,X)}\|f\|_{\mathscr{H}}\right], \end{aligned}$$

where we use Jensen's inequality in $(*)$. Hence, by Riesz representation theorem (see, *e.g.*, Theorem 2.4), there exists $\lambda \in \mathscr{H}$ such that $\mathbf{L}_{\mathbb{P}} f =$

$\langle f, \lambda \rangle_{\mathscr{H}}$. Choose $f = k(\mathbf{x}, \cdot)$ for some $\mathbf{x} \in \mathcal{X}$. Then, $\lambda(\mathbf{x}) = \mathbf{L}_{\mathbb{P}} k(\mathbf{x}, \cdot) = \int k(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}')$ which means $\lambda = \int k(\cdot, \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}') = \mu_{\mathbb{P}}$. ■

From the proof of Lemma 3.1, $\mathbb{E}_{X \sim \mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathscr{H}}$ for any $f \in \mathscr{H}$. This equality can essentially be viewed as a *reproducing property* of the expectation operation in the RKHS. That is, it allows us to compute the expectation of a function f in the RKHS w.r.t. the distribution $\mathbb{P}$ by means of an inner product between the function f and the embedding $\mu_{\mathbb{P}}$. This property has proven useful in certain applications such as graphical model and probabilistic inference that require an evaluation of expectation w.r.t. the model (Song et al. 2010a; 2011a, Boots et al. 2013, McCalman et al. 2013). This property can be extended to conditional distributions as well (see §4).

3.1.1 Explicit Representation of Mean Embeddings

It is important to understand what information of the distribution is retained by the kernel mean embedding. For a linear kernel $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$, it is clear that $\mu_{\mathbb{P}}$ becomes just the first moment of $\mathbb{P}$, whereas for the polynomial kernel $k(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^2$ the mean map retains both the first and the second moments of $\mathbb{P}$. Below we provide some explicit examples which can also be found in, *e.g.*, Smola et al. (2007), Fukumizu et al. (2008), Sriperumbudur et al. (2010), Gretton et al. (2012a), Schölkopf et al. (2015).

Example 3.1 (inhomogeneous polynomial kernel)**.** Let consider the inhomogeneous polynomial kernel $k(\mathbf{x}, \mathbf{x}) = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^p$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ of degree p. Using

$$\begin{aligned}(\langle \mathbf{x}, \mathbf{y} \rangle + 1)^p &= 1 + \binom{p}{1} \langle \mathbf{x}, \mathbf{y} \rangle + \binom{p}{2} \langle \mathbf{x}, \mathbf{y} \rangle^2 + \binom{p}{3} \langle \mathbf{x}, \mathbf{y} \rangle^3 + \cdots \\ &= 1 + \binom{p}{1} \langle \mathbf{x}, \mathbf{y} \rangle + \binom{p}{2} \langle \mathbf{x}^{(2)}, \mathbf{y}^{(2)} \rangle \\ &\quad + \binom{p}{3} \langle \mathbf{x}^{(3)}, \mathbf{y}^{(3)} \rangle + \cdots\end{aligned}$$

where $\mathbf{x}^{(i)}$ denotes the ith-order tensor product (Schölkopf and Smola 2001; Proposition 2.1), the kernel mean embedding can be written explicitly as

$$\begin{aligned}\mu_{\mathbb{P}}(\mathbf{t}) &= \int (\langle \mathbf{x}, \mathbf{t} \rangle + 1)^p \, \mathrm{d}\mathbb{P}(\mathbf{x}) \\ &= 1 + \binom{p}{1} \langle \mathbf{m}_{\mathbb{P}}(1), \mathbf{t} \rangle + \binom{p}{2} \langle \mathbf{m}_{\mathbb{P}}(2), \mathbf{t}^{(2)} \rangle \\ &\quad + \binom{p}{3} \langle \mathbf{m}_{\mathbb{P}}(3), \mathbf{t}^{(3)} \rangle + \cdots ,\end{aligned}$$

where $\mathbf{m}_{\mathbb{P}}(i)$ denotes the ith moment of the distribution $\mathbb{P}$.

As we can see from Example 3.1, the embedding incorporates up to the m-th moment of $\mathbb{P}$. As we increase p, more information about $\mathbb{P}$ is stored in the kernel mean embedding.

Example 3.2 (moment-generating function)**.** Consider $k(\mathbf{x}, \mathbf{x}') = \exp(\langle \mathbf{x}, \mathbf{x}' \rangle)$. Hence, we can write the kernel mean embedding as

$$\mu_{\mathbb{P}} = \mathbb{E}_{X \sim \mathbb{P}} \left[e^{\langle X, \cdot \rangle} \right],$$

which is essentially the *moment-generating function* (MGF) of a random variable X with distribution $\mathbb{P}$, given that the MGF exists.

The MGF is essentially the Laplace transformation of a random variable X, which does not need to exist as it requires in particular the existence of moments of any order. On the other hand, the characteristic function—which is the Fourier transformation—always exists for any probability distribution $\mathbb{P}$.

Example 3.3 (characteristic function)**.** First, consider the Fourier kernel $k(\mathbf{x}, \mathbf{y}) = \exp(i \mathbf{x}^\top \mathbf{y})$ using which we can express $\mu_{\mathbb{P}}$ as

$$\mu_{\mathbb{P}}(\mathbf{t}) = \mathbb{E}_{X \sim \mathbb{P}}[k(X, \mathbf{t})] = \mathbb{E}_{X \sim \mathbb{P}}[\exp(i X^\top \mathbf{t})].$$

It is essentially the characteristic function of $\mathbb{P}$, which we denote

by $\varphi_{\mathbb{P}}$ thereafter. Given i.i.d. sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the empirical estimate $\hat{\mu}_{\mathbb{P}}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \exp(i\mathbf{x}_i^\top \mathbf{t})$ becomes an empirical characteristic function (ECF).

Next, consider any translation invariant kernel $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where ψ is a positive definite function. Bochner's theorem (see Theorem 2.2) allows us to express the kernel mean embedding as

$$\begin{aligned}
\mu_{\mathbb{P}}(\mathbf{t}) &= \int \psi(\mathbf{x} - \mathbf{t}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \\
&= \iint_{\mathbb{R}^d} \exp\left(i\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{t})\right) \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \\
&= \iint_{\mathbb{R}^d} \exp\left(i\boldsymbol{\omega}^\top \mathbf{x}\right) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \exp\left(-i\boldsymbol{\omega}^\top \mathbf{t}\right) \, \mathrm{d}\Lambda(\, \mathrm{d}\boldsymbol{\omega}) \\
&= \int_{\mathbb{R}^d} \varphi_P(\boldsymbol{\omega}) \exp\left(-i\boldsymbol{\omega}^\top \mathbf{t}\right) \, \mathrm{d}\Lambda(\boldsymbol{\omega})
\end{aligned}$$

for some positive finite measure Λ (Sriperumbudur et al. 2010). It is not difficult to show that for $k(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$ we have $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathscr{H}} = \langle \varphi_P, \varphi_Q \rangle_{L^2(\mathbb{R}^d, \Lambda)}$.

3.1.2 Empirical Estimation of Mean Embeddings

In practice, we do not have access to the true distribution $\mathbb{P}$, and thereby cannot compute $\mu_{\mathbb{P}}$. Instead, we must rely entirely on the sample from this distribution. Given an i.i.d. sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the most common empirical estimate, denoted by $\hat{\mu}_{\mathbb{P}}$ of the kernel mean $\mu_{\mathbb{P}}$ is

$$\hat{\mu}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \cdot). \tag{3.12}$$

Clearly, $\hat{\mu}_{\mathbb{P}}$ is an unbiased estimate of $\mu_{\mathbb{P}}$, and by the law of large number, $\hat{\mu}_{\mathbb{P}}$ converges to $\mu_{\mathbb{P}}$ as $n \to \infty$. Sriperumbudur et al. (2012) provides a thorough discussion on several properties of this estimator. We will defer the detail of kernel mean estimation and further discussion on some other estimators, *e.g.*, shrinkage estimators, to §3.4.

Some authors might also consider explicitly the mean map of the sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We view this case as a special case of mean

map of distribution when the distribution is an empirical distribution associated with the sample $\mathbf{X}$. In which case, the mean embedding takes the form of an empirical estimate $\hat{\mu}_{\mathbb{P}}$.

In summary, under suitable assumptions on the kernel k, the Hilbert-space embedding of distributions allows us to apply RKHS methods to probability measures. Throughout this section, we restrict our attention to the space of marginal distributions $\mathbb{P}(X)$, and defer an extension to the space of conditional distributions $\mathbb{P}(Y|X)$ to §4.

3.2 Covariance Operators

In addition to the mean element, the *covariance* and *cross-covariance* operators on RKHSs are important concepts for modern applications of Hilbert space embedding of distributions. In principle, they are generalizations of covariance and cross-covariance matrices in Euclidean space to the infinite-dimensional elements in RKHSs. We give a brief review here; see Baker (1973), Fukumizu et al. (2004) for further detail.

Cross-covariance operators were introduced in Baker (1970) and then treated more extensively in Baker (1973). Let (X, Y) be a random variable taking values on $\mathcal{X} \times \mathcal{Y}$ and $(\mathscr{H}, k)$ and $(\mathscr{F}, l)$ be RKHSs with measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Throughout we assume the integrability

$$\mathbb{E}_X[k(X, X)] \leq \infty, \quad \mathbb{E}_Y[l(Y, Y)] \leq \infty, \tag{3.13}$$

which ensures that $\mathscr{H} \subset L^2(\mathbb{P}_X)$ and $\mathscr{F} \subset L^2(\mathbb{P}_Y)$, since $\int f^2(x)\,\mathrm{d}\mathbb{P}_X(x) = \int \langle f, k(\cdot, x)\rangle^2\,\mathrm{d}\mathbb{P}_X(x) \leq \int \|f\|^2 \|k(\cdot, X)\|^2\,\mathrm{d}\mathbb{P}_X \leq \|f\|^2 \mathbb{E}_X[k(X, X)]$, and similar to $\int g(y)^2\,\mathrm{d}\mathbb{P}_Y(y)$. The (uncentered) *cross-covariance operator* $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$ is defined as

$$\mathcal{C}_{YX} := \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] = \mu_{\mathbb{P}_{YX}}, \tag{3.14}$$

where $\mathbb{P}_{YX}$ denotes the joint distribution of (X, Y) and ϕ (resp. φ) is the canonical feature map corresponding to k (resp. l). The corresponding centered version of $\mathcal{C}_{YX}$ is given by $\tilde{\mathcal{C}}_{YX} := \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] - \mu_{\mathbb{P}_Y} \otimes \mu_{\mathbb{P}_X} = \mu_{\mathbb{P}_{YX}} - \mu_{\mathbb{P}_Y \otimes \mathbb{P}_X}$.

In the above definition, one needs to note that an operator can be identified with an element in the product space. More precisely, the

space of Hilbert-Schmidt operators $\mathrm{HS}(\mathscr{H}, \mathscr{F})$ (Section 2.3) is isomorphic as Hilbert spaces to the product space $\mathscr{H} \otimes \mathscr{F}$ given by the product kernel. The isomorphism is defined by

$$\begin{aligned} \mathscr{H} \otimes \mathscr{F} &\to \mathrm{HS}(\mathscr{H}, \mathscr{F}) \\ \sum_i f_i \otimes g_i &\mapsto [h \mapsto \sum_i \langle h, f_i \rangle_{\mathscr{H}} g_i]. \end{aligned} \tag{3.15}$$

Note that, given an orthonormal basis $\{\phi_i\}_a$ of $\mathscr{H}$ and $\{\psi_b\}_b$ of $\mathscr{F}$, we have $\| \sum_i \langle \cdot, f_i \rangle_{\mathscr{H}} g_i \|^2_{\mathrm{HS}} = \sum_a \sum_b \{ \sum_i \langle \phi_a, f_i \rangle_{\mathscr{H}} \langle \psi_b, g_i \rangle_{\mathscr{F}} \}^2 = \sum_a \sum_b \langle \sum_i f_i \otimes g_i, \phi_a \otimes \psi_b \rangle^2_{\mathscr{H} \otimes \mathscr{F}} = \| \sum_i f_i \otimes g_i \|^2_{\mathscr{H} \otimes \mathscr{F}}$, where the last equality is based on the fact that $\{\phi_a \otimes \psi_b\}_{a,b}$ is an orthonormal basis of $\mathscr{H} \otimes \mathscr{F}$. This implies the above map is an isometry.

To see that the cross-covariance operator $\mathcal{C}_{YX}$ is well-defined as an Hilbert-Schmidt operator, it suffices from (3.15) that $\| \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] \|_{\mathscr{H} \otimes \mathscr{F}} < \infty$ is shown. From $\| \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] \|_{\mathscr{H} \otimes \mathscr{F}} \leq \mathbb{E}_{YX} \| \varphi(Y) \otimes \phi(X) \|_{\mathscr{H} \otimes \mathscr{F}} \leq \mathbb{E}_{YX}[\sqrt{k(X,X) l(Y,Y)}] \leq \{ \mathbb{E}_X[k(X,X)] \mathbb{E}_Y[l(Y,Y)] \}^{1/2}$, the assumption (3.13) guarantees the existence.

Alternatively, we may define an operator $\mathcal{C}_{YX}$ as a unique bounded operator that satisfies

$$\langle g, \mathcal{C}_{YX} f \rangle = \mathrm{Cov}[g(Y), f(X)]$$

for all $f \in \mathscr{H}$ and $g \in \mathscr{F}$. It can be shown using the Hilbert-Schmidt theory (see §2.3) that—under the stated assumptions—this operator is of the form (3.14). These two equivalent definitions stem from the relations between the covariance operator and mean element of the joint measure $\mathbb{P}_{XY}$ (Baker 1973). To see that (3.14) leads to the result above, note that

$$\begin{aligned} \langle g, \mathcal{C}_{YX} f \rangle_{\mathscr{F}} &= \langle \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)], g \otimes f \rangle_{\mathscr{F} \otimes \mathscr{H}} \\ &= \mathbb{E}_{YX}[\langle g \otimes f, \varphi(Y) \otimes \phi(X) \rangle_{\mathscr{F} \otimes \mathscr{H}}] \\ &= \mathbb{E}_{YX}[\langle g, \varphi(Y) \rangle_{\mathscr{F}} \langle \phi(X), f \rangle_{\mathscr{H}}] \\ &= \mathbb{E}_{YX}[g(Y) f(X)] =: \mathrm{Cov}[g(Y), f(X)]. \end{aligned} \tag{3.16}$$

If $X = Y$, we call $\mathcal{C}_{XX}$ the *covariance operator*, which is self-adjoint and positive.

For any $f \in \mathscr{H}$, we can also write the integral expressions for $\mathcal{C}_{YX}$ and $\mathcal{C}_{XX}$, respectively, as

$$\begin{aligned}(\mathcal{C}_{YX} f)(\cdot) &= \int_{\mathcal{X}\times\mathcal{Y}} l(\cdot, y) f(x) \, \mathrm{d}\mathbb{P}_{XY}(x, y) \\ (\mathcal{C}_{XX} f)(\cdot) &= \int_{\mathcal{X}} k(\cdot, x) f(x) \, \mathrm{d}\mathbb{P}_{X}(x),\end{aligned}$$

where $\mathbb{P}_X$ denotes the marginal distribution of X and $\mathbb{P}_{XY}$ denotes the joint distribution of (X, Y). The first equation can be confirmed by plugging $g = l(\cdot, y')$ into (3.16), and similarly the second one. Note that $\langle g, \mathcal{C}_{YX} f\rangle_{\mathscr{F}} = \langle \mathcal{C}_{XY} g, f\rangle_{\mathscr{H}}$. Hence, $\mathcal{C}_{XY}$ is the adjoint of $\mathcal{C}_{YX}$.

Below we outline a basic result that will become crucial for the definition of conditional mean embedding presented in Section 4.

Theorem 3.2 (Fukumizu et al. (2004), Fukumizu et al. (2013; Theorem 1))**.** If $\mathbb{E}_{YX}[g(Y)|X = \cdot] \in \mathscr{H}$ for $g \in \mathscr{F}$, then

$$\mathcal{C}_{XX}\mathbb{E}_{YX}[g(Y)|X = \cdot] = \mathcal{C}_{XY} g.$$

Given an i.i.d. sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ on $\mathcal{X} \times \mathcal{Y}$, an empirical estimate of the *centered* $\mathcal{C}_{YX}$ can be obtained as

$$\begin{aligned}\widehat{\mathcal{C}}_{YX} &= \frac{1}{n}\sum_{i=1}^{n} \{l(\mathbf{y}_i, \cdot) - \hat{\mu}_{\mathbb{P}_Y}\} \otimes \{k(\mathbf{x}_i, \cdot) - \hat{\mu}_{\mathbb{P}_X}\} \\ &= \frac{1}{n}\Psi\mathbf{H}\Phi^\top, \qquad (3.17)\end{aligned}$$

where $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbb{1}_n$ is the centering matrix with $\mathbb{1}_n$ an $n \times n$ matrix of ones, $\Phi = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))^\top$ and $\Psi = (\varphi(\mathbf{y}_1), \ldots, \varphi(\mathbf{y}_n))^\top$. The empirical covariance operator $\widehat{\mathcal{C}}_{XX}$ can be obtained in a similar way, *i.e.*, $\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\Phi\mathbf{H}\Phi^\top$. It has been shown that $\|\widehat{\mathcal{C}}_{YX} - \mathcal{C}_{YX}\|_{\mathrm{HS}} = O_p(1/\sqrt{n})$ as $n \to \infty$ (Berlinet and Thomas-Agnan 2004). Since $\mathcal{C}_{YX}$ can be seen as an element in $\mathscr{H} \otimes \mathscr{F}$, this result is a sequel of the $\sqrt{n}$-consistency of the empirical mean $\hat{\mu}_{\mathbb{P}}$ (cf. Theorem 3.4) because the Hilbert-Schmidt norm of an operator from $\mathscr{H}$ to $\mathscr{F}$ corresponds to the norm in $\mathscr{H} \otimes \mathscr{F}$.

The following result due to Baker (1973) states that the cross-covariance operator can be decomposed into the covariance of the marginals and the correlation.

Theorem 3.3 (Baker (1973; Theorem 1))**.** There exists a unique bounded operator $\mathcal{V}_{YX} : \mathscr{H} \to \mathscr{F}$, $\|\mathcal{V}\| \leq 1$, such that

$$\mathcal{C}_{YX} = \mathcal{C}_{YY}^{1/2} \mathcal{V}_{YX} \mathcal{C}_{XX}^{1/2}, \tag{3.18}$$

where $\mathcal{R}(\mathcal{V}_{YX}) \subset \overline{\mathcal{R}(\mathcal{C}_{YY})}$ and $\mathcal{N}(\mathcal{V}_{YX})^{\perp} \subset \overline{\mathcal{R}(\mathcal{C}_{XX})}$.

The operator $\mathcal{V}_{YX}$ is often referred to as the *normalized cross-covariance operator* and has been used as a basis for conditional dependence measure. It is also essentially related to the canonical correlation. See, *e.g.*, Fukumizu et al. (2007; 2008). Compared to $\mathcal{C}_{YX}$, $\mathcal{V}_{YX}$ captures the same information about the dependence of X and Y, but with less influence of the marginal distributions.

The covariance operator serves as a basic building block in classical kernel-based methods such as kernel PCA (Schölkopf et al. 1998, Zwald et al. 2004), kernel Fisher discriminant, kernel CCA (Fukumizu et al. 2007), and kernel ICA (Bach and Jordan 2003). More recent applications of covariance operator include independence and conditional independence measures (Gretton et al. 2005b, Zhang et al. 2008; 2011, Doran et al. 2014). See §3.6 and §4.6 for more details.

Lastly, it is instructive to point out the connection between covariance operator $\mathcal{C}_{XX}$ and integral operator $\mathcal{T}_k$ defined in Theorem 2.1. Since $\mathcal{T}_k^{1/2}$ is an isometry from $L_2(\mathcal{X})$ to $\mathscr{H}$, one may define $\mathcal{C}_{XX}$ directly on $L_2(\mathcal{X})$ which happens to coincide with the operator $\mathcal{T}_k$. Hence, they have the same eigenvalues (Hein and Bousquet 2004, Rosasco et al. 2010). A detailed exposition on this connection is also given in Section 2.1 of Bach (2015).

3.3 Properties of the Mean Embedding

The following result, which appears in Lopez-Paz et al. (2015), is a slight modification of Theorem 27 from Song (2008) which establishes the convergence of the empirical mean embedding $\hat{\mu}_{\mathbb{P}}$ to the embedding

of its population counterpart $\mu_{\mathbb{P}}$ in RKHS norm:[1]

Theorem 3.4. Assume that $\|f\|_\infty \leq 1$ for all $f \in \mathscr{H}$ with $\|f\|_{\mathscr{H}} \leq 1$. Then with probability at least $1-\delta$ we have

$$\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathscr{H}} \leq 2\sqrt{\frac{\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x},\mathbf{x})]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}. \tag{3.19}$$

As we can see, the convergence happens at a rate $O_p(n^{-1/2})$. While various estimators have been studied for $\mu_{\mathbb{P}}$ (Muandet et al. 2016, Sriperumbudur 2016), Tolstikhin et al. (2016) recently showed that this rate is minimax optimal and so the empirical estimator $\hat{\mu}_{\mathbb{P}}$ is a minimax optimal estimator of $\mu_{\mathbb{P}}$. Moreover, $\sqrt{n}(\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}})$ converges to a zero mean Gaussian process on $\mathscr{H}$ (Berlinet and Thomas-Agnan 2004; Section 9.1).

3.3.1 Characteristic and Universal Kernels

In this section we discuss the notion of *characteristic kernel* which can be formally defined as follows.

Definition 3.2. A kernel k is said to be characteristic if the map $\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective. The RKHS $\mathscr{H}$ is said to be characteristic if its reproducing kernel is characteristic.

A characteristic kernel is essential for kernel mean embedding as for the characteristic kernel k, $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. In other words, there is no information loss when mapping the distribution into the Hilbert space. It was first introduced in Fukumizu et al. (2004) being the kernels that satisfy Definition 3.2.[2] Subsequently, Fukumizu et al. (2008) shows that Gaussian and Laplacian kernels are characteristic on $\mathbb{R}^d$. The properties of characteristic kernels were explored further in Sriperumbudur et al. (2008; 2010; 2011a).

[1] The similar result has also appeared in Smola et al. (2007; Theorem 2) and Gretton et al. (2012a).

[2] The terminology *probability determining* is used there.

Characterization of Characteristic Kernels

There are several characterizations of characteristic kernels. Intuitively, for the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ to be injective, the RKHS endowed with the kernel k should contain a sufficiently rich class of functions to represent all higher order moments of $\mathbb{P}$. Fukumizu et al. (2008; Lemma 1)—see also Fukumizu et al. (2009a; Prop. 5)—shows that, for $q \geq 1$, if the sum[3] $\mathscr{H}_k + \mathbb{R}$ is dense in $L^q(\mathcal{X}, \mathbb{P})$ for any probability $\mathbb{P}$ on $(\mathcal{X}, \mathcal{A})$, the kernel k is characteristic. As a result, universal kernels on compact domains (Steinwart 2002a) are characteristic as the associated RKHS is dense in $L^2(\rho)$ for any ρ; see also Gretton et al. (2007; Theorem 3). For non-compact spaces, Fukumizu et al. (2008; Theorem 2) proves that the kernel $k(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x} - \mathbf{y})$ is characteristic w.r.t. the Borel σ-field if for any $\xi \in \mathbb{R}^d$ there exists τ_0 such that its Fourier transform $\tilde{\varphi}(\cdot)$ satisfies $\int \frac{\tilde{\varphi}(\tau(t+\xi))^2}{\tilde{\varphi}(t)} \, \mathrm{d}t < \infty$ for all $\tau > \tau_0$. Gaussian and Laplacian kernels satisfy this condition, and hence are characteristic. As we can see, it is required that $\tilde{\varphi}(t) > 0$ for all t for such an integral to be finite. This property was studied further in Sriperumbudur et al. (2008; 2010; 2011a) who showed that any translation-invariant kernel is characteristic if the support of the Fourier transform of the kernel is an entire $\mathbb{R}^d$. Specifically, for $k(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x} - \mathbf{y})$, it follows that

$$\mu_{\mathbb{P}} = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi(\mathbf{x} - \cdot)] = \varphi * \mathbb{P}. \tag{3.20}$$

where $*$ denotes the convolution. Hence, $\widehat{\varphi * \mathbb{P}} = \Lambda \widehat{\mathbb{P}}$ where Λ is a Fourier transform of φ and $\widehat{\mathbb{P}}$ denotes the Fourier transform of $\mathbb{P}$, which is related to its characteristic function $\varphi_{\mathbb{P}}$.[4] Since the characteristic function $\varphi_{\mathbb{P}}$ uniquely determines $\mathbb{P}$, so does $\mu_{\mathbb{P}}$ if Λ is everywhere positive. Later, it was shown in Sriperumbudur et al. (2009) that *integrally strictly positive definite* kernels are characteristic.[5] Examples

[3] The sum of two RKHS's corresponds to the positive definite kernel given by the sum of the respective kernels.

[4] Generally speaking, we may view (3.20) as an *integral transform* of the distribution, *e.g.*, Fourier and Laplace transforms. Thus, from the continuity of $\widehat{\mathbb{P}}$, if the transform is everywhere positive, there will be a one-to-one correspondence between the distributions and their mean embeddings.

[5] A measurable and bounded kernel k is said to be strictly positive definite on a topological space $\mathcal{M}$ if and only if $\iint_{\mathcal{M}} k(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\nu(\mathbf{x}) \, \mathrm{d}\nu(\mathbf{y}) > 0$ for all finite non-zero signed Borel measures ν on $\mathcal{M}$.

of kernels that are strictly positive definite also include translation-variant kernels, kernels on non-compact domains, and Matern kernels. We summarize various characteristic kernels in Table 3.1. For kernels on non-standard input space such as group and semi-group, see, *e.g.*, Fukumizu et al. (2009b).

Given a characteristic kernel k, we can generate new characteristic kernels through a *conformal mapping* given by $\tilde{k}(\mathbf{x}, \mathbf{y}) := f(\mathbf{x})k(\mathbf{x}, \mathbf{y})f(\mathbf{y})$ for any bounded continuous function f such that $f(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$ and $k(\mathbf{x}, \mathbf{x})|f(\mathbf{x})|^2$ is bounded (Fukumizu et al. 2009b; Lemma 2), see, also Wu and Amari (2002).

The richness of RKHS has previously been studied through the notions of *universal* kernels (Steinwart 2002b). A continuous positive definite kernel k on a compact metric space $\mathcal{X}$ is said to be universal in the sense of Steinwart (2002b) if the corresponding RKHS $\mathscr{H}$ is dense in the space of bounded continuous functions on $\mathcal{X}$, $C_b(\mathcal{X})$. That is, for any $f \in C_b(\mathcal{X})$ and $\varepsilon > 0$ there exists a function $g \in \mathscr{H}$ such that $\|f - g\|_\infty < \varepsilon$. It implies that any kernel-based learning algorithms with universal kernels can in principle approximate any bounded continuous function f arbitrarily well. Note that approximation in the RKHS norm implies the approximation in the sup norm by the continuity of the evaluation functional. It follows from Definition 3.2 and Gretton et al. (2012a; Theorem 8) that all universal kernels are characteristic. Examples of universal kernels on a compact domain are Gaussian and Laplace kernels. On the discrete domain $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, any strictly positive definite kernel, *e.g.*, $k(\mathbf{x}, \mathbf{x}') = \mathbb{1}_{\{\mathbf{x}=\mathbf{x}'\}}$, is universal (Borgwardt et al. 2006; Section 2.3). Overall, universality is a stronger notion than characteristic property, but they match if the kernel is translation invariant, continuous and decays to zero at infinity. Sriperumbudur et al. (2011a) provides a comprehensive insight into the connection between universal and characteristic kernels.

For non-Euclidean spaces, sufficient and necessary conditions for the characteristic RKHSes on groups and semi-groups—*e.g.*, periodic domains, rotation matrices, and $\mathbb{R}_+^d$—are established in Fukumizu et al. (2009b). A universal kernel on non-Euclidean spaces and its relation to characteristic kernels are also discussed in Christmann and Steinwart

(2010). Nishiyama and Fukumizu (2014) connects the translation-invariant kernels on $\mathbb{R}^d$ to the infinitely divisible distributions. That is, the kernel is characteristics if it is generated by the bounded continuous density of a symmetric and infinitely divisible distribution, *e.g.*, α-stable distribution.

In certain applications such as two-sample testing, characteristic kernel is crucial as it ensures that in the population limit we obtain the desired statistics. In practice, we always incur an estimation error due to a finite sample. Moreover, there are many application domains in which the kernel is not necessarily required to be characteristic. For example, predictive learning on distributional data (Muandet et al. 2012, Muandet and Schölkopf 2013, Oliva et al. 2014, Szabó et al. 2015). In these cases, it may be more favourable to interpret kernel k as a *weight function* which determines which frequency component occurs in the embedding (see Example 3.3). A shape of the kernel k in the Fourier domain can therefore be more informative in such applications.

Non-characteristic kernels. In general, if the kernel k is *non*-characteristic, the embedding μ forms an equivalence class of distributions that correspond to the same mean embedding $\mu_{\mathbb{P}}$. Nevertheless, k may be characteristic for a more restricted class of distributions. Consider, for example, any translation-invariant kernel with the corresponding Λ whose support has non-empty interior. Then, we may conclude from (3.20) that k will be "characteristic" for any class of probability measures whose characteristic functions only agree outside the support of Λ. Sriperumbudur et al. (2010; Theorem 12) and Harmeling et al. (2013; Proposition 3) consider a more interesting class, namely, a class of *probability measures with compact support* on $\mathbb{R}^d$. It follows from the Paley-Wiener theorem (Rudin 1991) that the characteristic functions of such measures are entire functions on $\mathbb{C}^d$. As a result, if Λ has support with non-empty interior, the corresponding kernel will be characteristic for probability measures with compact support. Examples of kernels with such property include the sinc kernel $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}') = \frac{\sin \sigma(\mathbf{x}-\mathbf{x}')}{\mathbf{x}-\mathbf{x}'}$ with $\Lambda(\boldsymbol{\omega}) = \sqrt{\frac{\pi}{2}} \mathbb{1}_{[-\sigma,\sigma]}(\boldsymbol{\omega})$.

Table 3.1: Various characterizations of well-known kernel functions. The columns marked 'U', 'C', 'TI', and 'SPD' indicate whether the kernels are universal, characteristic, translation-invariant, and strictly positive definite, respectively, w.r.t. the domain $\mathcal{X}$. For the discrete kernel, $\#_s(\mathbf{x})$ is the number of times substrings s occurs in a string $\mathbf{x}$. K_ν is a modified Bessel function of the second kind of order ν and Γ is the Gamma function.

Kernel Function	$k(\mathbf{x}, \mathbf{y})$	**Domain** $\mathcal{X}$	**U**	**C**	**TI**	**SPD**
Dirac	$\mathbb{1}_{\mathbf{x}=\mathbf{y}}$	$\{1, 2, \ldots, m\}$	✓	✓	✗	✓
Discrete	$\sum_{s\in\mathcal{X}} w_s \#_s(\mathbf{x})\#_s(\mathbf{y})$ with $w_s > 0$ for all s	$\{s_1, s_2, \ldots, s_m\}$	✓	✓	✗	✓
Linear	$\langle \mathbf{x}, \mathbf{y} \rangle$	$\mathbb{R}^d$	✗	✗	✗	✗
Polynomial	$(\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$	$\mathbb{R}^d$	✗	✗	✗	✗
Gaussian	$\exp(-\sigma\|\mathbf{x} - \mathbf{y}\|_2^2),\ \sigma > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
Laplacian	$\exp(-\sigma\|\mathbf{x} - \mathbf{y}\|_1),\ \sigma > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
Rational quadratic	$(\|\mathbf{x} - \mathbf{y}\|_2^2 + c^2)^{-\beta},\ \beta > 0, c > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
B_{2l+1}-splines	$B_{2l+1}(\mathbf{x} - \mathbf{y})$ where $l \in \mathbb{N}$ with $B_i := B_i \otimes B_0$	$[-1, 1]$	✓	✓	✓	✓
Exponential	$\exp(\sigma\langle \mathbf{x}, \mathbf{y} \rangle),\ \sigma > 0$	compact sets of $\mathbb{R}^d$	✗	✓	✗	✓
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right) K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right)$	$\mathbb{R}^d$	✓	✓	✓	✓
Poisson	$1/(1 - 2\alpha\cos(\mathbf{x} - \mathbf{y}) + \alpha^2),\ 0 < \alpha < 1$	$([0, 2\pi), +)$	✓	✓	✓	✓

3.4 Kernel Mean Estimation and Approximation

In practice, the distribution $\mathbb{P}$ is generally unknown, and we must rely entirely on the sample drawn from $\mathbb{P}$. Given an independent and identically distributed (i.i.d.) sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbb{P}$, the standard estimator of the kernel mean is an empirical average

$$\hat{\mu}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \cdot) \,. \tag{3.21}$$

By the weak law of large number, the empirical estimator (3.21) is guaranteed to converge to the true mean embedding. Berlinet and Thomas-Agnan (2004) shows that the convergence happens at a rate $O_p(n^{-1/2})$. This result has also been reported in Smola et al. (2007), Shawe-Taylor and Cristianini (2004; Chapter 4), Song (2008; Theorem 27), Gretton et al. (2012a), and Lopez-Paz et al. (2015; Theorem 1), in slightly different forms.

One may argue that the estimator (3.21) is the "best" possible estimator of $\mu_{\mathbb{P}}$ if nothing is known about the underlying distribution $\mathbb{P}$. In fact, (3.21) is minimax in the sense of van der Vaart (1998; Theorem 25.21, Example 25.24). In similar respect, Lopez-Paz et al. (2015) also gives a lower bound showing that the $n^{-1/2}$ is an optimal rate matching the upper bound of Theorem 3.4. Nevertheless, given that a kernel mean is central to kernel methods in that it is used by many classical algorithms such as kernel principal component analysis (PCA), and it also forms the core inference step of modern kernel methods that rely on embedding probability distributions in RKHSs; it is compelling to ask whether the estimation of $\mu_{\mathbb{P}}$ can be improved.

3.4.1 Kernel Mean Shrinkage Estimators

Inspired by the James-Stein estimator of the mean of multivariate Gaussian distribution (Stein 1955, James and Stein 1961), Muandet et al. (2014a; 2016) proposed a shrinkage estimator of the kernel mean which has the following form

$$\hat{\mu}_{\alpha} := \alpha f^* + (1 - \alpha)\hat{\mu}_{\mathbb{P}}, \tag{3.22}$$

for some $f^* \in \mathscr{H}$ which is independent of the sample. The shrinkage parameter α specifies an amount by which the estimator $\hat{\mu}_{\mathbb{P}}$ is shrunk towards f^*. Note that the works of Muandet et al. (2014a; 2016) differ fundamentally from the Stein's seminal work and those along this line. That is, the setting of Muandet et al. (2016) involves a non-linear feature map ϕ associated with the kernel k. Consequently, the resulting kernel mean $\mu_{\mathbb{P}}$ may incorporate higher moments of $\mathbb{P}$. When k is a linear kernel, $\mu_{\mathbb{P}}$ becomes a mean of $\mathbb{P}$ and this setting coincides with that of Stein. A direct generalization of James-Stein estimator to an infinite dimensional Hilbert space has been considered in Berger and Wolpert (1983), Privault and Réveillac (2008), Mandelbaum and Shepp (1987).[6]

To understand the effect of the shrinkage estimator in (3.22), consider the bias-variance decomposition

$$\mathbb{E}\,\|\hat{\mu}_\alpha - \mu_{\mathbb{P}}\|^2_{\mathscr{H}} = \alpha^2\|f^* - \mu_{\mathbb{P}}\|^2_{\mathscr{H}} + (1-\alpha)^2\mathbb{E}[\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathscr{H}}], \quad (3.23)$$

where the expectation is taken w.r.t. the i.i.d. sample of size n from $\mathbb{P}$. The first term on the r.h.s. of (3.23) represents the squared-bias, whereas the second term represents the variance. Note that for $\alpha \in (0, 2)$, $(1-\alpha)^2 < 1$, which means the variance of $\hat{\mu}_\alpha$ is always smaller than that of $\hat{\mu}_{\mathbb{P}}$ at the expense of the increased bias. Hence, α which controls the bias-variance trade-off can be chosen to be the minimizer of the mean-squared error. However, since this minimizer will depend on the unknown $\mu_{\mathbb{P}}$, Muandet et al. (2016) consider an estimate of α which when plugged in (3.21) yields an estimate of $\mu_{\mathbb{P}}$. Muandet et al. (2016) also propose the positive-part version of (3.22), *i.e.*, $\hat{\mu}_\alpha := \alpha f^* + (1-\alpha)_+\hat{\mu}_{\mathbb{P}}$, which is similar in spirit to the positive-part James-Stein estimator.

It is known that the empirical estimator (3.22) can be considered as an *M-estimator* (Shawe-Taylor and Cristianini 2004, Kim and Scott

[6]The effect of shrinkage in the context of kernel mean embedding has previously been observed in the experimental study of Huszar and Duvenaud (2012) that investigates the connection between kernel Herding and Bayesian quadrature. In Bayesian quadrature, there is an indication that the Bayesian weight obtained by minimizing the posterior variance exhibits shrinkage when the number of sample is small.

2012), *i.e.*, it can be obtained as

$$\hat{\mu}_{\mathbb{P}} = \arg\min_{g \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \|g - k(\mathbf{x}_i, \cdot)\|^2_{\mathscr{H}}. \tag{3.24}$$

The population counterpart $\mu_{\mathbb{P}}$ can be obtained by replacing the sum with expectation. Owing to this interpretation, Muandet et al. (2016) constructs a shrinkage estimator by adding a regularizer $\Omega(\|g\|_{\mathscr{H}}) = \lambda\|g\|^2_{\mathscr{H}}$ to (3.24) whose minimizer is given by $\left(\frac{1}{1+\lambda}\right)\hat{\mu}_{\mathbb{P}}$. As can be seen that this estimator is the same as the one in (3.22) when $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$. Muandet et al. (2016) propose a data-dependent choice for λ using leave-one-out cross-validation, resulting in a shrinkage estimator for $\mu_{\mathbb{P}}$. As an aside, Kim and Scott (2012) also exploit this interpretation to robustify kernel density estimator (KDE) by replacing the squared loss $L(k(\mathbf{x}_i, \cdot), g) = \|g - k(\mathbf{x}_i, \cdot)\|^2_{\mathscr{H}}$ in (3.24) by the loss which is less sensitive to outliers such as Huber loss (Huber 1964) and k is assumed to be non-negative and integrate to one, *e.g.*, Gaussian kernel and Student's-t kernel.

Later, Muandet et al. (2014b) provides a *non-linear* extension of (3.22) by mean of spectral filtering algorithms (Bauer et al. 2007, De Vito et al. 2006). First, one can argue that, by virtue of representer theorem (Schölkopf et al. 2001), any solution g can be written as $g = \sum_{i=1}^{n} \beta_i k(\mathbf{x}_i, \cdot)$ for some $\boldsymbol{\beta} \in \mathbb{R}^d$. Hence, finding g amounts to solving a system of equations $\mathbf{K}\boldsymbol{\beta} = \mathbf{K}\mathbf{1}_n$. Using spectral filtering algorithms, Muandet et al. (2014b) proposes the following estimators:

$$\hat{\mu}_\lambda := \sum_{i=1}^{n} \beta_i k(\mathbf{x}_i, \cdot), \qquad \boldsymbol{\beta} := g_\lambda(\mathbf{K})\mathbf{K}\mathbf{1}_n \tag{3.25}$$

where g_λ is called a filter function such that $g_\lambda(\mathbf{K}) = \mathbf{U}g_\lambda(\mathbf{D})\mathbf{U}^\top$. There exist efficient algorithms, *e.g.*, Landweber iteration and iterated Tikhonov, for solving (3.25) which do not require an explicit computation of eigendecomposition of $\mathbf{K}$. See, *e.g.*, Bauer et al. (2007), De Vito et al. (2006), Muandet et al. (2014b) for detail of each algorithm.

Compared to (3.22), it was shown that the estimators of Muandet et al. (2014b) performs shrinkage by first projecting data onto

Table 3.2: The iterative update for $\boldsymbol{\beta}$ and associated filter function (see, *e.g.*, De Vito et al. (2006), Muandet et al. (2014b) for further detail). Here we define $\mathbf{z} := \mathbf{K}\mathbf{1}_n - \mathbf{K}\boldsymbol{\beta}^{t-1}$.

Algorithm	Iterative Update	Filter Function
L2 Boosting	$\boldsymbol{\beta}^t \leftarrow \boldsymbol{\beta}^{t-1} + \eta\mathbf{z}$	$g(\gamma) = \eta\sum_{i=1}^{t-1}(1-\eta\gamma)^i$
Acc. L2 Boosting	$\boldsymbol{\beta}^t \leftarrow \boldsymbol{\beta}^{t-1} + \omega_t(\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^{t-2}) + \frac{\kappa_t}{n}\mathbf{z}$	$g(\gamma) = p_t(\gamma)$
Iterated Tikhonov	$(\mathbf{K} + n\lambda\mathbf{I})\boldsymbol{\beta}_i = \mathbf{1}_n + n\lambda\boldsymbol{\beta}_{i-1}$	$g(\gamma) = \frac{(\gamma+\lambda)^t - \gamma^t}{\lambda(\gamma+\lambda)^t}$
Truncated SVD	None	$g(\gamma) = \gamma^{-1}\mathbb{1}_{\gamma\geq\lambda}$

the KPCA basis, and then shrinking each component independently according to a pre-defined shrinkage rule, which is specified by the filter function g_λ (Muandet et al. 2014b; Proposition 3). The shrinkage estimate is then reconstructed as a superposition of the resulting components. Unlike (3.22), the spectral shrinkage estimators (3.25) also incorporate the eigenspectrum of the kernel $\mathbf{K}$ into account. The idea has been extended to estimating covariance operator (Muandet et al. 2016, Wehbe and Ramdas 2015) which is ubiquitous in kernel independence and conditional independence tests (cf. §3.6 and §4.6).

3.4.2 Approximating the Kernel Mean

In many applications of kernel methods such as in genomics, astronomy, and social science, the computational cost may be a critical issue, especially in the era of "big data". Traditional kernel-based algorithms become computationally prohibitive as the volume of data has exploded because most existing algorithms scale at least quadratically with sample size. Likewise, the use of kernel mean embedding has also suffered from this limitation due to two fundamental issues. First, any estimators of the kernel mean involve the (weighted) sum of the feature map of the sample. Second, for certain kernel functions such as Gaussian RBF kernel, the kernel mean lives in an infinite dimensional space. We can categorize previous attempts in approximating the kernel mean into two basic approaches: 1) find a smaller subset of samples whose estimate approximate well the original estimate of the kernel mean, 2) find a finite approximation of the kernel mean directly.

The former has been studied extensively in the literature. For example, Cortes and Scott (2014) considers the problem of approximating the kernel mean as a sparse linear combination of the sample. The proposed algorithm relies on a subset selection problem using novel incoherence measure. The algorithm can be solved efficiently as an instance of the *k-center problem* and has linear complexity in the sample size. Similarly, Grünewälder et al. (2012) proposes a sparse approximation of the conditional mean embedding by relying on an interpretation of the conditional mean as a regressor. Note that the same idea can be adopted to find a sparse approximation of the standard kernel mean by imposing the sparsity-inducing norm on the coefficient $\boldsymbol{\beta}$, *e.g.*, $\|\boldsymbol{\beta}\|_1$ (Muandet et al. 2014a). An advantage of sparse representation is in applications where the kernel mean is evaluated repeatedly, *e.g.*, Kalman filter (Kanagawa et al. 2013, McCalman et al. 2013). The crucial drawback is that it requires solving an optimization to find an optimal subsample, which may not be trivial optimization problems.

An alternative approach to kernel mean approximation is to find a finite representation of the kernel mean directly. One of the most effective approaches depends on the *random feature map* (Rahimi and Recht 2007). That is, instead of relying on the implicit feature map provided by the kernel, the basic idea of random feature approximation is to explicitly map the data to a low-dimensional Euclidean inner product space using a randomized feature map $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathscr{H}} \approx \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}) \tag{3.26}$$

where $\mathbf{z}(\mathbf{x}) := \mathbf{W}^\top \mathbf{x}$ and $w_{ij} \sim p(\mathbf{w})$. If elements of $\mathbf{W}$ are drawn from appropriate distribution $p(\mathbf{w})$, the Johnson-Lindenstrauss Lemma (Dasgupta and Gupta 2003, Blum 2005) ensures that this transformation will preserve similarity between data points. In Rahimi and Recht (2007), $p(\mathbf{w})$ is chosen to be the Fourier transform of translation-invariant kernels $k(\mathbf{x} - \mathbf{y})$. Given a feature map $\mathbf{z}$, the finite approximation of the kernel mean can be obtained directly as

$$\tilde{\mu}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i) \in \mathbb{R}^m. \tag{3.27}$$

Since $\mathbf{z}(\mathbf{x}_i) \in \mathbb{R}^m$ for all i, so does $\tilde{\mu}_{\mathbb{P}}$. Hence, there is no need to store all the vector $\mathbf{z}(\mathbf{x}_i)$. In addition to giving us a compact representation of kernel mean, these randomized feature maps also accelerate the evaluation of the algorithms that use kernel mean embedding (see, *e.g.*, Kar and Karnick (2012), Le et al. (2013), Pham and Pagh (2013) and references therein for extensions). Note that the approximation (3.27) is so general that it can be obtained as soon as one know how to compute $\mathbf{z}(\mathbf{x})$. Other approaches such as low-rank approximation are also applicable. As we can see, the advantage of this approach is that given any finite approximation of $\phi(\mathbf{x})$, it is easy to approximate the kernel mean. Moreover, the resulting approximation has been shown to enjoy good empirical performance. The downside of this approach is that as the approximation lives in the finite dimensional space, theoretical guarantee relating this approximation back to the infinite-dimensional counterpart may be difficult to obtain. Preliminary result is given in Lopez-Paz et al. (2015; Lemma 1). Also, the random features are limited to only a certain class of kernel functions.

3.5 Maximum Mean Discrepancy

The kernel mean embedding can be used to define a metric for probability distributions which is important for problems in statistics and machine learning. Later, we will see that the metric defined in terms of mean embeddings can be considered as a particular instance of an *integral probability metric* (IPM) (Müller 1997). Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $\mathcal{X}$, an IPM is defined as

$$\gamma[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{F}} \left\{ \int f(\mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) - \int f(\mathbf{y}) \, \mathrm{d}\mathbb{Q}(\mathbf{y}) \right\} \tag{3.28}$$

where $\mathcal{F}$ is a space of real-valued bounded measurable functions on $\mathcal{X}$. The IPM are fully characterized by the function class $\mathcal{F}$. There is obviously a trade-off on the choice of $\mathcal{F}$. That is, on one hand, the function class must be rich enough so that $\gamma[\mathcal{F}, \mathbb{P}, \mathbb{Q}]$ vanishes if and only if $\mathbb{P} = \mathbb{Q}$. On the other hand, the larger the function class $\mathcal{F}$, the more difficult it is to estimate $\gamma[\mathcal{F}, \mathbb{P}, \mathbb{Q}]$. Thus, $\mathcal{F}$ should be restrictive enough for the empirical estimate to converge quickly (see, *e.g.*,

Sriperumbudur et al. (2012)).

For example, if $\mathcal{F}$ is chosen to be a space of all bounded continuous functions on $\mathcal{X}$, the IPM is a metric over a space of probability distributions, as stated in the following theorem (Müller 1997).

Theorem 3.5. $\gamma[C_b(\mathcal{X}), \mathbb{P}, \mathbb{Q}] = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Unfortunately, it is practically difficult to work with $C_b(\mathcal{X})$. A more restrictive function class is often used. For instance, let $\mathcal{F}_{\text{TV}} = \{f \,|\, \|f\|_\infty \leq 1\}$ where $\|f\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$. Then, $\gamma[\mathcal{F}_{\text{TV}}, \mathbb{P}, \mathbb{Q}] = \|\mathbb{P} - \mathbb{Q}\|_1$ is the *total variation distance*. If $\mathcal{F} = \{\mathbf{1}_{(\infty,t]}\}$, we get the *Kolmogorov (or L^∞) distance* between distributions, which is the max norm of the difference between their cumulative distributions. If $\|f\|_L := \sup\{|f(\mathbf{x}) - f(\mathbf{y})|/\,\mathrm{d}(\mathbf{x}, \mathbf{y}), \mathbf{x} \neq \mathbf{y} \in \mathcal{X}\}$ is the Lipschitz seminorm of a real-valued function f, setting $\mathcal{F} = \{f \,|\, \|f\|_L \leq 1\}$ yields the *earthmover distance*. In mathematics, this metric is known as *Wasserstein (or L^1) distance*.

The maximum mean discrepancy (MMD) considers functions in the unit ball of RKHS, *i.e.*, $\mathcal{F} := \{f \,|\, \|f\|_{\mathscr{H}} \leq 1\}$. In which case, the MMD can be expressed as the distance in $\mathscr{H}$ between mean embeddings as shown in Borgwardt et al. (2006), Gretton et al. (2012a; Lemma 4). That is,

$$\begin{aligned}
\text{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] &= \sup_{\|f\|\leq 1} \left\{ \int f(\mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}) - \int f(\mathbf{y})\, \mathrm{d}\mathbb{Q}(\mathbf{y}) \right\} \\
&= \sup_{\|f\|\leq 1} \left\{ \langle f, \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}) \rangle - \langle f, \int k(\mathbf{y}, \cdot)\, \mathrm{d}\mathbb{Q}(\mathbf{y}) \rangle \right\} \\
&= \sup_{\|f\|\leq 1} \{ \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle \} \\
&= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} \qquad (3.29)
\end{aligned}$$

where we use the reproducing property of $\mathscr{H}$ and the linearity of the inner product, respectively. Thus, we can express the MMD in terms of the associated kernel function k as

$$\text{MMD}^2[\mathscr{H}, \mathbb{P}, \mathbb{Q}] = \mathbb{E}_{X,\tilde{X}}[k(X, \tilde{X})] - 2\mathbb{E}_{X,Y}[k(X, Y)] + \mathbb{E}_{Y,\tilde{Y}}[k(Y, \tilde{Y})] \qquad (3.30)$$

where $X, \tilde{X} \sim \mathbb{P}$ and $Y, \tilde{Y} \sim \mathbb{Q}$ are independent copies. Note that in deriving (3.30), we use

$$\|\mu_{\mathbb{P}}\|^2_{\mathscr{H}} = \langle \mathbb{E}[k(\cdot, X)], \mathbb{E}[k(\cdot, \tilde{X})] = \mathbb{E}[k(X, \tilde{X}].$$

It follows that $\mathrm{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] = 0$ if and only if $\mathscr{H}$ is characteristic. Moreover, when k is translation invariant, *i.e.*, $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$, we have $\mathrm{MMD}(\mathscr{H}, \mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\boldsymbol{\omega}) - \varphi_{\mathbb{Q}}(\boldsymbol{\omega})|^2 \mathfrak{F}^{-1}\psi(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}$ where $\mathfrak{F}^{-1}$ denotes the inverse Fourier transform and $\varphi_{\mathbb{P}}, \varphi_{\mathbb{Q}}$ are characteristic functions of $\mathbb{P}, \mathbb{Q}$, respectively, (Sriperumbudur et al. 2010; Corollary 4). In other words, the MMD can be interpreted as the distance in $L^2(\nu)$ between $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$ where ν is the inverse Fourier transform of the kernel.

Given i.i.d. samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m, \}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ from $\mathbb{P}$ and $\mathbb{Q}$, respectively, a biased empirical estimate of MMD can be obtained as

$$\widehat{\mathrm{MMD}}^2_b[\mathscr{H}, \mathbf{X}, \mathbf{Y}] := \sup_{\|f\|_{\mathscr{H}} \leq 1} \left(\frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}_i) - \frac{1}{n} \sum_{j=1}^{n} f(\mathbf{y}_j) \right). \quad (3.31)$$

The empirical MMD can be expressed in terms of empirical mean embeddings as $\widehat{\mathrm{MMD}}_b[\mathscr{F}, X, Y] = \|\hat{\mu}_{\mathbf{X}} - \hat{\mu}_{\mathbf{Y}}\|^2_{\mathscr{H}}$. Moreover, we can write an unbiased estimate of the MMD entirely in terms of k as

$$\begin{aligned} \widehat{\mathrm{MMD}}^2_u[\mathscr{H}, \mathbf{X}, \mathbf{Y}] \;=\; & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathbf{x}_i, \mathbf{x}_j) \\ & + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(\mathbf{y}_i, \mathbf{y}_j) \\ & - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{y}_j). \end{aligned} \quad (3.32)$$

Note that (3.32) is an unbiased estimate which is a sum of two U-statistics and a sample average (Serfling 1981; Chapter 5). That is, assuming that $m = n$,

$$\widehat{\mathrm{MMD}}^2_u[\mathscr{H}, \mathbf{X}, \mathbf{Y}] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h(\mathbf{v}_i, \mathbf{v}_j), \quad (3.33)$$

where $h(\mathbf{v}_i, \mathbf{v}_j) := k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{y}_i, \mathbf{y}_j) - k(\mathbf{x}_i, \mathbf{y}_j) - k(\mathbf{x}_j, \mathbf{y}_i)$. We assume throughout that $\mathbb{E}[h^2] < \infty$. The biased counterpart $\widehat{\mathrm{MMD}}_b^2[\mathscr{H}, \mathbf{X}, \mathbf{Y}]$ can be obtained using V-statistics. The convergence of empirical MMD has been established in Gretton et al. (2012a; Theorem 7). Theorem 3.5 below describes an unbiased quadratic-time estimate of the MMD, and its asymptotic distribution under an alternative hypothesis that $\mathbb{P} \neq \mathbb{Q}$.

Theorem 3.6 (Gretton et al. (2012a; Lemma 6; Corollary 16))**.** Given i.i.d. samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m, \}$ and $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ from $\mathbb{P}$ and $\mathbb{Q}$, respectively. When $\mathbb{P} \neq \mathbb{Q}$, an unbiased empirical estimate $\widehat{\mathrm{MMD}}_u^2[\mathscr{H}, \mathbf{X}, \mathbf{Y}]$ given in (3.33) converges in distribution to a Gaussian distribution

$$\sqrt{m}(\widehat{\mathrm{MMD}}_u^2[\mathscr{H}, \mathbf{X}, \mathbf{Y}] - \mathrm{MMD}^2[\mathscr{H}, \mathbb{P}, \mathbb{Q}]) \xrightarrow{p} \mathcal{N}(0, \sigma_{XY}^2)$$

where

$$\sigma_{XY}^2 = 4(\mathbb{E}_{\mathbf{v}_1}[(\mathbb{E}_{\mathbf{v}_2} h(\mathbf{v}_1, \mathbf{v}_2))^2] - [\mathbb{E}_{\mathbf{v}_1, \mathbf{v}_2} h(\mathbf{v}_1, \mathbf{v}_2)]^2)$$

uniformly at rate $1/\sqrt{m}$.

Sriperumbudur et al. (2012) studied the convergence rate and upper bound of the empirical estimators of the IPM for different $\mathcal{F}$ including the MMD. They showed that the MMD enjoys fast convergence, *i.e.*, in the order of $1/\sqrt{n}$ and the rate is independent of the dimensionality d, whereas other metrics such as total variation and Wasserstein suffer from the slow rate that depends on d.

A natural application of the MMD is *two-sample testing*: a statistical hypothesis test for equality of two samples. In particular, we test the *null hypothesis* $H_0 : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} = 0$ against the *alternative hypothesis* $H_1 : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} \neq 0$. However, even if the two samples are drawn from the same distribution, the MMD criterion may still be non-zero due to the finite sample. Gretton et al. (2012a) proposes two distribution-free tests based on large deviation bounds (using Rademacher complexity and bound on U-statistics of Hoeffding (1948)) and the third one based on the asymptotic distribution of the test statistics. The tests based on

large deviation bounds are generally more conservative than the latter as they does not characterize the distribution of MMD explicitly. The MMD-based test can be viewed as a generalization of Kolmogorov-Smirnov test to the multivariate case (Gretton et al. 2012a).

The MMD test has several advantages over existing methods proposed in the literatures (Anderson et al. 1994, Biau and Györfi 2005, Nguyen et al. 2007). First, the MMD test is distribution free.[7] The assumption on the parametric form of the underlying distribution is not needed. Furthermore, like most of kernel-based tests, *e.g.*, Harchaoui et al. (2007), the test can be applied in structured domains like graphs and documents as soon as the positive definite kernel is well-defined. Moreover, an availability of the asymptotic distribution of the test statistics allows for an efficient computation without resorting to costly bootstrapping.

3.5.1 Scaling up the MMD

The MMD can be computed in quadratic time $O(n^2 d)$, which might prohibit its applications in large-scale problems. In Gretton et al. (2012a), the authors also propose the linear time statistics and test by using the subsampling of the term in (3.32), *i.e.*, drawing pairs from $\mathbf{X}$ and $\mathbf{Y}$ without replacement. This method reduces the time complexity of MMD from $O(n^2 d)$ to $O(nd)$. However, the test has high variance due to loss of information. The B-test of Zaremba et al. (2013) tradeoffs the computation and variance of the test by splitting two-sample sets into corresponding subsets and then compute the exact MMD in each block while ignoring between-block interactions with $O(n^{3/2} d)$ time complexity.[8]

Recall that when $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$ for some positive definite function ψ on $\mathbb{R}^d$, we have

$$\mathrm{MMD}(\mathscr{H}, \mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} |\varphi_{\mathbb{P}}(\boldsymbol{\omega}) - \varphi_{\mathbb{Q}}(\boldsymbol{\omega})|^2 \mathfrak{F}^{-1}\psi(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega} \tag{3.34}$$

[7] Note that even if a test is consistent, it is not possible to distinguish distributions with high probability at a given, fixed sample size.

[8] The B-test can be understood as a specific case of the tests using *incomplete* U-statistic (Blom 1976). While this kind of statistic can be obtained in numerous ways, it has not been explored much in the context of MMD.

where $\mathfrak{F}^{-1}$ denotes the inverse Fourier transform and $\varphi_{\mathbb{P}}, \varphi_{\mathbb{Q}}$ are characteristic functions of $\mathbb{P}, \mathbb{Q}$, respectively. Based on (3.34), Ji Zhao (2015) proposes an efficient test called *FastMMD* which employs the random Fourier feature to transform the MMD test with translation invariant kernel. In this case, the empirical MMD becomes the $L_2(\mathfrak{F}^{-1}\psi)$ distance between the empirical characteristic functions. Later, Chwialkowski et al. (2015) demonstrates that the original formulation fails to distinguish a large class of measures, and presents a "smoothed" version of the formulation using an analytic smoothing kernel. The resulting metric is shown to be "random metric" which satisfies all the conditions for a metric with qualification "almost surely". Interestingly, the proposed linear-time test can outperform the quadratic-time MMD in terms of power of the test. The time complexity also reduces to $O(nd)$. For kernels whose spectral distributions $\Lambda(\boldsymbol{\omega})$ are spherically invariant, *i.e.*, $\Lambda(\boldsymbol{\omega})$ only depends on $\|\boldsymbol{\omega}\|_2$, the cost reduces further to $O(n \log d)$ by using the Fastfood technique (Le et al. 2013). The disadvantage is that it is restricted to only translation invariant kernels. Another popular approach to reducing the cost of evaluating the empirical MMD estimate is by using a low-rank approximation of the Gram matrix.

As pointed out by some of the previous works, we may pose the problem of distribution comparison as a binary classification (see, *e.g.*, Gretton et al. (2012a; Remark 20) and Sriperumbudur et al. (2009)). That is, any classifiers for which uniform convergence bounds can be obtained such as neural network, support vector machine, and boosting, can be used for the purpose of distribution comparison. The benefit of this interpretation is that there is a clear definition of loss function which can be used for the purpose of parameter selection. A slightly different interpretation is to look at this problem as a learning problem on probability distributions (Muandet et al. 2012, Muandet and Schölkopf 2013, Szabó et al. 2015). For example, the goal of many hypothesis testing problems is to learn a function from an empirical distribution $\hat{\mathbb{P}}$ to $\{0, 1\}$ which, for example, indicates whether or not to reject the null hypothesis. If the training examples $(\hat{\mathbb{P}}_1, y_1), \ldots, (\hat{\mathbb{P}}_n, y_n)$ are available, we can consider a hypothesis testing

as a machine learning problem on distributions.

Lastly, a commonly used kernel for MMD test on $\mathbb{R}^d$ is the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2)$ whose bandwidth parameter is chosen via the *median heuristic*: $\sigma^2 = \text{median}\{\|\mathbf{x}_i - \mathbf{x}_j\|^2 : i, j = 1, \ldots, n\}$ (Gretton et al. 2005b). While this heuristic has been shown to work well in many applications, it may run into trouble when the sample size is small. In fact, it has been observed empirically that the median heuristic may not work well when estimating the kernel mean from the small sample and there is room for improvement, especially in the high-dimensional setting (Danafar et al. 2013, Muandet et al. 2014a;b, Reddi et al. 2015). An alternative is to choose the kernel that maximizes the test statistic, which is found to outperform the median heuristic empirically (Sriperumbudur et al. 2009). Gretton et al. (2012b) proposes a criterion to choose a kernel for two-sample testing using MMD. The kernel is chosen so as to maximize the test power, and minimize the probability of making a Type II error. The proposed method corresponds to maximizing the Hodges and Lehmann asymptotic relative efficiency (Hodges and Lehmann 1956). Despite these efforts, how to choose a good kernel function on its own remains an open question.

The MMD has been applied extensively in many applications, namely, clustering (Jegelka et al. 2009), density estimation (Song et al. 2007a; 2008), (conditional) independence tests (Fukumizu et al. 2008, Doran et al. 2014, Chwialkowski and Gretton 2014), causal discovery (Sgouritsa et al. 2013, Chen et al. 2014, Schölkopf et al. 2015), covariate shift (Gretton et al. 2009, Pan et al. 2011) and domain adaptation (Blanchard et al. 2011, Muandet et al. 2013), selection bias correction (Huang et al. 2007), herding (Chen et al. 2010, Huszar and Duvenaud 2012), Markov chain Monte Carlo (Sejdinovic et al. 2014), moment matching for training deep generative models (Li et al. 2015, Dziugaite et al. 2015), statistical model criticism (Lloyd and Ghahramani 2015), approximate Bayesian computation (Park et al. 2016), and model selection in generative models (Bounliphone et al. 2016), for example.

3.6 Kernel Dependency Measures

A dependence measure is one of the most fundamental tools in statistical analysis. Rényi (1959) outlines a list of seven desirable properties—known as *Rényi's axioms*—which should be satisfied by any measure of dependence between two random variables. The classical criteria such as Spearman's rho and Kendall's tau can detect only linear dependencies. Other non-linear criteria such as mutual information based measures and density ratio methods, requires certain assumptions regarding a parametric form of the underlying distribution which may be too restrictive in several applications.

When the dependence is non-linear, one of the most successful non-parametric measures is the *Hilbert Schmidt Independence Criterion (HSIC)* (Gretton et al. 2005a). Let $\mathscr{H}$ and $\mathscr{F}$ be separable RKHSs on X and Y with reproducing kernel k and l, respectively. The HSIC is defined as the square HS-norm of the associated cross-covariance operator:

$$\begin{aligned}\mathrm{HSIC}(\mathscr{H},\mathscr{F},k,l) &= \|\mathcal{C}_{XY}\|^2_{\mathrm{HS}} \\ &= \mathbb{E}_{\mathbf{x}\mathbf{x}'\mathbf{y}\mathbf{y}'}[k(\mathbf{x},\mathbf{x}')l(\mathbf{y},\mathbf{y}')] \\ &\quad +\mathbb{E}_{\mathbf{x}\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')] \\ &\quad -2\mathbb{E}_{\mathbf{x}\mathbf{y}}[\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')]] \quad (3.35)\end{aligned}$$

If the product kernel $k(\cdot,\cdot)\times l(\cdot,\cdot)$ is characteristic on $\mathcal{X}\times\mathcal{Y}$, it is not difficult to show that $\mathcal{C}_{XY}=\mathbf{0}$ and $\mathrm{HSIC}(X,Y,k,l)=0$ if and only if $X \perp\!\!\!\perp Y$. To see this, by definition we have $\mathcal{C}_{XY}=\mu_{\mathbb{P}_{YX}}-\mu_{\mathbb{P}_Y\otimes\mathbb{P}_X}$ (cf. §3.2). Hence, $\mathcal{C}_{XY}=\mathbf{0}$ implies $\mu_{\mathbb{P}_{YX}}=\mu_{\mathbb{P}_Y\otimes\mathbb{P}_X}$. Since the product kernel is characteristic, it follows that $\mathbb{P}_{YX}=\mathbb{P}_Y\otimes\mathbb{P}_X$. As an analogy, when X and Y are jointly Gaussian in $\mathbb{R}^d$, $V_{XY}=\mathbf{0}$ if and only if X and Y are independent where V_{XY} denotes the cross-covariance matrix.

An empirical unbiased estimate of HSIC statistic from an i.i.d. sample $(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)$ on $\mathcal{X}\times\mathcal{Y}$ is given by

$$\begin{aligned}\widehat{\mathrm{HSIC}}(X,Y,k,l) &= \|\widehat{\mathcal{C}}_{XY}\|^2_{\mathrm{HS}} \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i,\mathbf{x}_j)l(\mathbf{y}_i,\mathbf{y}_j)\end{aligned}$$

$$-\frac{2}{n^3}\sum_{i,j,p}^{n} k(\mathbf{x}_i,\mathbf{x}_j)l(\mathbf{y}_i,\mathbf{y}_p)$$
$$+\frac{1}{n^4}\sum_{i,j=1}^{n} k(\mathbf{x}_i,\mathbf{x}_j)\sum_{p,q=1}^{n} l(\mathbf{y}_p,\mathbf{y}_q). \quad (3.36)$$

or equivalently $\widehat{\text{HSIC}}(X,Y,k,l) = (1/n^2)\text{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}})$ where $\tilde{\mathbf{K}} = \mathbf{HKH}$, $\tilde{\mathbf{L}} = \mathbf{HLH}$, and $\mathbf{H} = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^\top$. The empirical estimate (3.36) is equivalent to the quadratic dependence measure of Achard et al. (2003), but the test of Achard et al. (2003) is not necessarily zero if and only if the random variables are independent (see Gretton et al. (2005a; Appendix B) for the proof of equivalence and discussion). It can be computed in $O(n^2)$ time and converges to the population HSIC at a rate of $1/\sqrt{n}$ (a slight improvement was later given in Song et al. (2007b) for an unbiased estimator of HSIC). Relying on the large deviation bound, Gretton et al. (2005a) defines the independence test as the indicator that HSIC is larger than a term in the form $C\sqrt{\log(1/\alpha)/n}$ where α is a significance level of a test and C is a suitable constant.

Due to its generality, HSIC has proven successful in both statistics and machine learning. Many learning problems can be interpreted as a maximization or minimization of the dependence between X and Y. For instance, Song et al. (2007b) considers a supervised feature selection as maximizing a dependence between the subsets of covariate X', obtained from backward elimination algorithm (Guyon and Elisseeff 2003), and the target Y, *e.g.*, outputs of binary, multi-class, and regression settings. Likewise, when Y denotes the cluster labels, Song et al. (2007a) proposes a HSIC-based clustering algorithm called CLUHSIC that clusters the data by maximizing the HSIC between X and Y. Interestingly, depending on the structure of the output kernel l, many classical clustering algorithms, *e.g.*, k-means, weighted k-mean, and hierarchical clustering, can be considered as a special case of CLUHSIC. Moreover, the MaxMMD of Jegelka et al. (2009) is in fact equivalent to dependence maximization framework. This concept has been extended to several applications including, for example, kernelized sorting (Quadrianto et al. 2009).

Alternatively, we may pose the independence test between X and

Y as a two-sample testing problem. Recall that X and Y are said to be independent if and only if their joint distribution factorizes as $\mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$. Let $\mu_{\mathbb{P}_{XY}}$ and $\mu_{\mathbb{P}_X \otimes \mathbb{P}_Y}$ be the kernel mean embedding of $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$, respectively. The MMD test statistic for testing independence can then be written as

$$\widehat{\mathrm{MMD}}(\mathscr{H} \otimes \mathscr{F}, \mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y) = \|\hat{\mu}_{\mathbb{P}_{XY}} - \hat{\mu}_{\mathbb{P}_X \otimes \mathbb{P}_Y}\|^2_{\mathscr{H} \otimes \mathscr{F}}, \quad (3.37)$$

where $\hat{\mu}_{\mathbb{P}_{XY}}$ and $\hat{\mu}_{\mathbb{P}_X \otimes \mathbb{P}_Y}$ denote the empirical estimates of $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$, respectively. Since we only have access to sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ from $\mathbb{P}_{XY}$, the corresponding sample from $\mathbb{P}_X \otimes \mathbb{P}_Y$ can be obtained approximately as $\{(\mathbf{x}_i, \mathbf{y}_{\pi(i)})\}_{i=1}^n$ where $\pi(\cdot)$ is a random permutation such that $\pi(i) \neq i$.[9] In light of (3.37), Doran et al. (2014) constructs a conditional independence test, relying on the learned permutation $\pi(\cdot)$ that additionally preserves the similarity on conditioning variable Z (see §4.6).

Several extensions of HSIC has been proposed for settings in which there are three random variables (X, Y, Z)—*e.g.*, conditional dependence measure, three-variable interaction, and relative dependency measure. For instance, Sejdinovic et al. (2013a) proposes kernel nonparametric test to detect Lancaster three-variable interaction, *e.g.*, V-structure. Bounliphone et al. (2015) constructs a consistent test for relative dependency which—unlike the Lancaster test of Sejdinovic et al. (2013a)—measures dependency between a source variable and two candidate target variables. Taking into account the correlation between two HSIC statistics when deriving the corresponding *joint* asymptotic distribution leads to more powerful, consistent test than the test based on two independent HSIC statistics (Bounliphone et al. 2015; Theorem 4)). Lastly, a nonparametric dependency test for arbitrary number of random variables has also been considered in the literature and is required in many applications. However, the prominent issue is that, as the number of random variables grows, the convergence of the estimators may be arbitrarily slow. We will postpone the discussion of

[9] As noted in Janzing et al. (2013) and Doran et al. (2014), this random permutation is only an approximation: while it removes the dependence between X and its corresponding Y, it introduces a dependence to one of the other Y variables, which becomes negligible in the limit $n \to \infty$.

conditional dependence measure to §4.6.

3.6.1 Extensions to Non-i.i.d. Random Variables

The aforementioned dependence measures only work when the data are i.i.d. For non-i.i.d. data, assumptions on the form of dependency—oftens in term of graphical models or *mixing* conditions—are needed. For example, to cope with non-i.i.d. data, Zhang et al. (2008) proposes a *Structured-HSIC* where $\mathbb{P}$ satisfies the conditional independence specified by an undirected graphical models, *e.g.*, a Markovian dependence for sequence data and grid structured data. Thus, the kernel mean embedding decomposes along the maximal cliques of the graphical model. The Hammersley-Clifford theorem (Hammersley and Clifford 1971) ensures the full support of $\mathbb{P}$ and hence the injectivity of the embedding. Zhang et al. (2008) applied the proposed measure to independent component analysis (ICA) and time series clustering and segmentation.

Kernel independence tests for time series—*e.g.*, financial data and brain activity data—were recently proposed in a series of works including Besserve et al. (2013), Chwialkowski and Gretton (2014), Chwialkowski et al. (2014). If one is interested in a serial dependency within a single time series, the test reduces to i.i.d. case under null hypothesis. In order to test dependence between one time series and another, Besserve et al. (2013) characterizes dependencies between time-series as the Hilbert-Schmidt norm of the kernel cross-spectral density (KCSD) operator which is the Fourier transform of the covariance operator at each time lag. In contrast, Chwialkowski and Gretton (2014) uses standard HSIC test statistic whose null estimate is obtained by making shift of one signal relative to the other, rather than the ordinary bootstrapping. The authors impose mixing conditions on random processes under investigation (see Chwialkowski and Gretton (2014) for detail).

3.7 Learning on Distributional Data

In many machine learning applications, it may be more preferable to represent training data as probability measures $\mathbb{P}$ rather than just

points x in some input space $\mathcal{X}$. For instance, many classical problems such as multiple-instance learning (Doran 2013), learning from noisy data, and group anomaly detection (Póczos et al. 2011, Xiong et al. 2011b;a, Muandet and Schölkopf 2013), may be viewed as empirical risk minimization when training data are probability distributions (Muandet et al. 2012). More recently, modern applications of learning on distributions have also proven successful in statistical estimation and causal inference (Póczos et al. 2013, Oliva et al. 2014, Szabó et al. 2015, Lopez-Paz et al. 2015). Furthermore, in the big data era, the benefits of representing data by distributions are two-fold: it reduces the effective sample size of the learning problems and also help conceal the identity of individual samples, which is an important milestone in privacy-preserving learning (Dwork 2008).

3.7.1 Kernels on Probability Distributions

Given distributions $\mathbb{P}_1, \mathbb{P}_2 \ldots, \mathbb{P}_n$, the basic idea is to treat the embedding $\mu_{\mathbb{P}_1}, \mu_{\mathbb{P}_2}, \ldots, \mu_{\mathbb{P}_n}$ as their feature representation on which learning algorithms operate. For example, one of the most popular approaches is to define a positive definite kernel on probability distributions by

$$\kappa(\mathbb{P}, \mathbb{Q}) := \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathscr{H}} = \iint_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \,. \tag{3.38}$$

The empirical counterpart of (3.38) is obtained by replacing both integrals with the finite sums over the i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \sim \mathbb{P}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m \sim \mathbb{Q}$, *i.e.*,

$$\kappa(\mathbb{P}, \mathbb{Q}) \approx \langle \hat{\mu}_{\mathbb{P}}, \hat{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} = \frac{1}{m^2} \sum_{i,j=1}^{m} k(\mathbf{x}_i, \mathbf{y}_j) \,. \tag{3.39}$$

For some probability distributions and kernels k, the kernel (3.38) can be evaluated analytically, see, *e.g.*, Song et al. (2008; Table 1) and Muandet et al. (2012; Table 1) which we reproduce here in Table 3.3. Since $\hat{\mu}_{\mathbb{P}}$ can be infinite dimensional, another popular approach—especially in large-scale learning—is to find a finite approximation of $\hat{\mu}_{\mathbb{P}}$ directly (see, *e.g.*, §3.4.2). The advantages of this approach are that learning is often more efficient and any off-the-shelf learning algorithm can be employed.

The benefits of kernel mean representation for distributional data are three-fold. Firstly, the representation can be estimated consistently without any parametric assumption on the underlying distributions, unlike those based on generative probability models (Jaakkola and Haussler 1998, Jebara et al. 2004). Secondly, compared to density estimation approaches (Póczos et al. 2011, Oliva et al. 2014; 2013), the kernel mean representation is less prone to the *curse of dimensionality* and admits fast convergence in terms of sample size (Shawe-Taylor and Cristianini 2004, Smola et al. 2007, Grünewälder et al. 2012). Lastly, being an element of the kernel feature space enables one to extend the whole arsenal of kernel methods to distributional data and to adopt existing tools for theoretical analysis, *e.g.*, generalization bounds (Szabó et al. 2015, Lopez-Paz et al. 2015, Muandet 2015).

3.7.2 Properties of Distributional Kernels

Note that the map $\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}$ is linear w.r.t. $\mathbb{P}$. Despite being non-linear in the input space $\mathcal{X}$, the function class induced by (3.38) is therefore comprised of only linear functions over the probability space M_+^1. Hence, it cannot be a universal kernel in the sense of Steinwart (2002b) for probability distributions. Let $\mathcal{F} := \{\mathbb{P} \mapsto \int_{\mathcal{X}} g \, \mathrm{d}\mathbb{P} : \mathbb{P} \in M_+^1, g \in C_b(\mathcal{X})\}$ where $C_b(\mathcal{X})$ denotes a class of bounded continuous functions on $\mathcal{X}$. It follows that $C_b(\mathcal{X}) \subset \mathcal{F} \subset C_b(M_+^1)$ where $C_b(M_+^1)$ is a class of bounded continuous functions on $M_+^1(\mathcal{X})$. It was shown in Muandet et al. (2012; Lemma 2) that a function space induced by the kernel κ is dense in $\mathcal{F}$ if k is universal (see, *e.g.*, Definition 4 in Steinwart (2002b)).

Lemma 3.7 (Muandet et al. (2012))**.** Assuming that $\mathcal{X}$ is compact, the RKHS $\mathscr{H}$ induced by the kernel k is dense in $\mathcal{F}$ if k is universal in the sense of Steinwart (2002b).

A more challenging question to answer is whether there exists the non-linear kernel κ on M_+^1 which is dense in $C_b(M_+^1)$. For example,

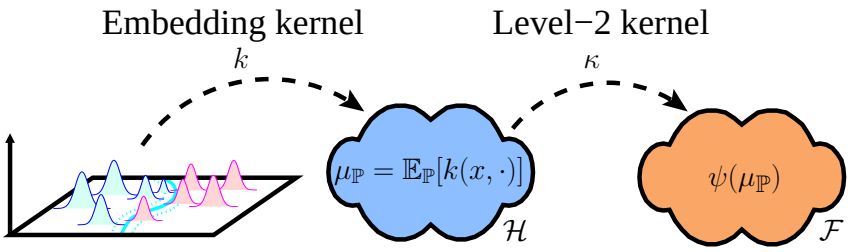


Figure 3.3: A visualization of a learning framework on distributional data. The embedding kernel k defines feature representation for distributions, while the level-2 kernel κ induces a class of non-linear functions over probability space based on such a representation.

Christmann and Steinwart (2010) considers a Gaussian-like kernel

$$\kappa(\mathbb{P}, \mathbb{Q}) = \exp\left(-\frac{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathscr{H}}}{2\sigma^2}\right). \tag{3.40}$$

Theorem 3.8 (Christmann and Steinwart (2010))**.** If the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective for a compact space $\mathcal{X}$, the function space $\mathscr{H}_\kappa$ is dense in $C_b(M_+^1)$.

Recently, the empirical kernel between distributions was employed in an unsupervised way for multi-task learning to generalize to a previously unseen task (Blanchard et al. 2011). Figure 3.3 summarizes the kernel-based framework for distributional data.

3.7.3 Distributional Risk Minimization

As also noted in Muandet et al. (2012), Muandet (2015) and Szabó et al. (2015), existing tools for theoretical analysis can be extended to learning framework on distributional data. Let assume that we have access to a training sample $(\mathbb{P}_1, y_1), \ldots, (\mathbb{P}_n, y_n) \in M_+^1 \times \mathcal{Y}$ generated from some unknown distribution over $M_+^1 \times \mathcal{Y}$, a strictly monotonically increasing function $\Omega : [0, \infty) \to \mathbb{R}$, and a loss function $l : (M_+^1 \times \mathbb{R}^2) \to \mathbb{R} \cup \{+\infty\}$, Muandet et al. (2012; Theorem 1) shows that any function $\hat{f} \in \mathscr{H}$ that minimizes a loss functional—called *dis-*

tributional risk minimization (DRM)

$$l(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_n, y_n, \mathbb{E}_{\mathbb{P}_n}[f]) + \lambda\Omega(\|f\|_{\mathscr{H}}) \tag{3.41}$$

admits a representation of the form

$$\hat{f} = \sum_{i=1}^{n} \alpha_i \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_i}[k(\mathbf{x}, \cdot)] = \sum_{i=1}^{n} \alpha_i \mu_{\mathbb{P}_i}. \tag{3.42}$$

Put differently, any solution $\hat{f}$ can be expressed in terms of the kernel mean embedding of $\mathbb{P}_1, \ldots, \mathbb{P}_n$. Note that if we restrict M_+^1 to contain only Dirac measures $\delta_{\mathbf{x}}$, then the solution reduces to $\hat{f} = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \cdot)$, *i.e.*, the classical representer theorem (Schölkopf et al. 2001).

On the one hand, it is instructive to compare the DRM in (3.41) to the vicinal risk minimization (VRM) of Chapelle et al. (2000). Specifically, they consider slightly different regularized functional

$$\mathbb{E}_{\mathbf{x}_1\sim\mathbb{P}_1} \cdots \mathbb{E}_{\mathbf{x}_n\sim\mathbb{P}_2} l(\mathbf{x}_1, y_1, f(\mathbf{x}_1), \ldots, \mathbf{x}_n, y_n, f(\mathbf{x}_n)) + \lambda\Omega(\|f\|_{\mathscr{H}}). \tag{3.43}$$

Intuitively, (3.43) amounts to the empirical risk minimization (ERM) on the samples drawn from $\mathbb{P}_1, \ldots, \mathbb{P}_n$. Note that (3.41) and (3.43) become equivalent only for a class of *linear* loss functional l. Arguably, (3.43) is ultimately what we want to minimize when learning from distributional data. However, it is computationally expensive and is not suitable for some applications. On the other hand, one may consider the following regularized functional

$$l(\mathbf{m}_1, y_1, f(\mathbf{m}_1), \ldots, \mathbf{m}_n, y_n, f(\mathbf{m}_n)) + \lambda\Omega(\|f\|_{\mathscr{H}}), \tag{3.44}$$

where $\mathbf{m}_i := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_i}[\mathbf{x}]$. That is, (3.44) corresponds to the ERM on the means of the distributions. Despite being more efficient, it throws away most information about high-level statistics. In some sense, the DRM can be viewed as something in between. Lastly, there is no specific assumption on the output space $\mathcal{Y}$, making it applicable to binary, real-valued, structured, or even distributional outputs.

Based on the DRM, Muandet et al. (2012) proposes the so-called *support measure machine* (SMM) which is a variant of an SVM that operates on distributions rather than points, permitting modeling of

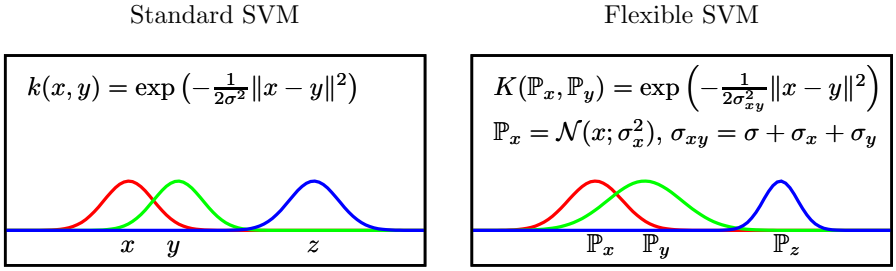


Figure 3.4: A picturial comparison between standard SVM algorithm and the flexible SVM algorithm. The flexible SVM allows us to put different kernel functions over training samples, as opposed to the standard SVM.

input uncertainties. In the special case of Gaussian input distributions and SVMs with Gaussian kernel, the SMM leads to a multi-scale SVM—akin to an RBF network with variable bandwidths—which is still trained by solving a QP problem (see Figure 3.4 for an illustration) (Muandet et al. 2012; Lemma 4). In Muandet et al. (2012), the SMM was applied to natural image categorization using bag-of-word (BoW) data representation, *i.e.*, each image is viewed as a distribution over codewords. Yoshikawa et al. (2014) proposed *latent SMM* which assumes that each codeword $t \in \mathcal{V}$ is represented by a q-dimensional latent vector $\mathbf{x}_t \in \mathbb{R}^q$ which is leant jointly with the SMM parameters (see also Yoshikawa et al. (2015) for a Gaussian process formulation). In unsupervised setting, one-class SMM (OCSMM) was studied in Muandet and Schölkopf (2013) with connection to variable KDE and application in group anomaly detection. Guevara et al. (2014) considers another equivalent characterization of the one-class problem leading to the *support measure data description* (SMDD).

Theorem 3.9 (Lopez-Paz et al. (2015))**.** Consider a class $\mathcal{F}_k$ of functionals mapping $\mathscr{H}$ to $\mathbb{R}$ with Lipschitz constants uniformly bounded by $L_{\mathcal{F}}$. Let $\varphi : \mathbb{R} \to \mathbb{R}^+$ be a L_φ-Lipschitz function such that $\varphi(z) \geq \mathbb{1}_{z>0}$. Let $\varphi(-f(h)l) \leq B$ for every $f \in \mathcal{F}_k$, $h \in \mathscr{H}$, and $l \in \mathcal{L}$. Define a surrogate φ-risk of any $f \in \mathcal{F}_k$ as

$$R_\varphi(f) = \mathbb{E}_{(z,l)\sim\mathbb{P}}[\varphi(-f(z)l)].$$

Then, with probability at least $1-\delta$,

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*) \leq 4L_\varphi R_n(\mathcal{F}_k) + 2B\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{4L_\varphi L_\mathcal{F}}{n}\sum_{i=1}^{n}\left(\sqrt{\frac{\mathbb{E}_{\mathbb{P}_i}[k(x,x)]}{n_i}} + \sqrt{\frac{\log(2n/\delta)}{2n_i}}\right),$$

where $R_n(\mathcal{F}_k)$ denotes a Rademacher complexity of $\mathcal{F}_k$.

The bound reveals important ingredients for the convergence of the excess risk to zero, *i.e.*, the consistency of the learning procedure on distributional data. Specifically, the upper bound converges to zero as both n (the number of distributions) and n_i (the size of the sample obtained from $\mathbb{P}_i$) tend to infinity, in such a way that $\log(n)/n_i = o(1)$. Moreover, n_i is only in the order of $\log(n)$ for the second term of the rhs of the bound to be small. Szabó et al. (2015) provides the detailed analysis of the ridge regression on distributional data.

Some recent works rely on a finite approximation of $\hat{\mu}_\mathbb{P}$, *i.e.*, the first step in Figure 3.3. For example, Lopez-Paz et al. (2015) proposes to solve a bivariate cause-effect inference between RVs X and Y as a classification problem on $\mu_{\mathbb{P}(X,Y)}$ approximated with random Fourier features Rahimi and Recht (2007). Inspired by the work of Eslami et al. (2014), Jitkrittum et al. (2015) proposes Just-In-Time kernel regression for learning to pass expectation propagation (EP) messages. To perform the message passing efficiently, the authors use two-stage random feature for κ, *i.e.*, they first construct random Fourier features to approximate $\mu_\mathbb{P}$ on which the second set of random features can be generated to approximate (3.40).

Related Works on Learning from Distributions

Several works in the past have attempted to leverage information from distributions (or generative models such as hidden Markov models) over complex objects in discriminative models. Jaakkola and Haussler (1998) provides a generic procedure for obtaining kernel functions from generative probability models. Given a generative probability model

Table 3.3: A closed-form solution of the kernel $K(\mathbb{P},\mathbb{Q}) = \iint k(x,y)\,\mathrm{d}\mathbb{P}(x)\,\mathrm{d}\mathbb{Q}(y)$ for a certain class of probability distributions $\mathbb{P},\mathbb{Q}$ and kernel function k (reproduced from Muandet et al. (2012)).

Distribution	$k(\mathbf{x},\mathbf{y})$	$K(\mathbb{P}_i,\mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathscr{H}}$
$\mathbb{P}(\mathbf{m};\Sigma)$	$\langle \mathbf{x},\mathbf{y}\rangle$	$\mathbf{m}_i^\mathsf{T}\mathbf{m}_j + \delta_{ij}\mathrm{tr}\boldsymbol{\Sigma}_i$
$\mathcal{N}(\mathbf{m},\Sigma)$	$\exp(-\frac{\gamma}{2}\|\mathbf{x}-\mathbf{y}\|^2)$	$\exp(-\frac{1}{2}(\mathbf{m}_i-\mathbf{m}_j)^\mathsf{T}(\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j+\gamma^{-1}\mathbf{I})^{-1}(\mathbf{m}_i-\mathbf{m}_j))$ $/\lvert\gamma\boldsymbol{\Sigma}_i+\gamma\boldsymbol{\Sigma}_j+\mathbf{I}\rvert^{\frac{1}{2}}$
$\mathcal{N}(\mathbf{m},\Sigma)$	$(\langle \mathbf{x},\mathbf{y}\rangle+1)^2$	$(\langle \mathbf{m}_i,\mathbf{m}_j\rangle+1)^2 + \mathrm{tr}\ \boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j + \mathbf{m}_i^\mathsf{T}\boldsymbol{\Sigma}_j\mathbf{m}_i + \mathbf{m}_j^\mathsf{T}\boldsymbol{\Sigma}_i\mathbf{m}_j$
$\mathcal{N}(\mathbf{m},\Sigma)$	$(\langle \mathbf{x},\mathbf{y}\rangle+1)^3$	$(\langle \mathbf{m}_i,\mathbf{m}_j\rangle+1)^3 + 6\mathbf{m}_i^\mathsf{T}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j\mathbf{m}_j$ $+3(\langle \mathbf{m}_i,\mathbf{m}_j\rangle+1)(\mathrm{tr}\ \boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j + \mathbf{m}_i^\mathsf{T}\boldsymbol{\Sigma}_j\mathbf{m}_i + \mathbf{m}_j^\mathsf{T}\boldsymbol{\Sigma}_i\mathbf{m}_j)$

$\mathbb{P}(\mathbf{x}|\boldsymbol{\theta})$, where $\mathbf{x}$ is a data point and $\boldsymbol{\theta}$ is a vector of the model parameters, they define *Fisher kernel*

$$K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{u}_i^\top \mathcal{I}^{-1}\mathbf{u}_j \tag{3.45}$$

where $\mathbf{u}_i$ is the *Fisher score vector* defined by $\mathbf{u}_i := \nabla_{\boldsymbol{\theta}} \log \mathbb{P}(\mathbf{x}_i|\boldsymbol{\theta})$ and $\mathcal{I}$ is a Fisher information matrix.[10] The benefit of Fisher kernel is that the resulting discriminative models such as SVMs are well-informed about the underlying generative models such as GMMs and HMMs, which makes it useful for structure data such as biological data, logical sequences, and documents. Similarly, Jebara et al. (2004) proposed the *probability product kernel* (PPK)

$$K_\rho(p,q) = \int_{\mathcal{X}} p(\mathbf{x})^\rho q(\mathbf{x})^\rho\,\mathrm{d}\mathbf{x}, \tag{3.46}$$

which is a generalized inner product between two input objects. That is, the probabilistic models p and q are learned for each example and used as a surrogate to construct the kernel between those examples. The kernel (3.46) can be evaluated for all exponential families such as multinomials and Gaussians. Moreover, the PPK (with a certain value of ρ) can be computed analytically for some distributions such as mixture models, HMMs, and linear dynamical systems. For intractable models, Jebara et al. (2004) suggested to approximate (3.46) using structured mean-field approximations. As a result, the PPK can be applied for a broader class of generative models.

[10] Intuitively, the gradient of log-likelihood $\mathbf{u}_i$ specifies how the parameter θ contributes to the process of generating the example $\mathbf{x}_i$ (Jaakkola and Haussler 1998).

The PPK is in fact closely related to well-known kernels such as the Bhattacharyya kernel (Bhattacharyya 1943) and the exponential symmetrized Kullback-Leibler (KL) divergence (Moreno et al. 2004). In Hein and Bousquet (2005), an extension of a two-parameter family of Hilbertian metrics of Topsøe was used to define Hilbertian kernels on probability measures. In Cuturi et al. (2005), the semi-group kernels were designed for objects with additive semi-group structure such as positive measures. Recently, Martins et al. (2009) introduced non-extensive information theoretic kernels on probability measures based on new Jensen-Shannon-type divergences. Although these kernels have proven successful in many applications, they are designed specifically for certain properties of distributions and application domains. Moreover, there has been no attempt in making a connection to the kernels on corresponding input spaces.

The kernel function $K(\mathbb{P}, \mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathscr{H}}$ considered earlier can in fact be understood as a special case of the Hilbertian metric (Hein and Bousquet 2005), with the associated kernel

$$K(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}, \tilde{\mathbf{x}} \sim \mathbb{Q}}[k(\mathbf{x}, \tilde{\mathbf{x}})], \tag{3.47}$$

and a generative mean map kernel (GMMK) proposed by Mehta and Gray (2010). In the GMMK, the kernel between two objects $\mathbf{x}$ and $\mathbf{y}$ is defined via $\hat{p}_{\mathbf{x}}$ and $\hat{p}_{\mathbf{y}}$, which are estimated probabilistic models of $\mathbf{x}$ and $\mathbf{y}$, respectively. Moreover, the empirical version of (3.47) and (3.38) coincides with that of classical set kernel of Haussler (1999) and Gärtner et al. (2002) (see also Kondor and Jebara (2003)). This connection reveals an interesting fact that—although originally presented as a similarity measure between sets of vectors—this class of kernels has a natural interpretation as similarity between the underlying probability distributions.

It has been shown that the PPK is a special case of GMMK when $\rho = 1$ (Mehta and Gray 2010). Consequently, GMMK, PPK with $\rho = 1$, and linear kernel $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathscr{H}}$ are equivalent when the embedding kernel is $k(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$.

Distributions can be used to capture input uncertainty. The use of the kernel (3.38) in dealing with the uncertainty in input data has a connection to robust SVMs. For example, a generalized form of the

SVM in Shivaswamy et al. (2006) incorporates the probabilistic uncertainty into the maximization of the margin. This results in a second-order cone programming (SOCP) that generalizes the standard SVM. In SOCP, one needs to specify the parameter τ_i that reflects the probability of correctly classifying the ith training example. In the context of this section, we may represent data point $\mathbf{x}_i$ by a distribution $\mathcal{N}(\mathbf{x}_i, \sigma_i^2\mathbf{I})$. Therefore, the parameter τ_i is closely related to the parameter σ_i, which specifies the variance of the distribution centered at the ith example. Anderson and Gupta (2011) showed the equivalence between SVMs using expected kernels and SOCP when $\tau_i = 0$. When $\tau_i > 0$, the mean and covariance of missing kernel entries have to be estimated explicitly, making the SOCP more involved for nonlinear kernels. Although achieving comparable performance to the standard SVM with expected kernels, the SOCP requires a more computationally extensive SOCP solver, as opposed to simple quadratic programming (QP). From a Bayesian perspective, the problem of regression with uncertain inputs has been studied in Gaussian process community, e.g., Girard et al. (2002).

A major drawback of the previous works in the literature is that they usually impose a strong *parametric assumption* on the form of probability distribution. The kernel mean representation, on the other hand, allows one to learn directly from distributions without such an assumption. For example, Szabó et al. (2015) has recently studied the *nonparametric* distributions regression problem based on kernel mean embedding and the kernel ridge regression algorithm. They establish the consistency and convergence rate of the resulting algorithm whose challenge arises from the *two-stage sampling*: a meta distribution generates i.i.d. sample of distributions from which i.i.d observations have been generated (see also Theorem 3.9). As a result, in practice we only observe samples from the distributions rather than the distributions themselves. The theoretical analysis uses the results of Caponnetto and De Vito (2007) who provides error bounds for regularized least-squares algorithm in standard setting.

In addition to the mean embedding approach, another line of research employs kernel density estimation (KDE) to perform regres-

sion on distributions with consistency guarantee (under the assumption that the true regressor is Hölder continuous, and the meta distribution have finite doubling dimension (Kpotufe 2011)) (Póczos et al. 2013, Oliva et al. 2014). In this case the covariates are nonparametric continuous distributions on $\mathbb{R}^d$ and the output are real-valued. Oliva et al. (2013) also considers the case when the output is also distribution. The basic idea is to approximate the density function by KDE and then apply kernels on top of it. Unlike the mean embedding approach, the kernels used are classical smoothing kernels and not the reproducing kernel. Although the parametric assumption is not needed, drawbacks of the KDE-based approach are that the convergence rate is slow in high-dimensional space and it is not applicable to learning over structured data such as documents, graphs, and permutations. The use of kernel mean embedding allows us to deal with any kind of data as long as the positive definite kernel on such data is well-defined.

3.8 Recovering Information from Mean Embeddings

Given a kernel mean embedding $\mu_{\mathbb{P}}$, can we recover essential properties of $\mathbb{P}$ from $\mu_{\mathbb{P}}$? We respond to this question by discussing two closely related problems, namely, distributional pre-image problem[11] (Kwok and Tsang 2004, Song et al. 2008, Kanagawa and Fukumizu 2014) and kernel herding (Chen et al. 2010). We consider these two problems to be related because both of them involve finding objects in the input space which correspond to specific kernel mean embedding in the feature space.

The classical pre-image problem in kernel methods involves finding patterns in input space that map to specific feature vectors in the feature space (Schölkopf and Smola 2001; Chapter 18). Recovering a pre-image is considered necessary in some applications such as image denoising using kernel PCA (Kwok and Tsang 2004, Kim et al. 2005) and visualizing the clustering solutions of a kernel-based clustering algorithm (Dhillon et al. 2004, Jegelka et al. 2009). Moreover, it

[11] We call this a *distributional pre-image* problem to distinguish it from the classical setting which does not involve probability distributions.

can be used as a reduced set method to compress a kernel expansion (Schölkopf and Smola 2001; Chapter 18). Schölkopf and Smola (2001; Proposition 18.1) shows that if the pre-image exists and the kernel is an invertible function of $\langle \mathbf{x}, \mathbf{x}' \rangle$, the pre-image will be easy to compute. Unfortunately, the exact pre-image typically does not exist, and the best one can do is to approximate it. There is a fair amount of works on this topic and the interested readers should consult Schölkopf and Smola (2001; Chapter 18) for further detail.

3.8.1 Distributional Pre-Image Problem

Likewise, in some applications of kernel mean embedding, it is important to recover the meaningful information of an underlying distribution from an estimate of its embedding. In state-space model, for example, we typically obtain a kernel mean estimate of the predictive distribution from the algorithm (Song et al. 2009, Nishiyama et al. 2012, McCalman et al. 2013). To obtain meaningful information, we need to extract the information of $\mathbb{P}$ from the estimate. Unfortunately, in these applications we only have access to the estimate $\hat{\mu}_X$ which lives in a high-dimensional feature space.

The idea is similar to the *approximate pre-image problem*. Let $\mathbb{P}_{\boldsymbol{\theta}}$ be an arbitrary distribution parametrized by $\boldsymbol{\theta}$ and $\mu_{\mathbb{P}_{\boldsymbol{\theta}}}$ be its mean embedding in $\mathscr{H}$. One can find $\mathbb{P}_{\boldsymbol{\theta}}$ by the following minimization problem

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \ \|\hat{\mu}_X - \mu_{\mathbb{P}_{\boldsymbol{\theta}}}\|^2_{\mathscr{H}} \tag{3.48}$$

subject to appropriate constraints on the parameter vector $\boldsymbol{\theta}$. Note that if $\mathbb{P}_{\boldsymbol{\theta}} = \delta_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{X}$, the distributional pre-image problem (3.48) reduces to the classical pre-image problem. The pre-image $\mathbf{x}$ can be viewed as a *point estimate* of the underlying distribution, but may not correspond to the MAP estimate.

Another example is a mixture of Gaussians $\mathbb{P}_{\boldsymbol{\theta}} = \sum_{i=1}^m \pi_i \mathcal{N}(\mathbf{m}_i, \sigma_i^2 \mathbf{I})$ where the parameter $\boldsymbol{\theta}$ consists of $\{\pi_1, \ldots, \pi_m\}$, $\{\mathbf{m}_1, \ldots, \mathbf{m}_m\}$, and $\{\sigma_1, \ldots, \sigma_m\}$. It is required that $\sum_{i=1}^m \pi_i = 1$ and $\sigma_i \geq 0$. Let assume that $\hat{\mu}_X = \sum_{i=1}^n \beta_i \phi(\mathbf{x}_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. In this

case, the optimization problem (3.48) reduces to

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \; \boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{Q}\boldsymbol{\pi} + \boldsymbol{\pi}^\top \mathbf{R}\boldsymbol{\pi}\,, \tag{3.49}$$

where

$$\begin{aligned}
\mathbf{K}_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j) \\
\mathbf{Q}_{ij} &= \int k(\mathbf{x}_i, \mathbf{x}')\, \mathrm{d}\mathcal{N}(\mathbf{x}'; \mathbf{m}_j, \sigma_j^2 \mathbf{I}) \\
\mathbf{R}_{ij} &= \iint k(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mathcal{N}(\mathbf{x}; \mathbf{m}_i, \sigma_i^2 \mathbf{I})\, \mathrm{d}\mathcal{N}(\mathbf{x}'; \mathbf{m}_j, \sigma_j^2 \mathbf{I})\,.
\end{aligned}$$

Note that (3.49) is quadratic in $\boldsymbol{\pi}$ and is also convex in $\boldsymbol{\pi}$ as $\mathbf{K}$, $\mathbf{Q}$, and $\mathbf{R}$ are positive definite. The integrals $\mathbf{Q}_{ij}$ and $\mathbf{R}_{ij}$ can be evaluated in close-form for some kernels (see Song et al. (2008; Table 1) and Muandet et al. (2012; Table 1)). Unfortunately, the problem is often non-convex in both $\mathbf{m}_i$ and σ_i, $i = 1, \ldots, m$. An derivative-free optimization is often used to find these parameters. In practice, π_i and $\{\mathbf{m}_i, \sigma_i\}$ are solved alternately until convergence (see, *e.g.*, Song et al. (2008), Chen et al. (2010)).

The reduced set problem is slightly more general than the pre-image problem because we do not just look for single pre-images, but for expansions of several input vectors. Interestingly, we may view the reduced set problem as a specific case of distributional pre-image problem. To understand this, assume we are given a function $g \in \mathscr{H}$ as a linear combination of the images of input points $\mathbf{x}_i \in \mathcal{X}$, *i.e.*, $g = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$. The function g is exactly the kernel mean embedding of the finite signed measure $\nu = \sum_{i=1}^n \alpha_i \delta_{\mathbf{x}_i}$ whose supports are the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. That is, $g = \int \phi(\mathbf{y})\, \mathrm{d}\nu(\mathbf{y})$. Given the reduced set vector $\mathbf{z}_1, \ldots, \mathbf{z}_m$ where $m \ll n$, the reduced set problem amounts to finding another finite signed measure $\mu = \sum_{j=1}^m \beta_j \phi(\mathbf{z}_j)$ whose supports are $\mathbf{z}_1, \ldots, \mathbf{z}_m$ that approximates well the original measure ν. From the distributional pre-image problem, the reduced set methods can be viewed as an approximation of a finite signed measure by another signed measure whose supports are smaller.

Although it is possible to find a distributional pre-image, it is not clear what kind of information of $\mathbb{P}$ this pre-image represents. Kanagawa and Fukumizu (2014) considers the recovery of the information of a distribution from an estimate of the kernel mean when

the Gaussian RBF kernel on Euclidean space is used. Specifically, they show that under some situations we can recover certain statistics of $\mathbb{P}$, namely its moments and measures on intervals, from $\hat{\mu}_{\mathbb{P}}$, and that the density of $\mathbb{P}$ can be estimated from $\hat{\mu}_{\mathbb{P}}$ without any parametric assumption on $\mathbb{P}$ (Kanagawa and Fukumizu 2014; Theorem 2). Moreover, they prove that the weighted average of function f in some *Besov space* converges to the expectation of f, *i.e.*, $\sum_i w_i f(X_i) \to \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ (Kanagawa and Fukumizu 2014; Theorem 1).[12] This result is a generalization of the known result for functions in an RKHS.

3.8.2 Kernel Herding

Instead of finding a distributional pre-image of the mean embedding, another common application is obtaining sample from the distribution or sampling. Chen et al. (2010) proposes a kernel herding algorithm that extends herding algorithm (Welling 2009a;b, Welling and Chen 2010) to continuous spaces by using the kernel trick. Herding can be understood concisely as a weakly chaotic non-linear dynamical system $\mathbf{w}_{t+1} = F(\mathbf{w}_t)$. In Chen et al. (2010), they re-interpret herding as an infinite memory process in the state space $\mathbf{x}$ by marginalizing out the parameter $\mathbf{w}$, resulting in a mapping $\mathbf{x}_{t+1} = G(\mathbf{x}_1, \ldots, \mathbf{x}_t; \mathbf{w}_0)$. Under some technical assumptions, herding can be seen to greedily minimize the squared error

$$\mathcal{E}_T^2 := \left\| \mu_{\mathbb{P}} - \frac{1}{T}\sum_{t=1}^{T} \phi(\mathbf{x}_t) \right\|_{\mathscr{H}}^2 = \|\mu_{\mathbb{P}} - \hat{\mu}_T\|_{\mathscr{H}}^2 , \tag{3.50}$$

where $\hat{\mu}_T$ denotes the empirical mean embedding obtained from herding. Following the result of Welling (2009a), kernel herding is shown to decrease the error of expectations of functions in the RKHS at a rate $O(1/T)$ as opposed to the random samples whose rate is $O(1/\sqrt{T})$. The fast rate is guaranteed even when herding is carried out with some error. This condition is reminiscent of Boosting algorithm and perceptron cycling theorem (Chen et al. 2010; Corollary 2). The reason for

[12] The Besov space is a complete quasinormed space, which also coincide with the more classical Sobolev spaces. For details of Besov spaces, see Adams and Fournier (2003; Chapter 7).

fast convergence is due to *negative autocorrelation*, *i.e.*, herding tends to find samples in an unexplored high-density region. This kind of behaviour can also be observed in Quasi Monte Carlo integration and Bayesian quadrature methods (Rasmussen and Ghahramani 2002).

Huszar and Duvenaud (2012) also investigates the kernel herding problem and suggests a connection between herding and Bayesian quadrature. Bayesian quadrature (BQ) (Rasmussen and Ghahramani 2002) estimates the integral $Z = \int f(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$ by putting a prior distribution on f and then inferring a posterior distribution over f conditioned on the observed evaluations. An estimate of Z can be obtained by a posterior expectation, for example. The sampling strategy of BQ is to select the sample so as to minimize the posterior variance. Huszar and Duvenaud (2012) shows that the posterior variance in BQ is equivalent to the criterion (3.50) minimized when selecting samples in kernel herding. An advantage of Bayesian interpretation of herding is that kernel parameters can be chosen by maximizing the marginal likelihood. As an aside, it is instructive to note that the herding sample exhibits similar behaviour to those defined under Determinantal Point Process (DPP) prior, *i.e.*, a DPP assigns high probability to sets of items that are *repulsive* (Kulesza and Taskar 2012).[13] Important applications of DPP are, for example, information retrieval and text summarization.

Herding can be problematic in high dimensional setting when optimizing over the new sample. Bach et al. (2012) also pointed out that the fast convergence rate is not guaranteed in an infinite dimensional Hilbert space. To alleviate this issue, Bach et al. (2012) shows that the herding procedure of Welling (2009a) takes the form of a convex optimization algorithm in which convergence results can be invoked. Lacoste-Julien et al. (2015) takes this interpretation and proposes the Frank-Wolfe optimization algorithm for particle filtering. Lastly, it is instructive to mention that kernel herding in RKHSs has an implicit connection to Quasi-Monte Carlo (QMC) theory, *i.e.*, one can show by

[13] Like kernel herding, a DPP prior is characterized via a marginal kernel $\mathbf{K}$. For a subset $A \subset \mathcal{X}$, it assigns a probability proportional to $\det(\mathbf{K}_A)$ where $\mathbf{K}_A$ denotes a restriction of $\mathbf{K}$ to the elements of A.

reproducing property and Cauchy-Schwartz inequality that the integration error

$$\mathcal{E}_{p,S}(f) := \left| \int_{\mathbb{R}^d} f(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \frac{1}{s} \sum_{\mathbf{w} \in S} f(\mathbf{w}) \right| \leq \|f\|_{\mathscr{H}} D(S)$$

where $D^2(\cdot)$ is a discrepancy measure:

$$D(S) := \left\| \mu_p - \frac{1}{s} \sum_{\mathbf{w} \in S} k(\mathbf{w}, \cdot) \right\|_{\mathscr{H}}^2 . \tag{3.51}$$

This connection has been noted by several authors including Bach et al. (2012) and Yang et al. (2014).

4

Hilbert Space Embedding of Conditional Distributions

In the previous section, we discuss the embedding of marginal distributions in RKHS and gives comprehensive reviews on various applications. Throughout this section we will extend the concept of kernel mean embedding to a *conditional distribution* $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|X = \mathbf{x})$ for some $\mathbf{x} \in \mathcal{X}$ (Song et al. 2009; 2013). Unlike the marginal distribution $\mathbb{P}(X)$, the conditional distribution $\mathbb{P}(Y|X)$ captures the functional relationship between two random variables, namely X and Y. Hence, the conditional mean embedding extends the capability of kernel mean embedding to model more complex dependency in various applications such as dynamical systems (Song et al. 2009, Boots et al. 2013), Markov decision processes and reinforcement learning (Grünewälder et al. 2012, Nishiyama et al. 2012), latent variable model (Song et al. 2010b; 2011a;b), kernel Bayes rule (Fukumizu et al. 2011), and causal discovery (Janzing et al. 2011, Sgouritsa et al. 2013, Chen et al. 2014). Figure 4.1 gives a schematic illustration of conditional mean embedding.

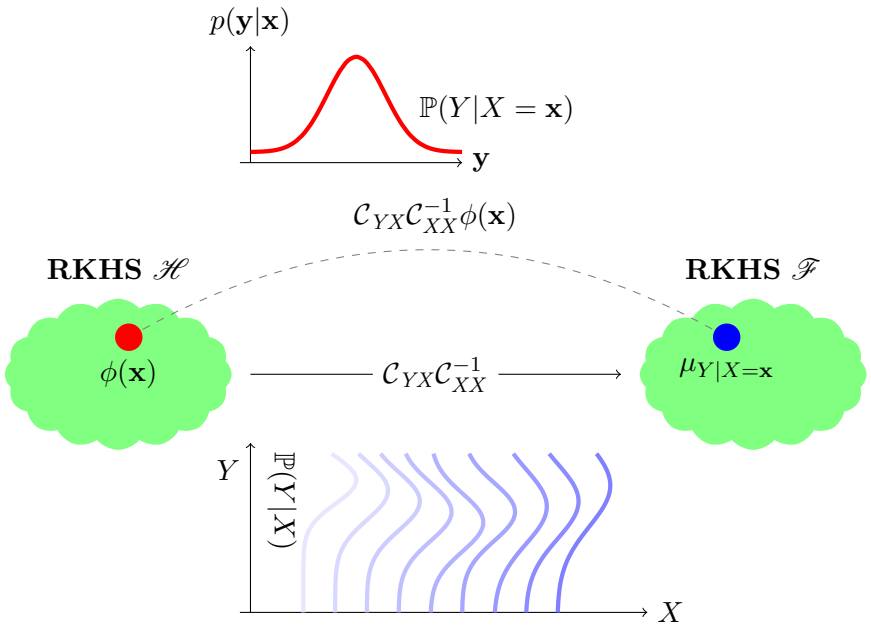


Figure 4.1: From marginal distribution to conditional distribution: unlike the embeddings discussed in the previous section, the embedding of conditional distribution $\mathbb{P}(Y|X)$ is not a single element in the RKHS. Instead, it may be viewed as a family of Hilbert space embeddings of the conditional distributions $\mathbb{P}(Y|X=\mathbf{x})$ indexed by the conditioning variable X. In other words, the conditional mean embedding can be viewed as an operator mapping from $\mathscr{H}$ to $\mathscr{F}$. We will see later in §4.2 that there is a natural interpretation in a vector-valued regression framework.

4.1 From Marginal to Conditional Distribution

To better understand the distinction between the kernel mean embedding of marginal and conditional distributions, and the problems that we may encounter in conditional mean embedding, we briefly summarize the concept of marginal, joint, and conditional distributions. Detailed materials should be widely available in most statistics textbooks, see, *e.g.*, Wasserman (2010). Readers already familiar with these concepts may wish to proceed directly to the definition of conditional mean embedding.

Given two random variables X and Y, probabilities defined on them may be either marginal, joint, or conditional. Marginal probabilities

$\mathbb{P}(X)$ and $\mathbb{P}(Y)$ are the (unconditional) probabilities of an event occurring. For example, if X denotes the level of cloudyness of the outside sky, $\mathbb{P}(X)$ describes how likely it is for the outside sky to be cloudy. Joint probability $\mathbb{P}(X, Y)$ is the probability of event $X = x$ and $Y = y$ occurring. If Y indicates whether or not it is raining, the joint distribution $\mathbb{P}(X, Y)$ explains the probability that it is both raining and cloudy outside. As we can see, joint distributions allow us to reason about the relationship between multiple events, which in this case are cloudyness and rain. Following the above definitions, one may subsequently ask given that it is cloudy outside, *i.e.*, $X = \texttt{cloudy}$, what is the probability that it is also raining? Conditional distribution $\mathbb{P}(Y|X)$ governs such a question. Formally, the conditional probability $\mathbb{P}(Y = y|X = x)$ is the probability of event $Y = y$ occurring, given that event $X = x$ has occurred. In other words, conditional probabilities allow us to reason about causality.[1]

The basic relationships between marginal, joint, and conditional distributions can be illustrated via the following equations:

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}. \tag{4.1}$$

As we can see in the first equation of (4.1), the conditional probability of Y given X is equal to the joint probability of X and Y divided by the marginal of X. Marginal, joint, and conditional distributions equipped with the formulation (4.1) provide the powerful language for statistical inference in statistics and machine learning.

4.1.1 Conditional Mean Embeddings

Suppose k and l are the positive definite kernels for the domains of X and Y, respectively, with corresponding RKHS's $\mathscr{H}$ and $\mathscr{F}$. Let $\mathcal{U}_{Y|X} : \mathscr{H} \to \mathscr{F}$ and $\mathcal{U}_{Y|\mathbf{x}} \in \mathscr{F}$ be conditional mean embeddings of the conditional distribution $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|X = \mathbf{x})$, respectively, such

[1] To be more precise, the fundamental question in causal inference/discovery from observational data is to identify conditions under which $\mathbb{P}(Y \,|\, \mathrm{do}(X = \mathbf{x}))$ is equal to $\mathbb{P}(Y \,|\, X = \mathbf{x})$ where $\mathrm{do}(X = \mathbf{x})$ denotes the operation of setting the value of X to be equal to $\mathbf{x}$ (Pearl 2000). Under such conditions, one is allowed to make a causal claim from the conditional distribution $\mathbb{P}(Y \,|\, X = \mathbf{x})$.

that they satisfy

$$\mathcal{U}_{Y|\mathbf{x}} = \mathbb{E}_{Y|\mathbf{x}}[\varphi(Y)|X=\mathbf{x}] = \mathcal{U}_{Y|X}k(\mathbf{x},\cdot) \quad (4.2)$$

$$\mathbb{E}_{Y|\mathbf{x}}[g(Y)|X=\mathbf{x}] = \langle g, \mathcal{U}_{Y|\mathbf{x}}\rangle_{\mathscr{F}}, \quad \forall g \in \mathscr{F}. \quad (4.3)$$

Note that $\mathcal{U}_{Y|X}$ is an operator from $\mathscr{H}$ to $\mathscr{F}$, whereas $\mathcal{U}_{Y|\mathbf{x}}$ is an element in $\mathscr{F}$. As an interpretation, condition (4.2) says that the conditional mean embedding of $\mathbb{P}(Y|X=\mathbf{x})$ should correspond to the conditional expectation of the feature map of Y given that $X=\mathbf{x}$ (as in the marginal embedding). Moreover, the embedding operator $\mathcal{U}_{Y|X}$ represents the *conditioning operation* that when applied to $\phi(\mathbf{x}) \in \mathscr{H}$ outputs the conditional mean embedding $\mathcal{U}_{Y|\mathbf{x}}$ (see also Figure 4.1). Condition (4.3) ensures the reproducing property of $\mathcal{U}_{Y|\mathbf{x}}$, *i.e.*, it should be a representer of conditional expectation in $\mathscr{F}$ w.r.t. $\mathbb{P}(Y|X=\mathbf{x})$ (as in the marginal embedding).

The following definition provides explicit form of $\mathcal{U}_{Y|X}$ and $\mathcal{U}_{Y|\mathbf{x}}$.

Definition 4.1 (Song et al. (2009; 2013))**.** Let $\mathcal{C}_{XX} : \mathscr{H} \to \mathscr{H}$ and $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$ be covariance operator on X and cross-covariance operator from X to Y, respectively. Then, the conditional mean embedding $\mathcal{U}_{Y|X}$ and $\mathcal{U}_{Y|\mathbf{x}}$ are defined as

$$\mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} \quad (4.4)$$

$$\mathcal{U}_{Y|\mathbf{x}} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(\mathbf{x},\cdot). \quad (4.5)$$

Under the assumption that $\mathbb{E}_{Y|X}[g(Y)|X] \in \mathscr{H}$, Song et al. (2009) shows that the conditional mean embedding given in Definition 4.1 satisfies both (4.2) and (4.3). This result follows from Fukumizu et al. (2004; Theorem 2) which is also given in Theorem 3.2 in this survey. Recall from Theorem 3.2 that for any $g \in \mathscr{F}$

$$\mathcal{C}_{XX}\mathbb{E}_{Y|X}[g(Y)|X=\cdot] = \mathcal{C}_{XY}g. \quad (4.6)$$

For some $\mathbf{x} \in \mathcal{X}$, we have by virtue of reproducing property that $\mathbb{E}_{Y|\mathbf{x}}[g(Y)|X=\mathbf{x}] = \langle \mathbb{E}_{Y|X}[g(Y)|X], k(\mathbf{x},\cdot)\rangle_{\mathscr{H}}$. Using relation (4.6) and taking the conjugate transpose of $\mathcal{C}_{XX}^{-1}\mathcal{C}_{XY}$ yields $\mathbb{E}_{Y|\mathbf{x}}[g(Y)|X=\mathbf{x}] = \langle g, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(\mathbf{x},\cdot)\rangle_{\mathscr{F}} = \langle g, \mathcal{U}_{Y|\mathbf{x}}\rangle_{\mathscr{F}}$.

One should also keep in mind that, unlike the marginal mean embedding, the operator $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ only acts as an approximation of the conditional mean embedding $\mathcal{U}_{Y|X}$ in the continuous domain because the assumption that for all $g \in \mathscr{F}$, the conditional expectation $\mathbb{E}_{Y|X}[g(Y)|X = \cdot]$ is an element of $\mathscr{H}$ may not hold in general (Fukumizu et al. 2004, Song et al. 2009).[2] This technical issue can be circumvented by resorting to a regularized version of (4.5), *i.e.*, $\mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon\mathbf{I})^{-1}k(\mathbf{x}, \cdot)$ where $\varepsilon > 0$ denotes a regularization parameter. Fukumizu et al. (2013; Theorem 8) showed that—under some mild conditions—its empirical estimator is a consistent estimator of $\mathbb{E}_{Y|X}[g(Y)|X = \mathbf{x}]$.

Theorem 4.1 (Song et al. (2009; Eq. 6), Fukumizu et al. (2013; Theorem 2))**.** Let μ_Π and μ_{Q_y} be the kernel mean embeddings of Π in $\mathscr{H}$ and Q_y in $\mathscr{F}$, respectively. If $\mathcal{C}_{XX}$ is injective, $\mu_\Pi \in \mathcal{R}(\mathcal{C}_{XX})$, and $\mathbb{E}[g(Y)|X = \cdot] \in \mathscr{H}$ for any $g \in \mathscr{F}$, then

$$\mu_{Q_y} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\mu_\Pi$$

where $\mathcal{C}_{XX}^{-1}\mu_\Pi$ denotes the function mapped to μ_Π by $\mathcal{C}_{XX}$.

4.1.2 Empirical Estimation of Conditional Mean Embeddings

Since the joint distribution $\mathbb{P}(X, Y)$ is unknown in practice, we cannot compute $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ directly. Instead, we must rely on the i.i.d. sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ from $\mathbb{P}(X, Y)$. With an abuse of notation, let $\Phi := [\varphi(\mathbf{y}_1), \ldots, \varphi(\mathbf{y}_n)]^\top$ and $\Upsilon := [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]^\top$. We define $\mathbf{K} = \Upsilon^\top\Upsilon$ and $\mathbf{L} = \Phi^\top\Phi$ as the corresponding Gram matrices. Then, the empirical estimator of the conditional mean embedding is given by

$$\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \lambda\mathcal{I})^{-1}k(\mathbf{x}, \cdot) \;=\; \frac{1}{n}\Phi\Upsilon^\top\left(\frac{1}{n}\Upsilon\Upsilon^\top + \lambda\mathcal{I}\right)^{-1}k(\mathbf{x}, \cdot)$$

[2] For example, when $\mathscr{H}$ and $\mathscr{F}$ are both Gaussian RKHSes, and X and Y are independent, $\mathbb{E}_{Y|X}[g(Y)|X = \mathbf{x}]$ is a constant function of $\mathbf{x}$, which is not contained in the Gaussian RKHS (Steinwart and Christmann 2008; Corollary 4.44). If X and Y are discrete random variables and the kernels are characteristic, $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathscr{H}$.

$$
\begin{aligned}
&= \Phi\Upsilon^\top \left(\Upsilon\Upsilon^\top + n\lambda\mathcal{I}\right)^{-1} k(\mathbf{x},\cdot) \\
&= \Phi \left(\Upsilon^\top\Upsilon + n\lambda\mathbf{I}_n\right)^{-1} \Upsilon^\top k(\mathbf{x},\cdot) \\
&= \Phi \left(\mathbf{K} + n\lambda\mathbf{I}_n\right)^{-1} \mathbf{k}_{\mathbf{x}}, \qquad (4.7)
\end{aligned}
$$

where $\mathcal{I}$ denotes the identity operator in $\mathscr{H}$ and $\mathbf{k}_{\mathbf{x}} := \Upsilon^\top k(\mathbf{x},\cdot)$. The most important step of the derivation uses the identity $\Upsilon^\top(\Upsilon\Upsilon^\top + n\lambda\mathcal{I})^{-1} = (\Upsilon^\top\Upsilon + n\lambda\mathbf{I}_n)^{-1}\Upsilon^\top$. Since $\widehat{\mathcal{C}}_{XX}$ is a compact operator, it has an arbitrary small positive eigenvalue when infinite dimensional RKHS's are assumed.[3] We thus need a regularizer $\lambda\mathcal{I}$ for the inverse of $\widehat{\mathcal{C}}_{XX}$ to be well-posed. Another possibility is to employ the spectral filtering algorithms, *i.e.*, $\hat{\mu} = \Phi g_\lambda(\mathbf{K})\mathbf{k}_{\mathbf{x}}$ where g_λ is a filter function, as also suggested by Muandet et al. (2014b). That is, we can construct a wide class of conditional mean estimators via different regularization strategies.

Theorem 4.2 gives a formal characterization on the empirical estimator of conditional mean embedding.

Theorem 4.2 (Song et al. (2009))**.** The conditional mean embedding $\mu_{Y|\mathbf{x}}$ can be estimated as

$$\hat{\mu}_{Y|\mathbf{x}} = \Phi(\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{k}_{\mathbf{x}}. \qquad (4.8)$$

Interestingly, we may write (4.8) as $\hat{\mu}_{Y|\mathbf{x}} = \Phi\boldsymbol{\beta} = \sum_{i=1}^n \beta_i\varphi(\mathbf{y}_i)$ where $\boldsymbol{\beta} := (\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{k}_{\mathbf{x}} \in \mathbb{R}^n$. That is, it is in the same form as the embedding of marginal distribution discussed previously, except that the values of coefficients $\boldsymbol{\beta}$ now depends on the value of the conditioning variable X instead of being uniform (Song et al. 2009). It is important to note that in this case the coefficient $\boldsymbol{\beta}$ need not be positive nor does it has to sum to one. In some applications of conditional mean embedding such as state-space model and reinforcement learning, however, one need to interpret $\boldsymbol{\beta}$ as probabilities, which is almost always

[3] If X and Y are Banach spaces and $\mathbf{T} : X \to Y$ is a bounded linear operator, $\mathbf{T}$ is said to be a compact operator if $\mathbf{T}$ maps the unit ball in X to a pre-compact set in Y. Equivalently, $\mathbf{T}$ is compact if and only if for every bounded sequences $\{x_n\} \subset X$, $\{\mathbf{T}x_n\}$ has a subsequence convergent in Y (Reed and Simon 1981; Chapter 6).

not the case for conditional embedding. In Song et al. (2009; Theorem 6), the rate of convergence is $O_p((n\lambda)^{-1/2} + \lambda^{1/2})$, suggesting that the conditional mean embeddings are harder to estimate than the marginal embeddings, which converge at a rate $O_p(n^{-1/2})$. In Fukumizu (2015), under the condition that the eigenvalues $(\gamma_m)_{m=1}^{\infty}$ of $\mathcal{C}_{XX}$ decays as $\gamma_m \leq \beta m^{-b}$ for some $\beta > 0$, it is shown that the convergence rate is of $O_p(n^{-b/(4b+1)})$ with appropriate choice of regularization coefficient.

4.2 Regression Interpretation

As illustrated in Figure 4.1, the conditional mean embedding has a natural interpretation as a solution to vector-valued regression problem. This observation has been made in Zhang et al. (2011) and later thoroughly in Grünewälder et al. (2012), which we review below.

Recall that the conditional mean embedding is defined via $\mathbb{E}[g(Y)|X = \mathbf{x}] = \langle g, \hat{\mu}_{Y|\mathbf{x}} \rangle_{\mathscr{F}}$. That is, for every $\mathbf{x} \in \mathcal{X}$, $\hat{\mu}_{Y|\mathbf{x}}$ is a function on $\mathcal{Y}$ and thereby defines a mapping from $\mathcal{X}$ to $\mathscr{F}$. Furthermore, the empirical estimator in (4.8) can be expressed as $\hat{\mu}_{Y|\mathbf{x}} = \Phi(\mathbf{K} + n\lambda \mathbf{I}_n)^{-1}\mathbf{k}_{\mathbf{x}}$, which already suggests that the conditional mean embedding is the solution to an underlying regression problem. Given a sample $(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n) \in \mathcal{X} \times \mathscr{F}$, a vector-valued regression problem can be formulated as

$$\widehat{\mathcal{E}}_{\lambda}(f) = \sum_{i=1}^{n} \|\mathbf{z}_i - f(\mathbf{x}_i)\|_{\mathscr{F}}^2 + \lambda \|f\|_{\mathscr{H}_\Gamma}^2 \tag{4.9}$$

where $\mathscr{F}$ is a Hilbert space and $\mathscr{H}_\Gamma$ denotes a RKHS of vector-valued functions from $\mathcal{X}$ to $\mathscr{F}$ (see Micchelli and Pontil (2005) for more detail). Grünewälder et al. (2012) shows that $\hat{\mu}_{Y|X}$ can be obtained as a minimizer of the optimization of the form (4.9).[4]

Following the analysis of Grünewälder et al. (2012), a natural optimization problem for the conditional mean embedding is to find a

[4] In fact, a regression view of conditional mean embedding has already been noted very briefly in Song et al. (2009; Section 6) with connections to the solutions of Gaussian process regression (Rasmussen and Williams 2005) and kernel dependency estimation (Cortes et al. 2005). Nevertheless, Grünewälder et al. (2012) gives a more rigorous account of this perspective.

function $\mu : \mathcal{X} \to \mathscr{F}$ that minimizes the following objective:

$$\mathcal{E}[\mu] = \sup_{\|g\|_{\mathscr{F}} \leq 1} \mathbb{E}_X \left[(\mathbb{E}_Y[g(Y)|X] - \langle g, \mu(X) \rangle_{\mathscr{F}})^2 \right] \tag{4.10}$$

Unfortunately, we cannot estimate $\mathcal{E}[\mu]$ because we do not observe $\mathbb{E}_Y[g(Y)|X]$. Grünewälder et al. (2012) shows that $\mathcal{E}[\mu]$ can be upper bounded by a *surrogate loss function* given by

$$\mathcal{E}_s[\mu] = \mathbb{E}_{(X,Y)} \left[\|l(Y, \cdot) - \mu(X)\|^2_{\mathscr{F}} \right], \tag{4.11}$$

which can then be replaced by its empirical counterpart

$$\widehat{\mathcal{E}}_s[\mu] = \sum_{i=1}^{n} \|l(\mathbf{y}_i, \cdot) - \mu(\mathbf{x}_i)\|^2_{\mathscr{F}} + \lambda \|\mu\|^2_{\mathscr{H}_\Gamma}. \tag{4.12}$$

The regularization term is added to provide a well-posed problem and prevent overfitting.

It follows from Micchelli and Pontil (2005; Theorem 4) that the solution to the above optimization problem can be written as $\hat{\mu} = \sum_{i=1}^{n} \Gamma_{\mathbf{x}_i} c_i$ for some coefficients $\{c_i\}_{i \leq n}, c_i \in \mathscr{F}$. Note that the kernel Γ associated with $\mathscr{H}_\Gamma$ is an *operator-valued kernel* (Álvarez et al. 2012). Grünewälder et al. (2012) considers $\Gamma(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}')Id$ where $Id : \mathscr{F} \to \mathscr{F}$ is the identity map on $\mathscr{F}$. Under this particular choice of kernel, $c_i = \sum_{j \leq n} W_{ij} l(\mathbf{y}_i, \cdot)$ where $\mathbf{W} = (\mathbf{K} + \lambda \mathbf{I})^{-1}$ and $\hat{\mu} = \sum_{i=1}^{n} \Gamma_{\mathbf{x}_i} (\mathbf{K} + \lambda \mathbf{I})^{-1} l(\mathbf{y}_j, \cdot)$ which is exactly the embedding in (4.8). It remains an interesting question whether one can also employ a more general kernel $\Gamma(\mathbf{x}, \mathbf{x}')$ that is useful in practice.

The advantages of vector-valued regression interpretation of conditional mean embedding are two-fold. First, since we have a well-defined loss function, we can use cross-validation procedure for parameter or model selection, *e.g.*, λ. Second, it improves the performance analysis of conditional mean embedding as one has access a rich theory of vector-valued regression (Micchelli and Pontil 2005, Carmeli et al. 2006, Caponnetto and De Vito 2007, Caponnetto et al. 2008). In particular, by applying the convergence results of Caponnetto and De Vito (2007), Grünewälder et al. (2012) derive minimax convergence rate which are $O(\log(n)/n)$. Since the analysis is done under the assumption that $\mathscr{F}$ is finite dimensional, the conditional mean embedding

is simply the ridge regression of feature vectors, and thus the better rate—compared to the rate of $O(n^{-1/4})$ of Song et al. (2009)—is quite natural.

Based on the new interpretation, Grünewälder et al. (2012) also derives a sparse formulation of the conditional mean embedding. Moreover, one can construct different estimators of conditional mean embedding by introducing new regularizer in (4.12) (see, *e.g.*, Muandet et al. (2014b; Table 1)) It may be of interest to investigate theoretical properties of these new estimators. Lastly, it is instructive to point out that the regression interpretation of the conditional mean embedding can be considered as an instance of a *smooth operator* framework proposed later in Grünewälder et al. (2013).

4.3 Basic Operations: Sum, Product, and Bayes' Rules

In this section we review basic operations in probabilistic inference and show how they can be carried out in terms of kernel mean embeddings. Sum and product rules are elementary rules of probability. Unlike traditional recipe, the idea is to perform these operations directly on the marginal and conditional embeddings to obtain a new element in the RKHS which corresponds to the embedding of the resulting distribution. One of the advantages of this idea is that the product and sum rules can be performed without making any parametric assumptions on the respective distributions.

Formally, sum and product rules describing the relations between $\mathbb{P}(X)$, $\mathbb{P}(Y|X)$, and $\mathbb{P}(X,Y)$ are given as follow:

$$\text{Sum rule:} \qquad \mathbb{P}(X) = \sum_{Y} \mathbb{P}(X,Y) \tag{4.13}$$

$$\text{Product rule:} \qquad \mathbb{P}(X,Y) = \mathbb{P}(Y|X)\mathbb{P}(X) \tag{4.14}$$

Combining (4.13) and (4.14) yields a renowned Bayes' rule: $\mathbb{P}(Y|X) = \mathbb{P}(X|Y)\mathbb{P}(Y)/\mathbb{P}(X)$. In the continuous case, the sum in (4.13) turns into an integral. Sum and product rules are very fundamental in machine learning and statistics, so much so that nearly all of the probabilistic inference and learning, no matter how complicate they are, amount to repeated application of these two equations.

Next, we will show how these operations can be achieved as an algebraic manipulation of the (conditional) mean embedding in the RKHS. These results are due to Song et al. (2009; 2013), Fukumizu et al. (2013), and Schölkopf et al. (2015).

4.3.1 Kernel Sum Rule

Using the law of total expectation[5], we have $\mu_X = \mathbb{E}_X[\phi(X)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y]]$. Plugging in the conditional mean embedding yields

$$\mu_X = \mathbb{E}_Y[\mathcal{U}_{X|Y}\varphi(Y)] = \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y)] = \mathcal{U}_{X|Y}\mu_Y. \tag{4.15}$$

Theorem 4.1 provides sufficient conditions for (4.15) to be well-defined. Alternatively, we can also use a tensor product feature $\phi(\mathbf{x}) \otimes \phi(\mathbf{x})$ to define the kernel sum rule as

$$\mathcal{C}_{XX} = \mathcal{C}_{(XX)|Y}\mu_Y \tag{4.16}$$

where we used a conditional embedding operator $\mathbb{E}_{X|\mathbf{y}}[\phi(X) \otimes \phi(X)] = \mathcal{C}_{(XX)|Y}\phi(\mathbf{y})$ (Song et al. 2013).

Suppose we have an empirical estimate of μ_Y and $\mathcal{U}_{X|Y}$, *i.e.*, $\hat{\mu}_Y = \sum_{i=1}^m \alpha_i \varphi(\tilde{\mathbf{y}}_i)$ with a sample $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$, and $\widehat{\mathcal{U}}_{X|Y} = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}$ obtained from a sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ drawn i.i.d. from $\mathbb{P}(X, Y)$. Note that the samples $(\mathbf{y}_i)$ and $(\tilde{\mathbf{y}}_j)$ to estimate the covariance operators and the kernel mean $\hat{\mu}_Y$, respectively, can be different. Applying (4.7) to these estimates, the kernel sum rule can be expressed in terms of kernel matrices as

$$\hat{\mu}_X = \widehat{\mathcal{U}}_{X|Y}\hat{\mu}_Y = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}\hat{\mu}_Y = \Upsilon(\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha}. \tag{4.17}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$, $\mathbf{L}_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$, and $\tilde{\mathbf{L}}_{ij} = l(\mathbf{y}_i, \tilde{\mathbf{y}}_j)$. Note that we can write $\hat{\mu}_X = \sum_{j=1}^n \beta_j \phi(\mathbf{x}_j)$ with $\boldsymbol{\beta} = (\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha}$, which enables us to use it in subsequent operations.

4.3.2 Kernel Product Rule

Consider a tensor product of the joint feature map $\phi(X) \otimes \varphi(Y)$. We can then factorize $\mu_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ according to the law of

[5]The law of total expectation states that for any integrable random variable X and any random variable Y, $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]]$

total expectation as

$$\begin{aligned}\mathbb{E}_Y[\mathbb{E}_{X|Y}[\phi(X)|Y] \otimes \varphi(Y)] &= \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)] \\ \mathbb{E}_X[\mathbb{E}_{Y|X}[\varphi(Y)|X] \otimes \phi(X)] &= \mathcal{U}_{Y|X}\mathbb{E}_X[\phi(X) \otimes \phi(X)].\end{aligned}$$

Let $\mu_X^{\otimes} := \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\mu_Y^{\otimes} := \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$, *i.e.*, the kernel embeddings using $\phi(\mathbf{x}) \otimes \phi(\mathbf{x})$ and $\varphi(\mathbf{y}) \otimes \varphi(\mathbf{y})$ features, respectively. Then, we can write the product rule in terms of kernel mean embeddings as

$$\mu_{XY} = \mathcal{U}_{X|Y}\mu_Y^{\otimes} = \mathcal{U}_{Y|X}\mu_X^{\otimes}. \tag{4.18}$$

We can see that—unlike the sum rule in (4.15)—the input to the product rule in (4.18) is a tensor product $\mathbb{E}_X[\phi(X) \otimes \phi(X)]$. If we identify the tensor product space $\mathscr{H} \otimes \mathscr{F}$ with the Hilbert-Schmidt operators $\mathrm{HS}(\mathscr{H}, \mathscr{F})$ as in Section 3.2, we can rewrite (4.18) by

$$\mathcal{C}_{XY} = \mathcal{U}_{X|Y}\mathcal{C}_{YY} \quad \textit{and} \quad \mathcal{C}_{YX} = \mathcal{U}_{Y|X}\mathcal{C}_{XX}. \tag{4.19}$$

Assume that—for some $\boldsymbol{\alpha} \in \mathbb{R}^m$—the embedding $\hat{\mu}_X^{\otimes}$ and $\hat{\mu}_Y^{\otimes}$ are given respectively by

$$\begin{aligned}\sum_{i=1}^{m} \alpha_i \phi(\tilde{\mathbf{x}}_i) \otimes \phi(\tilde{\mathbf{x}}_i) &= \tilde{\Upsilon}\Lambda\tilde{\Upsilon}^\top, \\ \sum_{i=1}^{m} \alpha_i \varphi(\tilde{\mathbf{y}}_i) \otimes \varphi(\tilde{\mathbf{y}}_i) &= \tilde{\Phi}\Lambda\tilde{\Phi}^\top,\end{aligned}$$

where $\Lambda := \mathrm{diag}(\boldsymbol{\alpha})$, $\tilde{\Upsilon} = [\phi(\tilde{\mathbf{x}}_1), \ldots, \phi(\tilde{\mathbf{x}}_m)]^\top$, and $\tilde{\Phi} = [\varphi(\tilde{\mathbf{y}}_m), \ldots, \varphi(\tilde{\mathbf{y}}_m)]^\top$. Consequently, we obtain

$$\hat{\mu}_{XY} = \widehat{\mathcal{U}}_{X|Y}\hat{\mu}_Y^{\otimes} = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{YY}^{-1}\hat{\mu}_Y^{\otimes} = \Upsilon(\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\Lambda\tilde{\Phi}^\top. \tag{4.20}$$

The operation in term of $\hat{\mu}_X^{\otimes}$ has a similar expression.[6] Corresponding to (4.19), the formula (4.20) can be also expressed as

$$\widehat{\mathcal{C}}_{XY} = \widehat{\mathcal{U}}_{X|Y}\widehat{\mathcal{C}}_{YY} = \Upsilon(\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\Lambda\tilde{\Phi}^\top. \tag{4.21}$$

Note that we can also write $\hat{\mu}_{XY}$ or $\widehat{\mathcal{C}}_{XY}$ in terms of the coefficient matrix, *i.e.*, $\Upsilon\mathbf{B}\tilde{\Phi}^\top$ where $\mathbf{B} := (\mathbf{L} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\Lambda$.

[6] $\hat{\mu}_{XY} = \widehat{\mathcal{U}}_{Y|X}\hat{\mu}_X^{\otimes} = \widehat{\mathcal{C}}_{YX}\widehat{\mathcal{C}}_{XX}^{-1}\hat{\mu}_X^{\otimes} = \Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{K}}\Lambda\tilde{\Upsilon}^\top$

Although both (4.15) and (4.18) do not require any parametric assumption on the underlying distributions, these operations can sometimes be both statistically difficult and computationally costly in some applications (more below).

4.3.3 Kernel Bayes' Rule

The kernelization of Bayes' rule, called *kernel Bayes' rule* (KBR)—proposed in Fukumizu et al. (2013)—realizes Bayesian inference in completely nonparametric settings without any parametric models. The inference is thus not to obtain the posterior of parameters, which is often the purpose of Bayesian inference, but rather more general posterior of a variable given some observation of another variable. The likelihood and prior, which are the probabilistic ingredients of Bayes rule, are expressed in terms of covariance operators and kernels mean, respectively. Namely, the KBR provides a mathematical method for obtaining an embedding $\mu^{\Pi}_{Y|X}$ of posterior $\mathbb{P}(Y|X)$ from the embeddings of prior $\Pi(Y)$ and the likelihood $\mathbb{P}(X|Y)$. The kernel sum and product rules form the backbone of the kernel Bayes' rule. See Fukumizu et al. (2013) and Song et al. (2013) for more technical details.

Essentially, the embedding of the posterior $\mathbb{P}(Y|X = \mathbf{x})$ can be obtained as $\mu^{\Pi}_{Y|\mathbf{x}} = \mathcal{C}^{\Pi}_{Y|X}\phi(\mathbf{x})$ where $\mathcal{C}^{\Pi}_{Y|X}$ denotes the conditional mean embedding which depends on the prior distribution $\Pi(Y)$. Like classical conditional mean embedding, we may view $\mathcal{C}^{\Pi}_{Y|X}$ as an embedding of the whole posterior $\mathbb{P}(Y|X)$. The *kernel Bayes' rule* is given by

$$\mu^{\Pi}_{Y|\mathbf{x}} = \mathcal{C}^{\Pi}_{Y|X}\phi(\mathbf{x}) = \mathcal{C}^{\Pi}_{YX}(\mathcal{C}^{\Pi}_{XX})^{-1}\phi(\mathbf{x}) \tag{4.22}$$

where the operators $\mathcal{C}^{\Pi}_{YX}$ and $\mathcal{C}^{\Pi}_{XX}$ are given by

$$\mathcal{C}^{\Pi}_{YX} = (\mathcal{C}_{X|Y}\mathcal{C}^{\Pi}_{YY})^{\top}, \quad \mathcal{C}^{\Pi}_{XX} = \mathcal{C}_{(XX)|Y}\mu^{\Pi}_{Y},$$

where the former uses the kernel product rule (4.19), and the latter uses the kernel sum rule (4.16). The embeddings μ^{Π}_{Y} and $\mathcal{C}^{\Pi}_{YY}$ correspond to the embeddings of $\Pi(Y)$ using features $\varphi(\mathbf{y})$ and $\varphi(\mathbf{y})\otimes\varphi(\mathbf{y})$, respectively.

Let $\hat{\mu}^{\Pi}_{Y} = \sum_{i=1}^{m}\alpha_i\varphi(\tilde{\mathbf{y}}_i)$ and $\widehat{C}^{\Pi}_{YY} = \sum_{i=1}^{m}\alpha_i\varphi(\tilde{\mathbf{y}}_i)\otimes\varphi(\tilde{\mathbf{y}}_i)$ be the empirical estimates obtained from the weighted sample $\{\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_m\}$.

Then, the empirical counterpart of (4.22) can be written as

$$\begin{aligned}\hat{\mu}_{Y|\mathbf{x}} &= \widehat{\mathcal{C}}^{\Pi}_{YX}((\widehat{\mathcal{C}}^{\Pi}_{XX})^2 + \tilde{\lambda}\mathbf{I})^{-1}\widehat{\mathcal{C}}^{\Pi}_{XX}\phi(\mathbf{x}) \\ &= \tilde{\Phi}\Omega^{\top}((\mathbf{DK}) + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{KDk}_{\mathbf{x}},\end{aligned}$$

where $\Omega := (\mathbf{L} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\Lambda$ and $\mathbf{D} := \mathrm{diag}((\mathbf{L} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{L}}\boldsymbol{\alpha})$. In the above equation, the regularization is different from the standard one used in the conditional kernel mean or kernel sum rule. In essence, we use $(B^2 + \lambda I)^{-1}Bz$ to regularize $B^{-1}z$, instead of $(B + \lambda I)^{-1}z$ used in the previous discussions, since in the above case the estimator $\widehat{\mathcal{C}}^{\Pi}_{XX}$ may not be positive definite and thus $\widehat{\mathcal{C}}^{\Pi}_{XX} + \lambda I$ may not be invertible. Note also that we can write $\hat{\mu}_{Y|\mathbf{x}} = \sum_{i=1}^{m}\beta_i\varphi(\tilde{\mathbf{y}}_i)$ where $\boldsymbol{\beta} := \Omega^{\top}((\mathbf{DK}) + \tilde{\lambda}\mathbf{I})^{-1}\mathbf{KDk}_{\mathbf{x}}$.

For any function $g \in \mathscr{F}$, we can evaluate the expectation of g w.r.t. the posterior $\mathbb{P}(Y|\mathbf{x})$ by means of $\hat{\mu}_{Y|x}$ as $\langle g, \hat{\mu}_{Y|\mathbf{x}}\rangle_{\mathscr{F}}$. If we assume specifically that $g = \sum_{i=1}^{n}\alpha_i\varphi(\mathbf{y}_i)$ for some $\boldsymbol{\alpha} \in \mathbb{R}^n$, then $\langle g, \hat{\mu}_{Y|\mathbf{x}}\rangle_{\mathscr{F}} = \boldsymbol{\beta}^{\top}\tilde{\mathbf{L}}\boldsymbol{\alpha}$. Moreover, the parametric form of posterior $\mathbb{P}(Y|\mathbf{x})$ can be reconstructed from $\hat{\mu}_{Y|\mathbf{x}}$ using one of the distributional pre-image techniques discussed previously in §3.8.

Fukumizu et al. (2013) demonstrates consistency of the posterior mean $\hat{\mu}_{Y|\mathbf{x}}$ (and the posterior expectation $\langle f, \hat{\mu}_{Y|\mathbf{x}}\rangle$), *i.e.*, $\|\hat{\mu}_{Y|\mathbf{x}} - \mu_{Y|\mathbf{x}}\|_{\mathscr{F}} \to 0$ in probability as $n \to \infty$ (Fukumizu et al. 2013; Theorem 5), and shows that if $\|\hat{\mu}^{\Pi}_{Y} - \mu^{\Pi}_{Y}\|_{\mathscr{F}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 \leq \alpha \leq 1/2$, then—under certain specific assumptions—we have for any $\mathbf{x} \in \mathcal{X}$,

$$\langle f, \hat{\mu}_{Y|\mathbf{x}}\rangle_{\mathscr{F}} - \mathbb{E}[f(Y)|X = \mathbf{x}] = O_p(n^{-\frac{8}{27}\alpha})$$

as $n \to \infty$ (Fukumizu et al. 2013; Theorem 6). In $L^2(\mathbb{P})$, the rate improves slightly to $O_p(n^{-\frac{1}{3}\alpha})$ (Fukumizu et al. 2013; Theorem 7). Fukumizu et al. (2013) claims that—while these seem to be slow rates—the rate of convergence can in practice be much faster.

A common application of KBR is in a situation where computing the likelihood of observed data is intractable, while generating samples from the corresponding density is relatively easy. Kanagawa et al. (2016), for example, proposes a filtering method with KBR in the situation where the observation model must be estimated nonlinearly and nonparametrically with data.

As an aside, Song et al. (2013) also provides the formulations of sum, product, and Bayes' rules as a *multi-linear algebraic operation* using a tensor product feature $\phi(\mathbf{x}) \otimes \phi(\mathbf{x})$—instead of the standard feature $\phi(\mathbf{x})$.

4.3.4 Functional Operations on Mean Embeddings

Functional operations—*e.g.*, the multiplication and exponentiation—on random variables are quite common in certain domains such as probabilistic programming.[7] Given two independent random variables X and Y with values in $\mathcal{X}$ and $\mathcal{Y}$, and a measurable function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, we are interested in estimating the distribution of $Z = f(X, Y)$. Assume that we have the consistent estimates of $\mu[X]$ and $\mu[Y]$, *i.e.*,

$$\hat{\mu}[X] = \sum_{i=1}^{m} \alpha_i \phi_x(\mathbf{x}_i), \qquad \hat{\mu}[Y] = \sum_{j=1}^{n} \beta_j \phi_y(\mathbf{y}_i)$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are mutually independent i.i.d. samples from $\mathbb{P}(X)$ and $\mathbb{P}(Y)$, respectively. Note that the embedding of conditional random variables can also be written in this form (see the discussion following Theorem 4.2). Schölkopf et al. (2015) proposed to estimate $\mu[f(X, Y)]$ by the following estimator

$$\hat{\mu}[f(X, Y)] := \frac{1}{\sum_{i=1}^{m} \alpha_i \sum_{j=1}^{n} \beta_j} \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j \phi_z(f(\mathbf{x}_i, \mathbf{y}_j)). \tag{4.23}$$

and showed that (4.23) is a consistent estimate of $\mu[f(X, Y)]$ and the convergence happens at a rate $O_p(\sqrt{\sum_i \alpha_i^2} + \sqrt{\sum_j \beta_j^2})$, *i.e.*, see Schölkopf et al. (2015; Theorem 3). Note that the feature maps ϕ_x, ϕ_y, and ϕ_z may correspond to different kernels. The estimators and theoretical analysis can be extended to functional operations on a larger set of random variables.

As we can see from (4.23), the gain here is that the embedding of $f(X, Y)$ can be obtained directly from the embedding of X and Y without resorting to density estimation of $f(X, Y)$. Many well-established

[7] In probabilistic programming (PP), instead of representing probabilistic models by graphical models or Bayesian networks, it uses *programs* to represent the models which are more expressive and flexible (Gordon et al. 2014).

framework for arithmetic operations on independent random variables involves *integral transform* methods such as Fourier and Mellin transforms (Springer 1979) which only allow for some simple cases, *e.g.*, the sum, difference, product, or quotient of independent random variables. Moreover, the functional f is applicable on any domains—no matter how complicated they are—as long as useful positive definite kernels are introduced on such domains.

Remark 4.1. We provide crucial remarks on the conditional mean embedding.

i) To guarantee the convergence of the empirical conditional mean embedding to the population one, some old literatures use a strong condition such as $k(\mathbf{x}, \cdot) \in \mathcal{R}(\mathcal{C}_{XX})$ or $\mathbb{E}[g(Y)|X = \cdot] \in \mathscr{H}$ so that $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(\mathbf{x}, \cdot)$ can be well-defined. Such a condition, however, is often impractically strong: as noted in Song et al. (2009) and Fukumizu et al. (2013), Gaussian kernel do not satisfy such an assumption for a wide class of distributions. If X and Y are independent, for example, $\mathbb{E}[g(Y)|X = \mathbf{x}]$ is a constant function of $\mathbf{x}$, and as a result it is not included in the RKHS given by a Gaussian kernel (Steinwart and Christmann 2008; Corollary 4.44).

ii) To alleviate such a strong assumption, a regularized version $\mathcal{C}_{YX}(\mathcal{C}_{XX} + \lambda_n\mathcal{I})^{-1}k(\mathbf{x}, \cdot)$ is used instead. For a separable Hilbert space $\mathscr{H}$, it has been shown that $\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \lambda_n\mathcal{I})^{-1}k(\mathbf{x}, \cdot)$ converges to $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(\mathbf{x}, \cdot)$ if λ_n decays to zero sufficiently slowly as $n \to \infty$. In practice, a cross-validation procedure is commonly used to find the best λ_n for the problems at hand.

Below we review some applications that employ the above operations on kernel mean embedding.

4.4 Graphical Models and Probabilistic Inference

Conditional mean embedding has enjoyed successful applications in graphical models and probabilistic inference (Song et al. 2009; 2010a; 2011a;b; 2010b). Probabilistic graphical models are ubiquitous in many

fields including natural language processing, computational biology, computer vision, and social science. Most of the traditional algorithms for inference often specifies explicitly the parametric distributions underlying the observations and then applies basic operations such as sum, product, and Bayes rules on these distributions to obtain the posterior distribution over desired quantities, *e.g.*, parameters of the model. On the other hand, the philosophy behind embedding-based algorithms is to represent distributions by their mean embedding counterparts, and then to apply the operations given in Section 4.3 on these embeddings instead. This method leads to several advantages over the classical approach. First, an inference can be performed in a non-parametric fashion; one does not need a parametric assumption about the underlying distribution as well as prior-posterior conjugacy. Second, most algorithms do not require density estimation which is difficult in high-dimensional spaces (Wasserman 2006; Section 6.5). Lastly, many models are only restricted to deal with discrete latent variables, *e.g.*, a hidden Markov model (HMM) (Baum and Petrie 1966). The embedding approach allows for (possibly structured) non-Gaussian continuous variables, which makes these models applicable for a wider class of applications. Nevertheless, there are some disadvantages as well. First, relying on the kernel function, the resulting algorithms are usually sensitive to the choice of kernel and its parameters which needs to be chosen carefully. Second, the algorithms only have access to the embedding of posterior distribution rather than the distribution itself. Hence, to recover certain information such as the shape of the distribution, one need to resort to a pre-image problem to obtain an estimate of the full posterior distribution (cf. Section 3.8 and Song et al. (2008), Kanagawa and Fukumizu (2014), McCalman et al. (2013)) Lastly, the algorithms can become computationally costly. Many approximation techniques such as low-rank approximation is often used to speed up the computation time and to reduce memory storage. See also Song et al. (2013) for a unified view of nonparametric inference in graphical models with conditional mean embedding.

The conditional mean embedding was first introduced in Song et al. (2009) with application in dynamical systems. In dynamical systems,

one is interested in a joint distribution $\mathbb{P}(\mathbf{s}_1, \ldots, \mathbf{s}_T, \mathbf{o}_1, \ldots, \mathbf{o}_T)$ where $\mathbf{s}_t$ is the hidden state at timestep t and $\mathbf{o}_t$ is the corresponding observation. A common assumption is that a dynamical system follows a *partially observable* Markov model under which the joint distribution factorizes as

$$\mathbb{P}(\mathbf{o}_1, \mathbf{s}_1) \prod_t \mathbb{P}(\mathbf{o}_t|\mathbf{s}_t)\mathbb{P}(\mathbf{s}_t|\mathbf{s}_{t-1}).$$

Thus, the system is characterized by two important models, namely, a *transition model* $\mathbb{P}(\mathbf{s}_t|\mathbf{s}_{t-1})$ which describes the evolution of the system and the *observation model* $\mathbb{P}(\mathbf{o}_t|\mathbf{s}_t)$ which captures the uncertainty of a noisy measurement process. Song et al. (2009) focuses on *filtering* which aims to query the posterior distribution of state conditioned on all past observations, *i.e.*, $\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{h}_{t+1})$ where $\mathbf{h}_t = (\mathbf{o}_1, \ldots, \mathbf{o}_t)$. The distribution $\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{h}_{t+1})$ can be obtained in two steps. First, we *update* the distribution by

$$\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{h}_t) = \mathbb{E}_{\mathbf{s}_t|\mathbf{h}_t}[\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{s}_t)|\mathbf{h}_t].$$

Then, we *condition* the distribution on a new observation $\mathbf{o}_{t+1}$ using Bayes rule to obtain

$$\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{h}_t\mathbf{o}_{t+1}) \propto \mathbb{P}(\mathbf{o}_{t+1}|\mathbf{s}_{t+1})\mathbb{P}(\mathbf{s}_{t+1}|\mathbf{h}_t).$$

Song et al. (2009) propose the exact updates for prediction (Song et al. 2009; Theorem 7) and conditioning (Song et al. 2009; Theorem 8) which can be formulated entirely in terms of kernel mean embeddings. Despite the exact updates, one still needs to estimate the conditional cross-covariance operator in each conditioning step, which is both statistically difficult and computationally costly. This problem is alleviated by using approximate inference under some simplifying assumptions. See Song et al. (2009; Theorem 9) for technical detail. Empirically, although requiring labeled sequence of observations to perform filtering, for strongly nonlinear dynamics it has been shown to outperform standard Kalman filter which requires the exact knowledge of the dynamics. McCalman et al. (2013) also considers the filtering algorithm based on kernel mean embedding, *i.e.*, kernel Bayes rule (Fukumizu et al. 2011), to address the multi-modal nature of posterior distribution in robotics.

As mentioned earlier, one of the advantages of mean embedding approach in graphical models is that it allows us to deal with (possibly structured) non-Gaussian continuous variables. For example, Song et al. (2010a) extends *spectral algorithm* of Hsu et al. (2009) for learning traditional hidden Markov models (HMMs), which are restricted to discrete latent state and discrete observations, to structured and non-Gaussian continuous distributions (see also Jaeger (2000) for a formulation of discrete HMMs in terms of *observation operator*

$$\mathcal{O}_{ij} = \mathbb{P}(\mathbf{h}_{t+1} = i|\mathbf{h}_t = j)\mathbb{P}(X_t = \mathbf{x}_t|\mathbf{h}_t = j)).$$

In Hsu et al. (2009), HMM is learned by performing a singular value decomposition (SVD) on a matrix of joint probabilities of past and future observations. Song et al. (2010a) relies on the embeddings of the distributions over observations and latent states, and then construct an operator that represents the joint probabilities in the feature space. The advantage of spectral algorithm for learning HMMs is that there is no need to perform a local search when finding the distribution of observation sequences, which usually leads to more computationally efficient algorithms. Unlike Song et al. (2009), the algorithm only requires access to unlabeled sequence of observations.

A nonparametric representation of tree-structured graphical models was introduced in Song et al. (2010b). Inference in this kind of graphical models relies mostly on message passing algorithms. In case of discrete variable, or Gaussian distribution, the message passing can be carried out efficiently using the sum-product algorithm. Minka (2001) proposes the expectation-propagation (EP) algorithm which requires an estimation of only certain moments of the messages. Sudderth et al. (2010) considers messages as mixture of Gaussians. The drawback of this method is that the number of mixture components grows exponentially as the message is propagated. Ihler and McAllester (2009) considers a particle belief propagation (BP) where the messages are expressed as a function of a distribution of particles at each node. Unlike these algorithms, the embedding-based algorithm proposed in Song et al. (2010b) expresses the message $\mathbf{m}_{ts}(s)$ between pairs of nodes as RKHS functions on which sum and product steps can be performed using linear operation in RKHS to obtain a new message.

In addition, Song et al. (2010b) also proves the consistency of the conditional mean embedding estimator, *i.e.*, $\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\|_{\text{HS}}$ converges in probability under some reasonable assumptions (Song et al. 2010b; Theorem 1). The algorithm was applied in cross-lingual document retrieval and camera orientation recovery from images. The idea has been used later for latent tree graphical models (Song et al. 2011b), which are often used for expressing hierarchical dependencies among many variables in computer vision and natural language processing; and for belief propagation algorithm (Pearl 1988, Song et al. 2011a) for pairwise Markov random fields.

It is instructive to note that by assuming that the latent structure underlying the data-generating process has a low-rank structure, *e.g.*, latent tree, Song and Dai (2013) constructs an improved estimator of kernel mean embedding for multivariate distribution using truncated SVD (TSVD) algorithm.

In a similar manner, Sejdinovic et al. (2014) introduces a kernel adaptive Metropolis-Hasting algorithm for sampling from a target distribution with strongly non-linear support. When the target distributions are strongly non-linear, the classical samplers may suffer from low acceptance probability and slow mixing. The idea proposed in Sejdinovic et al. (2014) is to embed the trajectory of the Markov chain into the RKHS and then to construct the proposal using the information encoded in the empirical covariance operator. Intuitively, the empirical covariance operator contains the information about the non-linear support of the distribution. Hence, kernel PCA (Schölkopf et al. 1998) direction may be used to construct the proposal distribution. The proposed algorithm first obtains the RKHS sample

$$f = k(\mathbf{y}, \cdot) + \sum_{i=1}^{n} \beta_i [k(\mathbf{z}_i, \cdot) - \mu_{\mathbf{z}}]$$

from the Gaussian measure in the RKHS and then finds a point $\mathbf{x}^* \in \mathcal{X}$ whose canonical feature map $k(\mathbf{x}^*, \cdot)$ is close to f in an RKHS norm. Consequently, the resulting sampler is adaptive to the local structure of the target at the current chain state, is oriented towards nearby region of high density, does not suffer from wrongly scaled proposal distribution.

4.5 Markov Decision Processes and Reinforcement Learning

A Markov decision process (MDP) is a discrete time stochastic control which is widely studied and applied in robotics, automated control, economics, and manufacturing (Bellman 1957). See Figure 4.2 for a graphical illustration of MDP. Formally, an MDP is a 5-tuple $(\mathcal{S}, \mathcal{A}, P(\cdot,\cdot), R(\cdot,\cdot), \gamma)$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $P_{\mathbf{a}}(\mathbf{s}, \mathbf{s}')$ is a *state-transition probability* $P(\mathbf{s}_{t+1} = \mathbf{s}' \,|\, \mathbf{s}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a})$, $R_{\mathbf{a}}(\mathbf{s}, \mathbf{s}')$ is the immediate reward received after transition from state $\mathbf{s}$ to state $\mathbf{s}'$, and $\gamma \in [0, 1]$ is the discount factor. At each time step, the process is in some state $\mathbf{s}$, and we may choose any action $\mathbf{a}$ that is available in state $\mathbf{s}$. The process moves into the new state $\mathbf{s}'$ according to the state-transition function $P_{\mathbf{a}}(\mathbf{s}, \mathbf{s}')$ and gives a corresponding reward $R_{\mathbf{a}}(\mathbf{s}, \mathbf{s}')$. The discount factor γ represents the difference in importance between future rewards and present rewards. The goal of MDP is to obtain an optimal policy that maximizes a value function $V(\cdot)$ defined on beliefs (distributions over states), and determined by the reward function R, see, *e.g.*, Sutton and Barto (1998) for further detail.

Like in MDP, any problems that involve *control optimal theory* can be solved by analyzing the appropriate Bellman equation. According to Bellman's *principle of optimality*, an optimal policy has the property that whatever the intial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision (Bellman 2003). In MDP, this principle is reflected in the *Bellman optimality equation*:

$$(\mathcal{B}V)(\mathbf{x}) := \max_{\mathbf{a} \in \mathcal{A}} \left\{ r(\mathbf{x}, \mathbf{a}) + \gamma \mathbb{E}_{X \sim P(\cdot|\mathbf{x},\mathbf{a})}[V(X)] \right\}. \tag{4.24}$$

The operator $\mathcal{B}$ is known as *Bellman operator*. If the image of $\mathcal{B}$ is always a measurable function, $V_{t+1} = \mathcal{B}V_t$ converges in sup-norm.

The Bellman equation is in general difficult to solve. Traditional approaches for solving Bellman equation such as dynamic programming (Bellman 2003), parametric methods (Engel et al. 2003), Monte Carlo methods (Sutton and Barto 1998), and piecewise linear and convex (PWLC)-based methods (Smallwood and Sondik 1973) have several drawbacks. Dynamic programming is well-understood mathemati-

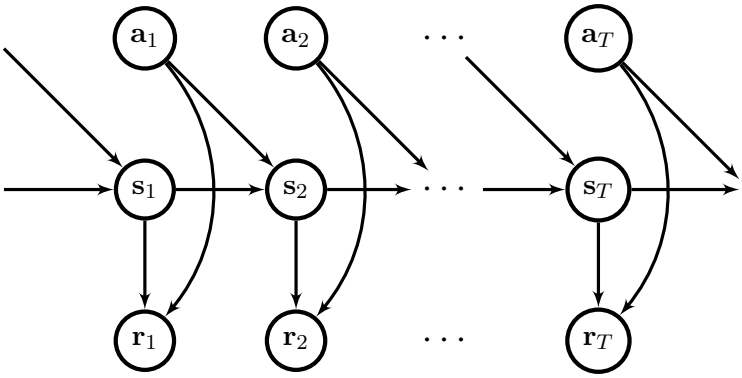


Figure 4.2: A graphical illustration of a Markov decision process. The state is dependent on the previous state and action, and the reward depends on the current state and action.

cal properties, but requires a complete and accurate model of the environment. Monte Carlo simulation is conceptually simple and requires no model, but can result in high-variance estimates. To overcome some of these difficulties, Grünewälder et al. (2012) proposes to learn transition dynamics, *i.e.*, value iteration and optimal policy learning, in MDPs by representing the stochastic transitions as conditional mean embeddings. That is, they are interested in the embedding of the expectation operator corresponding to the state transition probability kernel $\mathbb{P}$ such that $\langle \mu_{(\mathbf{x},\mathbf{a})}, f \rangle_{\mathscr{H}} = \mathbb{E}[f(X_{t+1}) \,|\, X_t = \mathbf{x}, A_t = \mathbf{a}]$. Specifically, the (empirical) Bellman operators $\mathcal{B}$ are first estimated using the kernel mean embedding:

$$(\mathcal{B}V)(\boldsymbol{\alpha}) := \max_{\mathbf{a} \in A} \left\{ \boldsymbol{\alpha}^\top \mathbf{R}_\mathbf{a} + \gamma \boldsymbol{\beta}^\top \mathbf{V}(\cdot) \right\}, \tag{4.25}$$

where $\mathbf{R}_\mathbf{a} \in \mathbb{R}^n$ is the reward vector on sample for action a and $\mathbf{V}(\cdot)$ denotes the vector of value function defined in terms of RKHS quantities (see Grünewälder et al. (2012), Nishiyama et al. (2012) for precise definitions). Then, these estimated operators are used in standard approach for solving MDPs. While the ordinary Bellman operator in (4.24) has *contractive* and *isotonic* properties which guarantee that the value iteration converges monotonically (Porta et al. 2006), the kernel Bellman operator in (4.25) may not have these two properties due to

the unnormalized coefficients of the conditional mean embedding (see also Theorem 4.2 and the following discussion). To alleviate this problem, Grünewälder et al. (2012) instead considers the normalized version of the conditional mean estimate and show that a certain contraction mapping for $\mathcal{B}$ can be achieved.

Nishiyama et al. (2012) extends the kernel-based algorithm of Grünewälder et al. (2012) to a partially observable MDP (POMDP). In POMDP, we do not assume that the state $\mathbf{s}$ is known when action $\mathbf{a}$ is to be taken. This approach also considers the distributions over states, observations, and actions, and employ kernel Bayes' rule (Fukumizu et al. 2013) to update the embeddings. The major contribution of Nishiyama et al. (2012) is to formulate POMDPs in feature space (kernel Bellman equation) and to propose a kernelized value iteration algorithm. However, the algorithm requires the labeled training data for the true latent states.

Boots et al. (2013) proposes to model *predictive state representations* (PSRs) using conditional mean embeddings. The PSR represents state as a set of predictions of future observable events which differ fundamentally from the latent variable model as it depends only on the observable quantities. For that reason, learning PSRs should be easier. The key idea of this work is to represent the state as nonparametric conditional mean embeddings in an RKHS. The benefit of mean embeddings here is that it allows for learning algorithms which work well for *continuous* actions and observations, unlike traditional algorithms which often run into trouble due to lack of data. Moreover, previous approaches often use kernel density estimation and exponential family, which suffer from slow rate of convergence and difficult numerical integration. The algorithm also updates the states using kernel Bayes' rule (Fukumizu et al. 2013). Lastly, it is claimed that although all prediction functions in the proposed algorithm are linear PSRs—*i.e.*, linear relationship between conditional probabilities of tests—it can still represent non-linear dynamic (Boots et al. 2013).

Finally, it should be emphasized that traditional approach for MDP and POMDP involves a conditional density estimation which is difficult for high dimensional problems. In addition, computation of high

dimensional integrals to obtain expectation can be very costly. On the other hand, the kernel mean embedding turns the expectation operator into an RKHS inner product, which has linear complexity in the number of training points. Moreover, the kernel mean estimates do not scale poorly with the dimension of the underlying space.

4.6 Conditional Dependency Measures

A task of determining the conditional dependency is ubiquitous in Bayesian network learning, gene expression, and causal discovery. Like an unconditional case, a joint distribution $\mathbb{P}(X, Y, Z)$ over variables X, Y, and Z satisfies a conditional independence relationship $X \perp\!\!\!\perp Y|Z$ if $\mathbb{P}(X, Y, Z)$ factorizes as $\mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)$. Several other equivalent characterizations of conditional independence are given in Dawid (1979), for example. Testing for conditional independence is generally a challenging problem due to the "curse of dimensionality" in terms of the dimensionality of the conditioning variable Z (Bergsma 2004).

There exist numerous tests for conditional dependency such as *partial correlation* tests (Baba et al. 2004), conditional densities based tests (Su and White 2008), and permutation-based tests (Tsamardinos and Borboudakis 2010). These classical tests requires either the parametric assumption of the underlying distribution, *e.g.*, Gaussianity, or linear relationship between random variables, or the discretization of the conditioning variable Z or all. Thus, they still suffer from the curse of dimensionality which makes their application domains quite limited. In what follows, we will focus on the kernel-based conditional dependency tests that have been proposed for non-linear and non-Gaussian data to remedy some of the aforementioned issues.

Similar to the idea of HSIC described in §3.6, Fukumizu et al. (2008) proposes a nonparametric conditional dependence measure based on the *normalized conditional cross-covariance operator*

$$\mathcal{V}_{YX|Z} := \mathcal{V}_{YX} - \mathcal{V}_{YZ}\mathcal{V}_{ZX} \tag{4.26}$$

where $\mathcal{V}_{YX}$ is a unique bounded operator such that $\mathcal{C}_{YX} = \mathcal{C}_{YY}^{1/2}\mathcal{V}_{YX}\mathcal{C}_{XX}^{1/2}$ (Baker 1973).[8] The operators $\mathcal{V}_{YZ}$ and $\mathcal{V}_{ZX}$ can be de-

[8] See also Theorem 3.3 for a brief discussion on the operator $\mathcal{V}_{YX}$.

fined in a similar way. Substituting $\mathcal{V}_{YX} = \mathcal{C}_{YY}^{-1/2}\mathcal{V}_{XY}\mathcal{C}_{XX}^{-1/2}$ (as well as $\mathcal{V}_{YZ}$ and $\mathcal{V}_{ZX}$) in (4.26) yields

$$\mathcal{V}_{YX|Z} = \mathcal{C}_{YY}^{-1/2}(\underbrace{\mathcal{C}_{YX} - \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\mathcal{C}_{ZX}}_{\mathcal{C}_{YX|Z}})\mathcal{C}_{XX}^{-1/2}. \tag{4.27}$$

Intuitively, the operator $\mathcal{C}_{YX|Z}$ can be viewed as a non-linear reminiscence of the *conditional covariane matrix* $\mathcal{C}_{YX|Z} = \mathcal{C}_{YX} - \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\mathcal{C}_{ZX}$ of Gaussian random variables. Given extended variables $\ddot{X} = (X, Z)$ and $\ddot{Y} = (Y, Z)$, they show that $\mathcal{C}_{\ddot{Y}\ddot{X}|Z} = \mathbf{0}$ if and only if X and Y are conditionally independent given Z, and propose the squared Hilbert-Schmidt norm $\|\mathcal{V}_{\ddot{Y}\ddot{X}|Z}\|_{\text{HS}}^2$ as a measure of conditional dependency and $\|\mathcal{V}_{YX}\|_{\text{HS}}^2$ as an unconditional counterpart. Note that $\mathcal{V}_{\ddot{Y}\ddot{X}|Z} = \mathbf{0}$ and $\mathcal{V}_{YX} = \mathbf{0}$ imply $\mathcal{C}_{YX} = \mathbf{0}$ and $\mathcal{C}_{\ddot{Y}\ddot{X}|Z} = \mathbf{0}$, respectively. Although mathematically rigorous, it is unknown how to analytically compute the null distribution of the test statistics. In Fukumizu et al. (2008), they resort to a bootstrapping approach which is computationally expensive. See Fukumizu et al. (2008; Section 2.2 and 2.3) for a rigorous treatment of this measure.

Zhang et al. (2011) resorts to the characterization based on *partial association* of Daudin (1980) which says that $X \perp\!\!\!\perp Y|Z$ if and only if $\mathbb{E}[fg] = 0$ for all suitable f chosen to be a function from Z to X and for all suitable g chosen to be a function from Z to Y, see, *e.g.*, Zhang et al. (2011; Lemma 2) for detail. This characterization is reminiscent of the partial correlation based characterization of conditional independence for Gaussian variables. Most importantly, Zhang et al. (2011) further derives the asymptotic distribution of the test under the null hypothesis, and provide ways to estimate such a distribution. In a similar vein, Flaxman et al. (2015a) adopts Gaussian process regression in conditional independence test on non-i.i.d. observations. First, the residual variables $\varepsilon_X, \varepsilon_Y, \varepsilon_Z$ are obtained by regressing from the latent variable T, *e.g.*, time or spatial location, to variables X, Y, and Z to remove their dependency on T. In the second step, the residuals ε_{XZ} and ε_{YZ} are obtained by regressing from Z to X and Y, respectively. Finally, the conditional independence $X \perp\!\!\!\perp Y|Z$ can be casted as an unconditional test between ε_{XZ} and ε_{YZ}.

Recall that we can view the unconditional independence testing as a two-sample testing between $\mathbb{P}(X,Y)$ and $\mathbb{P}(X)\mathbb{P}(Y)$, see, *e.g.*, (3.37). Likewise for the conditional independence, $X \perp\!\!\!\perp Y|Z$ holds if and only if $\mathbb{P}(X,Y,Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)$. Hence, we can in principle turn any unconditional independence tests into conditional ones. Doran (2013) extends this view to unconditional independence test using kernel mean embeddings. That is, given the i.i.d. sample $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^n$ from $\mathbb{P}(X,Y,Z)$, the sample from the corresponding $\mathbb{P}(X|Z)\mathbb{P}(Y|Z)\mathbb{P}(Z)$ under the null hypothesis can be obtained approximately as $\{(\mathbf{x}_i, \mathbf{y}_{\pi(i)}, \mathbf{z}_i\}_{i=1}^n$ where $\pi(\cdot)$ denotes a random permutation. Unlike in the unconditional case, Doran (2013) proposes to learn $\pi(\cdot)$ from the data in such a way that if $\pi(i) = j$, then $\mathbf{z}_i \approx \mathbf{z}_j$. To get the intuition, when Z is a discrete random variable, *i.e.*, $z \in \{1, 2, 3, \ldots, m\}$, we have that $X \perp\!\!\!\perp Y|Z$ if and only if $X \perp\!\!\!\perp Y|Z = i$ for all i. The latter can be achieved by performing an unconditional independence test in each bin using only $\{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^{n_i}$ whose z_k are the same.

4.7 Causal Discovery

Causal inference from observational data concerns how the distribution of effect variable would have changed if we were to intervene on one of the cause variables while keeping all others fixed (Pearl 2000). There has recently been a great deal of research and interest on causality in machine learning community (Schölkopf et al. 2012, Guyon 2013, Mooij et al. 2014, Guyon 2014). The kernel mean embedding has proven one of the successful mathematical tools in this research area.

Given a set of random variables $X = (X_1, X_2, \ldots, X_d)$, the goal of causal discovery is to uncover the underlying causal *directed acyclic graph* (DAG) denoted by $G(V, E)$. Each vertex $V_i \in V$ corresponds to the random variable X_i and an edge E_{ij} from V_i to V_j indicates a direct causal relationship from X_i to X_j (denoted by $X_i \to X_j$). The causal relationships in the DAG are usually parametrized by a *structural equation model* (SEM)

$$X_i \leftarrow f_i(\mathrm{Pa}(X_i), E_i)$$

where f_i is an arbitrary function and $\mathrm{Pa}(X_i)$ is a parental set of V_i. It

is often assumed that E_i are mutually independence, *i.e.*, there is no hidden confounder.

Under Markov and faithfulness assumptions (Pearl 2000), the PC algorithm (Spirtes et al. 2000) exploits conditional dependencies to recover the Markov equivalence class of G. Owing to this approach, several modern kernel-based conditional independence tests have been developed with this particular application in mind (Fukumizu et al. 2008, Zhang et al. 2011, Doran et al. 2014). Nevertheless, this approach is not suitable for bivariate case in which we observe only two random variables X and Y. In this case, Schölkopf et al. (2015) recently proposes a *kernel probabilistic programming* (KPP) which provides an expression of a functional of random variables, *e.g.*, $f(X, Y) = X \times Y$, by means of kernel mean embedding. In particular, they apply it to causal inference by embedding the SEM associated with the additive noise model (ANM) and constructing a test based on such embeddings (Schölkopf et al. 2015; Theorem 4). For a comprehensive review of research along this direction, see, *e.g.*, Mooij et al. (2014).

Inspired by the competition organized by Guyon (2013; 2014), Lopez-Paz et al. (2015) proposes a "data-driven" method for bivariate causal inference using kernel mean embedding. In contrast to previous works, this work assumes access to a large collection of cause-effect samples $\mathcal{S} = \{(S_i, l_i)\}_{i=1}^n$ where $S_i = \{(\mathbf{x}_{ij}, \mathbf{y}_{ij})\}_{j=1}^{n_i}$ is drawn from $\mathbb{P}(X_i, Y_i)$ and l_i is a label indicating the causal direction between X_i and Y_i. Lopez-Paz et al. (2015) avoids the handcrafted features of S_i by resorting to the kernel mean representation $\hat{\mu}[\hat{\mathbb{P}}(X_i, Y_i)]$ where $\hat{\mathbb{P}}(X_i, Y_i)$ is the empirical distribution of S_i. Then, inferring causal direction proceeds as a classification problem on distributions (Muandet et al. 2012).

A major challenge of most causal inference algorithms is the presence of hidden confounders. Sgouritsa et al. (2013) proposes the method to detect finite confounders based on the kernel mean embedding of the joint distribution $\mathbb{P}(X_1, X_2, \ldots, X_d)$. Specifically, they show that if the the joint distribution decomposes as $\mathbb{P}(X_1, X_2, \ldots, X_d) = \sum_{i=1}^m \mathbb{P}(Z_i) \prod_{j=1}^d \mathbb{P}(X_j|Z_i)$, then the tensor rank of the joint embedding

$$\mathcal{U}_{X_{[d]}} = \sum_{i=1}^{m} \mathbb{P}(Z_i) \otimes_{j=1}^{d} \mathbb{E}_{X_j}[\phi_j(X_j)|Z_i]$$

is exactly m (Sgouritsa et al. 2013; Theorem 1). This allows them to use clustering to identify confounders under various scenarios.

Recently, Schölkopf et al. (2012) postulates that under a specific causal assumption there is an asymmetry in the decomposition

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

In other words, if $X \to Y$, it is believed that $\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ are "independent", whereas in an anti-causal direction $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y)$ may contain information about each other. Owing to this postulate, Chen et al. (2014) defines the uncorrelatedness criterion between $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$ and formulate it in term of the complexity metric using Hilbert space embedding of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$.

5

Relationships between KME and Other Methods

In this section, we discuss the relationships between kernel mean embedding and other approaches ranging from kernel density estimation (KDE) to probabilistic models for Bayesian inference.

5.1 Beyond Density Estimation and Characteristic Function

Kernel density estimation has long been a popular method of choice for approximating the density function of the underlying distribution (Silverman 1986). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be a random sample from a distribution $\mathbb{P}$ with a density f_p. The kernel density estimate of this density is

$$\hat{f}_p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \kappa_\sigma(\mathbf{x}_i - \mathbf{x}) \tag{5.1}$$

where κ_σ satisfies $\int_{\mathcal{X}} \kappa_\sigma(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$ and $\kappa_\sigma(\mathbf{x}) \geq 0$ with bandwidth σ. Well-known examples of kernels satisfying all of the above properties are the Gaussian kernel, the multivariate Student kernel, and the Laplacian kernel. Anderson et al. (1994) constructs the two-sample test statistic using the L_2 distance between the kernel density estimates, *i.e.*, $\|\hat{f}_p - \hat{f}_q\|_2$ where $\hat{f}_p(\mathbf{x})$ and $\hat{f}_q(\mathbf{x})$ are the kernel density

estimates of densities p and q, respectively. More generally, we may define $D_r(p,q) = \|f_p - f_q\|_r$ as a distance between p and q, which subsumes seveal well-known distance measures such as Lévy distance ($r = 1$) and the Renyi entropy based measure ($r = 2$).

As shown in Gretton et al. (2012a), the L_2 distance between kernel density estimates $\hat{f}_p$ and $\hat{f}_q$ is a special case of the biased MMD in (3.31) between $\hat{\mu}_p$ and $\hat{\mu}_q$. That is, let $\hat{f}_p$ and $\hat{f}_q$ be given by $\hat{f}_p(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\kappa_\sigma(\mathbf{x}_i - \mathbf{x})$ and $\hat{f}_q(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\kappa_\sigma(\mathbf{y}_i - \mathbf{x})$, respectively. Then, the L_2 distance between $\hat{f}_p$ and $\hat{f}_q$ can be written as

$$\begin{aligned}
\|\hat{f}_p - \hat{f}_q\|_2^2 &= \int \left(\frac{1}{n}\sum_{i=1}^{n}\kappa_\sigma(\mathbf{x}_i - \mathbf{z}) - \frac{1}{m}\sum_{i=1}^{m}\kappa_\sigma(\mathbf{y}_i - \mathbf{z})\right) \mathrm{d}\mathbf{z} \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i - \mathbf{x}_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(\mathbf{y}_i - \mathbf{y}_j) \\
&\quad - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\mathbf{x}_i - \mathbf{y}_j) \\
&= \left\| \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i - \cdot) - \frac{1}{m}\sum_{i=1}^{m} k(\mathbf{y}_i - \cdot) \right\|_{\mathscr{H}_k}^2 \\
y &= \|\hat{\mu}_p - \hat{\mu}_q\|_{\mathscr{H}_k}^2
\end{aligned}$$

where $k(\mathbf{x}-\mathbf{y}) = \int \kappa_\sigma(\mathbf{x}-\mathbf{z})\kappa_\sigma(\mathbf{y}-\mathbf{z})\,\mathrm{d}\mathbf{z}$. By definition, $k(\mathbf{x}-\mathbf{y})$ is positive definite, and thereby is a reproducing kernel. As mentioned earlier in the survey, an RKHS-based approach provides advantages over the L_2 statistic in a number of important respects. Most notably, it bypasses the density estimation problem which is often considered more difficult especially in high dimensions. Similar approach used in estimating divergences using kernel density estimate (*e.g.*, L_2 divergence, Rényi, Kullback-Leibler, or many other divergences) avoids the need to estimate the density by adopting direct divergence approximation using, *e.g.*, density ratio estimation and regression.

Another popular approach for representing probability distribution is through *empirical characteristic function (ECF)* (Feuerverger and Mureika 1977, Alba Fernàndez et al. 2008). There is a one-to-one correspondence between characteristic functions and distributions. Moreover, it has been shown under some general restrictions

that the ECF converges uniformly almost surely to the population characteristic function. As discussed in §3.3.1, we may view both the kernel mean embedding and the characteristic function as *integral transforms* of the distribution. Kankainen and Ushakov (1998), for instance, proposes a consistent and asymptotically distribution-free test for independence based on the empirical characteristic function; see also Kankainen (1995). Unfortunately, ECF can be difficult to obtain in many cases, especially for conditional distributions.

5.2 Classical Kernel Machines

Kernel mean has long appeared in the literature since the beginning of the field of kernel methods itself (Schölkopf et al. 1998, Schölkopf and Smola 2001). Classical algorithms for classification and anomaly detection employed a mean function in the RKHS as their building block. Given a data set $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where $y_i \in \{-1, +1\}$, Shawe-Taylor and Cristianini (2004; Chapter 4), for example, considers a simple classifier that classifies a data point $\mathbf{x}_*$ by measuring the RKHS distance between $\phi(\mathbf{x}_*)$ and the class-conditional means in the feature space

$$\begin{aligned} \hat{\mu}_p &:= \frac{1}{|\{i \,|\, y_i = +1\}|} \sum_{y_i=+1} \phi(\mathbf{x}_i) \\ \hat{\mu}_n &:= \frac{1}{|\{i \,|\, y_i = -1\}|} \sum_{y_i=-1} \phi(\mathbf{x}_i). \end{aligned}$$

This algorithm is commonly known as a *Parzen window classifier* (Duda and Hart 1973). Likewise, anomaly detection algorithm can be obtained by constructing a high-confident region around the kernel mean $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)$ and consider points outside of this region as outliers. Although original works did not provide a link to the embedding of distributions, one can naturally interpret $\hat{\mu}_p$ and $\hat{\mu}_n$ as kernel mean embeddings of conditional distributions $\mathbb{P}(X|Y = +1)$ and $\mathbb{P}(X|Y = -1)$, respectively. Furthermore, Sriperumbudur et al. (2009) links the distance between kernel means (its MMD) to empirical risk minimization and large-margin principle in classification. Lastly, centering operation commonly used in many kernel algorithms involves

an estimation of the mean function in RKHS (Schölkopf and Smola 2001). As a result, reinterpreting these algorithms from distribution embedding perspective may shed light on novel methodologies of these learning algorithms.

5.3 Distance-based Two-Sample Test

The *energy distance* and *distance covariance* are among important classes of statistics used in two-sample and independence testing that have had a major impact in the statistics community. Sejdinovic et al. (2012; 2013b) shows that these statistics are in fact equivalent to distance between embedding of distributions with specific choice of kernels. The *energy distance* between probability distributions $\mathbb{P}$ and $\mathbb{Q}$ as proposed in Székely and Rizzo (2004; 2005) is given by

$$D_E(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{XY}\|X - Y\| - \mathbb{E}_{XX'}\|X - X'\| - \mathbb{E}_{YY'}\|Y - Y'\|, \quad (5.2)$$

where $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. The distance covariance was later introduced in Székely et al. (2007), Székely and Rizzo (2009) for independence test as a weighted L_2-distance between characteristic functions of the joint and product distributions. The *distance kernel* is a positive definite kernel obtained as $k(\mathbf{z}, \mathbf{z}') = \rho(\mathbf{z}, \mathbf{z}_0) + \rho(\mathbf{z}', \mathbf{z}_0) - \rho(\mathbf{z}, \mathbf{z}')$ where ρ is a semi-metric of negative type.[1] Sejdinovic et al. (2012) shows that the distance covariance, originally defined with Euclidean distance, can be generalized with a semi-metric of negative type. This generalized distance covariance is simply a special case of the HSIC with the corresponding distance kernel. In this respect, the kernel-based methods provides more flexible choice for dependence measures.

5.4 Fourier Optics

Harmeling et al. (2013) establishes a link between Fourier optics and kernel mean embedding from computer vision viewpoint. A simple

[1] A function ρ is said to be semi-metric if the "distance" function need not satisfy the triangle inequality. It is of negative type if it is also negative definite (see Definition 2 and 3 in Sejdinovic et al. (2012)).

imaging system can be described by the so-called *incoherent imaging equation*

$$q(\mathbf{u}) = \int f(\mathbf{u} - \boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \tag{5.3}$$

where both $q(\mathbf{u})$ and $p(\boldsymbol{\xi})$ describe image intensities. The function f represents the impulse response function, *i.e.*, the point spread function (PSF), of the imaging system. In this case, the image $p(\boldsymbol{\xi})$ induces, up to normalization, a probability measure which represents the light distribution of the object being imaged. The kernel $f(\mathbf{u}-\boldsymbol{\xi})$ in (5.3), which is shift-invariant, can be interpreted physically as the point response of an optical system. Based on this interpretation, Harmeling et al. (2013) asserts that the Fraunhofer diffraction is in fact a special case of kernel mean embedding and that in theory an object $p(\boldsymbol{\xi})$ with bounded support can be recovered completely from its diffraction-limited image, using an argument from the injectivity of mean embedding (Fukumizu et al. 2004, Sriperumbudur et al. 2008). In other words, the Fraunhofer diffraction does not destroy any information. A simple approach to compute the inversion in practice is also given in Harmeling et al. (2013).

5.5 Probabilistic Perspective

The kernel mean embedding can also be understood probabilistically. Let consider the following example.[2] Assume the data generating process $\mathbf{x} \sim \mathbb{P}$ and the GP model $f \sim \mathrm{GP}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. It follows that $\mathbb{E}[f(\mathbf{x}_i)] = 0$ and $\mathbb{E}[f(\mathbf{x}_i)f(\mathbf{x}_j)] = k(\mathbf{x}_i, \mathbf{x}_j)$. Consequently, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P}(\mathbf{x})\mathrm{GP}(f)}[f(\mathbf{x})f(\cdot)] &= \iint f(\mathbf{x})f(\cdot)\mathrm{GP}(f)p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}f \\ &= \int k(\mathbf{x}, \cdot)p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \mu_{\mathbb{P}}. \end{aligned}$$

In other words, the kernel mean can be viewed as an expected covariance of the functions induced by the GP prior whose covariance function is k. Note that unlike what we have seen so far the function

[2] This example was obtained independently via personal communication with Zoubin Gharamani.

f is drawn from a GP prior which is almost surely outside of $\mathscr{H}_k$. It turns out that this interpretation coincides with the one given in Shawe-Taylor and Dolia (2007). Specifically, let $\mathscr{H}$ be a set of functions and Π be a probability distribution over $\mathscr{H}$. Shawe-Taylor and Dolia (2007) defines the distance between two distributions $\mathbb{P}$ and $\mathbb{Q}$ as

$$D(\mathbb{P}, \mathbb{Q}) := \mathbb{E}_{f \sim \Pi(f)} \left| \mathbb{E}_X[f(X)] - \mathbb{E}_Y[f(Y)] \right|,$$

where $X \sim \mathbb{P}, Y \sim \mathbb{Q}$. That is, we compute the average distance between $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the distribution over *test function* f (see also Gretton et al. (2012a; Lemma 27, Section 7.5) for the connection to MMD). Nevertheless, a fully Bayesian interpretation of kernel mean embedding remains an open question.

6

Future Directions

We have seen that kernel mean embedding of distributions has made a considerable impact in machine learning and statistics. Nevertheless, some questions remain open and there are numerous potential research directions. In this section, we give suggestions on some of the directions we believe to be important.

6.1 Kernel Choice Problem

How to choose the "right" kernel function remains an ultimate problem in kernel methods and there is no exception to the kernel mean embedding. For some applications, one may argue that characteristic kernels should be sufficient as they has been shown theoretically to preserve all necessary information about the distributions. However, this property may not hold empirically because we only have access to finite samples. In which case, prior knowledge about the problem becomes more relevant in choosing the "right" kernel. It is also not trivial how to choose parameter values of such kernel. These issues have been addressed in Sriperumbudur et al. (2009), Gretton et al. (2012b), for example. Finally, understanding how to interpret the associated representation of

distributions is essential for several applications in statistics.

6.2 Stochastic Processes and Bayesian Interpretation

A definition of probability distribution is restricted to finite dimensional objects. Highly structured data can be taken into account using kernel, but it is not clear how to employ kernel mean embedding for *stochastic processes—i.e.*, distributions over infinite dimensional objects—in a similar manner to the distribution. See Chwialkowski and Gretton (2014) and Chwialkowski et al. (2014) for some preliminary results. This line of research has been quite fruitful in Bayesian nonparametrics. Hence, extending the kernel mean embedding to Bayesian nonparametric framework provides potential applications of probabilistic inference. It is also interesting to obtain a Bayesian interpretation of the kernel mean embedding. Having an elegant interpretation could potentially lead to several extensions of the previous works along the line of Bayesian inference, *e.g.*, Gaussian processes.

6.3 Scalability

In the era of "big data", it is imperative that modern learning algorithms are able to deal with increasingly complex and large-scale data. Recently, there has been a growing interest in developing large-scale kernel learning, which is probably inspired by the lack of theoretical insight of a deep neural network despite its success in various application domains. The advances along this direction will benefit the development of algorithms using kernel mean embedding. In the context of MMD, many recent works have addressed this issue (Zaremba et al. 2013, Cortes and Scott 2014, Ji Zhao 2015, Chwialkowski et al. 2015).

6.4 High-dimensional Inference

It has been observed that the improvement of shrinkage estimator tends to increase as the ambient dimensionality increases. Kernel-based methods are known to be less prone to the *curse of dimensionality* compared to classical approaches such as kernel density estimation, but little is

known about underlying theoretical insight. Better understanding of the role of ambient dimension in kernel methods may shed light in novel applications of kernel mean embedding in high-dimensional space. There are some recent works on high-dimensional analysis of MMD and related tests: see Ramdas et al. (2015) for example.

6.5 Causality

Causal inference involves the investigation of how the distribution of outcome changes as we intervene on some other variable. Several frameworks for causal inference using kernel mean embedding have been proposed. There have been some recent works in this direction (Zhang et al. 2011, Sgouritsa et al. 2013, Chen et al. 2014, Lopez-Paz et al. 2015), but it remains a challenging problem. For example, Lopez-Paz et al. (2015) considers bivariate causal inference as a classification task on the joint distributions of cause and effect variables. In potential outcome framework, the causal effect is defined as the difference between the distributions of outcome under *control* and *treatment* populations. Due to the *fundamental problem of causal inference*, either one of them would never be observed in practice. In this case, the question is whether we can use the kernel mean embedding to represent the counterfactual distribution.

6.6 Privacy-Preserving Embedding

Privacy is an essential requirement in medicine and finance—especially in the age of big data—because we now have access to an unprecedented amount of data of individuals that could compromise the integrity of their privacy. Unlike most parametric models, most applications of kernel mean embedding such as two sample testing and dependency measures often requires a direct access to individual data points which renders these algorithms vulnerable to adversarial attack. As a result, it is imperative to ask—to what extent—we can still achieve the same level of performance under an additional requirement that the privacy must be preserved. Another challenge in the context of kernel mean embedding is to preserve the privacy at the distributional level.

6.7 Statistics, Machine Learning, and Invariant Distribution Embeddings

Fundamental problems in statistics such as parameter estimation, hypothesis testing, and testing for model families, can be viewed as an estimation of the functional, *i.e.*, the estimator, which maps the empirical distribution to the value of the estimate. Classical approach to constructing such an estimator is to choose estimators which satisfy certain criteria, *e.g.*, maximum likelihood estimators. On the other hand, there have been ... That is, it is assumed that we have access to the training set $\{(\widehat{\mathbb{P}}_1(X), l_1), (\widehat{\mathbb{P}}_2(X), l_2), \ldots, (\widehat{\mathbb{P}}_n(X), l_n)\}$ where $\widehat{\mathbb{P}}_1(X)$ denotes an empirical distribution over a random variable X and l_i is the desired value of the estimate. For example, l_i may be the entropy computed from $\widehat{\mathbb{P}}$. Thus, finding an estimator in this case can be casted as a supervised learning problem on probability distributions. Using kernel mean embedding of $\widehat{\mathbb{P}}$, Szabó et al. (2015) provides theoretical analysis of the distribution regression framework and demonstrates the framework on entropy estimation problem. Flaxman et al. (2015b) employs distribution regression in ecological inference. Lastly, Lopez-Paz et al. (2015) casts bivariate causal inference problem as a classification problem on distributions. One of the future directions is to characterize the *learned* estimators and to understand their connection to most of the classical estimators in statistics.

Many statistical properties of a probability distribution are invariant to the input space on which it is defined. For example, independence implies $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ regardless of $\mathcal{X}$ and $\mathcal{Y}$. Therefore, there is a need to develop an invariant representation for distributions which will allow us to deal with such distributions simultaneously across different domains. This kind of knowledge is known as the *domain-general knowledge* in cognitive science (Goodman et al. 2011).

7

Conclusions

A kernel mean embedding of distributions has emerged as a powerful machinery for probabilistic modeling, statistical inference, machine learning, and causal discovery. The basic idea is to map distributions into a reproducing kernel Hilbert space (RKHS) where the whole arsenal of kernel methods can be extended to probability measures. It has given rise to a great deal of research and novel applications of positive definite kernels in both statistics and machine learning. In this survey, we gave a comprehensive review of existing works and recent advances in this research area. In the course of the survey, we also discussed some of the most challenging issues and open problems that could potentially lead to new research directions.

The survey began with a brief introduction to the reproducing kernel Hilbert space (RKHS) which forms the backbone of this survey (Section 2). As the name suggests, the kernel mean embedding owes its success to a powerful concept of a positive definite function commonly known as *kernel function*. Hence, we provided a short review of its theoretical properties and classical applications. Finally, we reviewed the Hilbert-Schmidt operators which are vital ingredients in modern applications of kernel mean embedding.

Next, we provided a thorough discussion of the Hilbert space embedding of marginal distributions, theoretical guarantees, and review of its applications (Section 3). To summarize, the Hilbert space embedding of distributions allows us to apply RKHS methods to probability measures. This extension prompts a wide range of applications such as kernel two-sample testing, independent testing, group anomaly detection, MCMC methods, predictive learning on distributions, and causal inference.

The survey then generalized the idea of embedding to conditional distributions, gave theoretical insights, and reviewed some applications (Section 4). The conditional mean embedding allows us to perform sum, product, and Bayes' rules which are ubiquitous in graphical model, probabilistic inference, and reinforcement learning, for example, in a non-parametric way. Furthermore, the conditional mean embedding admits a natural interpretation as a solution to vector-valued regression problem. Then, we drew some relationships between the kernel mean embedding and other related areas (Section 5). Lastly, we gave some suggestions on future research directions (Section 6).

We hope that this survey will become a useful reference for graduate students and researchers in machine learning and statistics who are interested in the theory and applications of kernel mean embedding of distributions.

Acknowledgement

We thank Anant Raj for reading the first draft of this survey and providing several helpful comments.

References

S. Achard, D. T. Pham, and C. Jutten. Quadratic dependence measure for nonlinear blind source separation. In *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, pages 263–268, 2003.

R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003. ISBN 0-12-044143-8.

R. P. Adams. *Kernel Methods for Nonparametric Bayesian Inference of Probability Densities and Point Processes*. PhD thesis, University of Cambridge, Cambridge, UK, 10/2009 2009.

V. Alba Fernàndez, M. Jimènez Gamero, and J. Muñoz Garcia. A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics & Data Analysis*, 52(7):3730–3748, 2008.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundation and Trends in Machine Learning*, 4(3): 195–266, Mar. 2012. ISSN 1935-8237. .

H. Anderson and M. Gupta. Expected kernel for missing features in support vector machines. In *Statistical Signal Processing Workshop*, pages 285–288, 2011.

N. H. Anderson, P. Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, July 1994.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657+, 2004.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In S. Shalev-Shwartz and I. Steinwart, editors, *COLT*, volume 30 of *JMLR Proceedings*, pages 185–209. JMLR.org, 2013.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions, 2015.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*, 2012.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003. ISSN 1532-4435.

C. R. Baker. Mutual information for Gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970. .

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:pp. 273–289, 1973. ISSN 00029947.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007. ISSN 0885-064X.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6): 1554–1563, 12 1966. .

F. Bavaud. On the Schoenberg transformations in data analysis: Theory and illustrations. *Journal of Classification*, 28(3):297–314, 2011.

R. Bellman. A Markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

R. E. Bellman. *Dynamic Programming*. Dover Publications, Incorporated, 2003. ISBN 0486428095.

J. Berger and R. Wolpert. Estimating the mean function of a Gaussian process and the Stein effect. *Journal of Multivariate Analysis*, 13(3):401–424, 1983.

W. P. Bergsma. Testing conditional independence for continuous random variables. *EURANDOM-report*, 2004.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

M. Besserve, N. K. Logothetis, and B. Schölkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems 26*, pages 2535–2543. Curran Associates, Inc., 2013.

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math Soc.*, 1943.

G. Biau and L. Györfi. On the asymptotic properties of a nonparametric l/sub 1/-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186. 2011.

G. Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):pp. 573–580, 1976.

A. Blum. Random projection, margins, kernels, and feature-selection. In C. Saunders, M. Grobelnik, S. R. Gunn, and J. Shawe-Taylor, editors, *SLSFS*, volume 3940 of *Lecture Notes in Computer Science*, pages 52–68. Springer, 2005. ISBN 3-540-34137-4.

S. Bochner. Monotone funktionen, Stieltjessche integrale und harmonische analyse. *Math. Ann.*, 108(1):378–410, 1933. ISSN 0025-5831. .

B. Boots, A. Gretton, and G. J. Gordon. Hilbert space embeddings of predictive state representations. In *Proc. 29th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. B. Blaschko. A low variance consistent test of relative dependency. In *Proceedings of the 32nd International Conference on Machine Learning*, JMLR Proceedings, pages 20–29, Lille, France, July 2015. JMLR.org.

W. Bounliphone, E. Belilovsky, M. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998. ISSN 1384-5810. .

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. ISSN 1615-3375. .

A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 04(04):377–408, 2006.

O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, pages 416–422. MIT Press, 2000.

Y. Chen, M. Welling, and A. J. Smola. Super-samples from kernel herding. In *UAI*, 2010.

Z. Chen, K. Zhang, L. Chan, and B. Schölkopf. Causal discovery via reproducing kernel Hilbert space embeddings. *Neural Computation*, 26(7): 1484–1517, 2014.

A. Christmann and I. Steinwart. Universal kernels on Non-Standard input spaces. In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414. 2010.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In E. P. Xing and T. Jebara, editors, *Proceedings of The 31st International Conference on Machine Learning*, volume 32, pages 1422–1430. JMLR, 2014.

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3608–3616, 2014.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representation of probability measures. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 1972–1980, 2015.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, Sept. 1995. ISSN 0885-6125. .

C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 153–160, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. .

E. C. Cortes and C. Scott. Scalable sparse approximation of a sample mean. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pages 5274–5278, 2014.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1), 2002.

M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

S. Danafar, P. M. V. Rancoita, T. Glasmachers, K. Whittingstall, and J. Schmidhuber. Testing hypotheses by regularized maximum mean discrepancy. 2013.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, Jan. 2003. ISSN 1042-9832. .

J. J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980. .

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):pp. 1–31, 1979.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2005. ISBN 0262541858.

E. De Vito, L. Rosasco, and R. Verri. Spectral methods for regularization in learning theory, 2006.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 551–556, New York, NY, USA, 2004.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley, 2000.

G. Doran. Distribution kernel methods for multiple-instance learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2012, Bellevue, Washington, USA*. AAAI Press, 2013. .

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *30th Conference on Uncertainty in Artificial Intelligence (UAI2014)*, 2014.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. A Wiley Interscience Publication. Wiley, 1973.

C. Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC'08, pages 1–19, Berlin, Heidelberg, 2008. Springer-Verlag.

G. K. Dziugaite, R. D. M., and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267, 2015.

Y. Engel, S. Mannor, and R. Meir. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In T. Fawcett and N. Mishra, editors, *ICML*, pages 154–161. AAAI Press, 2003.

S. M. A. Eslami, D. Tarlow, P. Kohli, and J. Winn. Just-in-time learning for fast and flexible inference. In *Advances in Neural Information Processing Systems 27*, pages 154–162. Curran Associates, Inc., 2014.

A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5(1):88–97, 01 1977. .

S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

S. R. Flaxman, D. B. Neill, and A. J. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Trans. Intell. Syst. Technol.*, 7(2), Nov. 2015a.

S. R. Flaxman, Y.-X. Wang, and A. J. Smola. Who supported obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 289–298, New York, NY, USA, 2015b. ACM.

G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-31716-0. Modern techniques and their applications, A Wiley-Interscience Publication.

K. Fukumizu. *Modern Methodology and Applications in Spatial-Temporal Modeling*, chapter Nonparametric Bayesian Inference with Kernel Mean Embedding, pages 1–24. Springer Japan, Tokyo, 2015. ISBN 978-4-431-55339-7.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems 20*, pages 489–496, Red Hook, NY, USA, 9 2008. Curran Associates, Inc.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *Ann. Statist.*, 37(4):1871–1905, 08 2009a.

K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In *Advances in neural information processing systems 21*, pages 473–480. Curran, 6 2009b.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1737–1745. 2011.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14: 3753–3783, 2013.

T. Gärtner. A survey of kernels for structured data. *SIGKDD Explor. Newsl.*, 5(1):49–58, July 2003. ISSN 1931-0145. .

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *In Proceeding of the 19th International Conference on Machine Learning (ICML)*, pages 179–186. Morgan Kaufmann, 2002.

M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2002.

A. Girard, C. E. Rasmussen, J. Q. Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems 15*, pages 529–536, 2002.

L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla. Mean map kernel methods for semisupervised cloud classification. *IEEE Transaction on Geoscience and Remote Sensing*, 48(1-1):207–220, 2010.

N. D. Goodman, T. D. Ullman, , and J. B. Tenenbaum. Learning a theory of causality. *Psychological review*, 2011.

A. D. Gordon, T. A. Henzinger, A. V. Nori, and S. K. Rajamani. Probabilistic programming. In *Proceedings of the on Future of Software Engineering*, FOSE 2014, pages 167–181, New York, NY, USA, 2014. ACM.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-schmidt norms. In *ALT*, ALT'05, pages 63–77. Springer-Verlag, 2005a.

A. Gretton, R. Herbrich, A. Smola, B. Schölkopf, and A. Hyvärinen. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. *Covariate Shift by Kernel Mean Matching*, pages 131–160. MIT Press, Cambridge, MA, USA, 2 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a.

A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1214–1222, 2012b.

S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *ICML*, 2012.

S. Grünewälder, G. Lever, A. Gretton, L. Baldassarre, S. Patterson, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

S. Grünewälder, A. Gretton, and J. Shawe-Taylor. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

J. Guevara, S. Canu, and R. Hirata. Support measure data description. Technical report, July 2014.

I. Guyon. Cause-effect pairs kaggle competition, 2013. URL `https://www.kaggle.com/c/cause-effect-pairs/`.

I. Guyon. Chalearn fast causation coefficient challenge, 2014. URL `https://www.codalab.org/competitions/1381`.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, Mar. 2003. ISSN 1532-4435.

J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. Curran Associates, Inc., 2007.

S. Harmeling, M. Hirsch, and B. Schölkopf. On a link between kernel mean maps and Fraunhofer diffraction, with an application to super-resolution beyond the diffraction limit. In *CVPR*, pages 1083–1090. IEEE, 2013.

D. Haussler. Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10, University of California at Santa Cruz, Santa Cruz, CA, USA, 1999.

M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max-Planck-Gesellschaft, July 2004.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability. In *Proceedings of The 12th International Conference on Artificial Intelligence and Statistics*, pages 136–143, 2005.

J. L. Hodges and E. L. Lehmann. The efficiency of some nonparametric competitors of the t-test. *Ann. Math. Statist.*, 27(2):324–335, 06 1956. .

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948. .

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 06 2008. .

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24, 1933.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 601–608, Cambridge, MA, USA, 9 2007. MIT Press.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964. .

F. Huszar and D. K. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In N. de Freitas and K. P. Murphy, editors, *UAI*, pages 377–386. AUAI Press, 2012.

A. Ihler and D. McAllester. Particle belief propagation. *International Conference on Artificial Intelligence and Statistics*, 5:256–263, 2009.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *In Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1998.

H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

W. James and J. Stein. Estimation with quadratic loss. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. University of California Press, 1961.

D. Janzing, E. Sgouritsa, O. Stegle, J. Peters, and B. Schölkopf. Detecting low-complexity unobserved causes. In *UAI*, pages 383–391. AUAI Press, 2011.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Supplement to: Quantifying causal influences. *The Annals of Statistics*, 2013.

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

S. Jegelka, A. Gretton, B. Schölkopf, B. Sriperumbudur, and U. von Luxburg. Generalized clustering via kernel embeddings. In B. Mertsching, M. Hund, and Z. Aziz, editors, *KI 2009: AI and Automation, Lecture Notes in Computer Science, Vol. 5803*, pages 144–152, Berlin, Germany, 9 2009. Max-Planck-Gesellschaft, Springer. .

D. M. Ji Zhao. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, June 2015.

W. Jitkrittum, A. Gretton, N. Heess, S. M. A. Eslami, B. Lakshminarayanan, D. Sejdinovic, and Z. Szabo. Kernel-based just-in-time learning for passing expectation propagation messages. In *UAI2015*, pages 405–414. AUAI Press, 2015.

M. Kanagawa and K. Fukumizu. Recovering distributions from Gaussian RKHS embeddings. In *Artificial Intelligence and Statistics (AISTATS)*, pages 457–465. JMLR, 2014.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Kernel Monte Carlo filter. Master's thesis, 2013.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Filtering with state-observation examples via kernel Monte Carlo filter. *Neural Computation*, 28(2):382–444, 2016.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*. PhD thesis, University of Jyvaskyla, 1995.

A. Kankainen and N. G. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciences*, 89(5):1486–1494, 1998. ISSN 1573-8795. .

P. Kar and H. Karnick. Random feature maps for dot product kernels. In N. D. Lawrence and M. Girolami, editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 583–591. JMLR.org, 2012.

J. Kim and C. D. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13:2529–2565, Sep 2012.

K. Kim, M. Franz, and B. Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1351–1366, 9 2005. .

R. Kondor and T. Jebara. A kernel between sets of vectors. In T. Fawcett and N. Mishra, editors, *ICML*, pages 361–368. AAAI Press, 2003.

S. Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012. ISSN 1935-8237. .

J. T. Y. Kwok and I. W. H. Tsang. The pre-image problem in kernel methods. *Neural Networks, IEEE Transactions on*, 15(6):1517–1525, Nov. 2004. .

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38, May 2015.

Q. Le, T. Sarlos, and A. Smola. Fastfood–approximating kernel expansions in loglinear time. In *30th International Conference on Machine Learning (ICML)*, 2013.

Y. Li, K. Swersky, and R. S. Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML2015)*, volume 37, pages 1718–1727. JMLR.org, 2015.

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning (ICML2015)*, 2015.

A. Mandelbaum and L. A. Shepp. Admissibility as a touchstone. *Annals of Statistics*, 15(1):252–268, 1987.

A. F. T. Martins, N. A. Smith, E. P. Xing, P. M. Q. Aguiar, and M. A. T. Figueiredo. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

L. McCalman, S. O'Callaghan, and F. Ramos. Multi-modal estimation with kernel embeddings for learning motion models. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2845–2852, May 2013. .

R. E. Megginson. *An Introduction to Banach Space Theory*. Springer-Verlag New York, Inc., 1998.

N. A. Mehta and A. G. Gray. Generative and latent mean map kernels. *CoRR*, abs/1005.0188, 2010.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209 (441-458):415–446, 1909. ISSN 0264-3952. .

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X, 9780262018258.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Proceedings of Advances in Neural Information Processing Systems*. MIT Press, 2004.

K. Muandet. *From Points to Probability Measures: Statistical learning on distributions with kernel mean embedding*. PhD thesis, Department of Computer Science, Tübingen University, 2015.

K. Muandet and B. Schölkopf. One-class support measure machines for group anomaly detection. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2013.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18. 2012.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel mean estimation and Stein effect. In E. P. Xing and T. Jebara, editors, *Volume 32: Proceedings of The 31st International Conference on Machine Learning*, pages 10–18. JMLR, 2014a.

K. Muandet, B. Sriperumbudur, and B. Schölkopf. Kernel mean estimation via spectral filtering. In *Advances in Neural Information Processing Systems 27*, pages 10–18. Curran Associates, Inc., 2014b.

K. Muandet, B. Sriperumburdur, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 2016.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):pp. 429–443, 1997. ISSN 00018678.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, 2007.

Y. Nishiyama and K. Fukumizu. Characteristic kernels and infinitely divisible distributions. 2014.

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert space embeddings of POMDPs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 644–653, 2012.

J. B. Oliva, B. Póczos, and J. G. Schneider. Distribution to distribution regression. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR Proceedings*, pages 1049–1057. JMLR.org, 2013.

J. B. Oliva, B. Poczos, and J. Schneider. Fast distribution to real regression. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *JMLR Proceedings*, pages 706–714. JMLR.org, 2014.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22 (2):199–210, 2011.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with infinite dimensional summary statistics via kernel embeddings. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 239–247, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7. .

B. Póczos, L. Xiong, and J. G. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 599–608, 2011.

B. Póczos, A. Singh, A. Rinaldo, and L. A. Wasserman. Distribution-free distribution regression. In *AISTATS*, volume 31 of *JMLR Proceedings*, pages 507–515. JMLR.org, 2013.

J. M. Porta, N. Vlassis, M. T. Spaan, and P. Poupart. Point-based value iteration for continuous POMDPs. *Journal of Machine Learning Research*, 7:2329–2367, Dec. 2006. ISSN 1532-4435.

N. Privault and A. Réveillac. Stein estimation for the drift of Gaussian processes using the Malliavin calculus. *Annals of Statistics*, 36(5):2531–2550, 2008.

N. Quadrianto, L. Song, and A. J. Smola. Kernelized sorting. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1289–1296. Curran Associates, Inc., 2009.

M. H. Quang, M. S. Biagio, and V. Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 388–396. Curran Associates, Inc., 2014.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS 2007 - Advances in Neural Information Processing Systems*, Dec. 2007.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. A. Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *CoRR*, abs/1508.00655, 2015. URL `http://arxiv.org/abs/1508.00655`.

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, pages 489–496. MIT Press, 2002. ISBN 0-262-02550-7.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

S. Reddi, A. Ramdas, B. Poczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, AISTATS (2015), pages 772–780, 2015.

M. Reed and B. Simon. *I: Functional Analysis, Volume 1 (Methods of Modern Mathematical Physics)*. Academic Press, 1 edition, 1981. ISBN 0125850506.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959.

L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Volume 1. , Foundations*. Cambridge mathematical library. Cambridge University Press, Cambridge, U.K., New York, 2000a. ISBN 0-521-77594-9.

L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000b. ISBN 0-521-77593-0.

L. Rosasco, M. Belkin, and E. D. Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, second edition, 1991.

I. J. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811–841, 1938.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42343-5.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *ICML*, 2012.

B. Schölkopf, K. Muandet, K. Fukumizu, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 2015.

D. Sejdinovic, A. Gretton, B. Sriperumbudur, and K. Fukumizu. Hypothesis testing using pairwise distances and associated kernels. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning*, pages 1111–1118, New York, NY, USA, 2012. Omnipress.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems 26*, pages 1124–1132. Curran Associates, Inc., 2013a.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 10 2013b. .

D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton. Kernel adaptive Metropolis-Hastings, Feb. 2014.

R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 1981.

E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In A. Nicholson and P. Smyth, editors, *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 556–565, Oregon, USA, 2013. AUAI Press Corvallis.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

J. Shawe-Taylor and A. N. Dolia. A framework for probability density estimation. In M. Meila and X. Shen, editors, *AISTATS*, volume 2 of *JMLR Proceedings*, pages 468–475. JMLR.org, 2007.

P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

L. Song. *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, The University of Sydney, 2008.

L. Song and B. Dai. Robust low rank kernel embeddings of multivariate distributions. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3228–3236. 2013.

L. Song, A. Smola, A. Gretton, and K. M. Borgwardt. A dependence maximization view of clustering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 815–822, New York, NY, USA, 2007a. ACM. ISBN 978-1-59593-793-3. .

L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, ICML '07, pages 823–830, New York, NY, USA, 2007b. ACM.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 992–999, 2008.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, June 2009.

L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. J. Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010a.

L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models via kernel embeddings. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2010b.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, May 2011a.

L. Song, A. P. Parikh, and E. P. Xing. Kernel embeddings of latent tree graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2708–2716, 2011b.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2000.

M. D. Springer. *The Algebra of Random Variables*. John Wiley & Sons, 1979.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 08 2016. .

B. Sriperumbudur and Z. Szabo. Optimal rates for random fourier features. In *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc., 2015.

B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758, Red Hook, NY, USA, 2009. Curran.

B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. .

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *The 21st Annual Conference on Learning Theory (COLT)*, 2008.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, July 2011a. ISSN 1532-4435.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *NIPS*, pages 1773–1781, 2011b.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, 1955.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, Mar. 2002a.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002b. ISSN 1532-4435. .

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012. ISSN 0176-4276. .

L. Su and H. White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24:829–864, 8 2008.

E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. *Communications of the ACM*, 53(10): 95–103, 2010. ISSN 0001-0782. .

D. J. Sutherland and J. G. Schneider. On the error of random fourier features. In *UAI*, pages 862–871. AUAI Press, 2015. ISBN 978-0-9966431-0-8.

R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, AISTATS (2015), pages 948–957, 2015.

G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimensions. *InterStat*, 2004.

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58 – 80, 2005. ISSN 0047-259X. .

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *Ann. Appl. Stat.*, 3(4):1236–1265, 12 2009. .

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 12 2007. .

I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings, 2016.

I. Tsamardinos and G. Borboudakis. Permutation testing improves Bayesian network learning. In *ECML/PKDD (3)*, volume 6323 of *Lecture Notes in Computer Science*, pages 322–337. Springer, 2010.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer Verlag, New York, 2nd edition, 2000.

L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010. ISBN 1441923225, 9781441923226.

L. Wehbe and A. Ramdas. Nonparametric independence testing for small sample sizes. In *Proceedings of the 2015 International Joint Conference on Artificial Intelligence (IJCAI2015)*, pages 3777–3783, 2015.

M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1121–1128, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-516-1. .

M. Welling. Herding dynamic weights for partially observed random field models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 599–606, Arlington, Virginia, United States, 2009b. AUAI Press. ISBN 978-0-9749039-5-8.

M. Welling and Y. Chen. Statistical inference using weak chaos and infinite memory. In *Proceedings of the International Workshop on Statistical-Mechanical Informatics*, 2010.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

S. Wu and S.-I. Amari. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. *Neural Processing Letters*, 15(1):59–67, 2002.

L. Xiong, B. Poczos, and J. Schneider. Group anomaly detection using flexible genre models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011a.

L. Xiong, B. Póczos, J. G. Schneider, A. Connolly, and J. VanderPlas. Hierarchical probabilistic models for group anomaly detection. *Journal of Machine Learning Research - Proceedings Track*, 15:789–797, 2011b.

J. Yang, V. Sindhwani, H. Avron, and M. W. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 485–493, 2014.

Y. Yoshikawa, T. Iwata, and H. Sawada. Latent support measure machines for bag-of-words data classification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1961–1969. Curran Associates, Inc., 2014.

Y. Yoshikawa, T. Iwata, and H. Sawada. Non-linear regression for bag-of-words data via gaussian process latent variable set model. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 3129–3135, 2015.

W. Zaremba, A. Gretton, and M. B. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS*, pages 755–763, 2013.

H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In F. G. Cozman and A. Pfeffer, editors, *UAI*, pages 804–813. AUAI Press, 2011. ISBN 978-0-9749039-7-2.

X. Zhang, L. Song, A. Gretton, and A. Smola. Kernel measures of independence for Non-IID data. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In J. Shawe-Taylor and Y. Singer, editors, *COLT*, volume 3120 of *Lecture Notes in Computer Science*, pages 594–608. Springer, 2004.