From Points to Probability Measures

Statistical Learning on Distributions with Kernel Mean Embedding

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
KRIKAMOL MUANDET
aus Songkhla/Thailand

Tübingen
2015

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	30 September 2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Bernhard Schölkopf
2. Berichterstatter:	Prof. Dr. Andreas Schilling
3. Berichterstatter, falls zutreffend	

To mankind

Abstract

The dissertation presents a novel kernel-based learning framework on probability measures which has abundant real-world applications. In classical setup, it is assumed that the data are points in a vector space that have been drawn independent and identically (i.i.d.) from some unknown distribution. In many scenarios, however, representing these data as distributions over such a vector space may be more preferable. For instance, when the measurement is noisy, we may incorporate the uncertainty by treating the data themselves as distributions. This is often the case for microarray data and astronomical data where the measurement process is imprecise. In order to obtain reliable data, the measurement or the experiment has to be replicated which is often costly and time consuming. Moreover, distributions not only embody individual data points, but also contain information about their interactions which can be beneficial for structural learning in fields such as high-energy physics, cosmology, and causality. Lastly, classical problems in statistics such as statistical estimation, hypothesis testing, and causal inference, may be interpreted in a decision-theoretic sense as learning a function that maps empirical distributions to the desired statistics, which is in contrast to standard estimation based on "plug-in" estimators. Rephrasing these problems in this way leads to novel approach for statistical inference and statistical estimation. Hence, allowing learning algorithms to operate directly on distributions prompts a wide range of future applications for machine learning.

To work with distributions, the key methodology adopted in this thesis is the kernel mean embedding of distributions that represents each distribution as a function in a reproducing kernel Hilbert space (RKHS). Successful applications of kernel mean embedding in the literature suggest that it is a powerful representation of distributions. Due to the dependence on the kernel function, it is adaptable to any domains and is eligible to the whole arsenal of kernel methods. Moreover, we can model the distribution underlying the data without making any parametric assumption. Finally, its simplicity eases theoretical analysis and lends itself to good computational efficiency. These characteristics render kernel mean embedding increasingly appealing in the community compared to existing approaches based on density estimation, divergence measures, and information geometry, for example. In particular, the kernel mean embedding has been applied successfully in two-sample testing, graphical model, and probabilistic inference. On the other hand, this thesis will focus mainly on the predictive learning on distributions, *i.e.*, when the observations are distributions and the goal is to make prediction about the previously unseen distributions. More importantly, the thesis investigates kernel mean estimation which is one of the most fundamental problems of kernel methods.

The dissertation begins with the introduction into foundation of kernel methods and literature review of applications of kernel mean embedding in the past few years. Then, it presents the kernel mean estimation problem. 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. A new class of estimators called kernel mean shrinkage estimators (KMSEs) that improve upon the standard kernel mean estimator is proposed. Owing to the kernel mean embedding and its estimators, the subsequent two chapters then present the learning framework on probability measures. In these chapters, I argue that many problems in machine learning and statistics can be formulated as a learning problem on distributions with some concrete examples such as group anomaly detection and domain generalization problems. In particular, the thesis provides an extension of well-known support vector machine (SVM) to a space of probability

distributions which we call a support measure machine (SMM). The presented applications not only demonstrate the benefits of the proposed framework, but also reveal its limitations which could potentially lead to new research directions.

To conclude, I found that representing data as distributions and learning from them can improve the performance of learning systems in certain applications. Probability distributions, as opposed to data points, constitute high-level information about aggregate behavior of the data, how the underlying process evolves over time and environments, or a complex concept that cannot be described merely by individual points. Since most intelligent organisms have the ability to recognize and exploit such information naturally, I believe that insights obtained from the theoretical and experimental results in this thesis may shed light on future development of intelligent machines, and most importantly, may provide clues on the true meaning of intelligence.

Acknowledgments

First and foremost, I want to thank my advisor Prof. Bernhard Schölkopf who is, and always has been, much more than just a Ph.D. advisor. He is a great mentor. His guidance is like a compass that shows me the way, yet I am free to choose my own path. I could not come this far without his support and guidance. He is undoubtedly a profound scientist from whom I have inherited a unique way of thinking. It is no exaggeration to say that exposing to his way of thinking is somewhat analogous to watching the *Inception* for the first time. He is an active colleague with whom I really enjoy collaborating. His views and perspectives have always proven valuable not only for our works, but for the the whole research community. Especially, I want to thank for his belief in me for co-organizing the *Empirical Inference Symposium* in honour of Vladimir Vapnik's 75th birthday. Both Vladimir and Bernhard are among very few people who are the reason I get into machine learning. It is incredibly honour that, at least once in my lifetime, I get to meet both of them in person, not to mention working directly with one. Last but not least, I want to thank him for his generosity as a friend and for all the good times we share together. I have to apologize, however, for my terrible skills at *Age of Empire*.

My Ph.D. works would have been impossible without collaborations from incredible colleagues. I want to thank Kenji Fukumizu who has been there since the beginning until the end of my Ph.D. He has been very influential to how I think about research. It was also an honour to visit his research group at the Institute of Statistical Mathematics in Japan which I am thankful for his hospitality. I want to thank Bharath Sriperumbudur who tirelessly put a tremendous effort into our works, yet continue to enjoy the collaboration. I am very impressed by his productivity. Finally, I want to thank Arthur Gretton for fruitful discussions and valuable suggestions. He is the one who always provide novel and complementary angles to our works. It was a pleasant experience working with them and I am really looking forward to our future collaboration.

I had opportunities to visit several laboratories and research groups during my Ph.D. First of all, I want to thank David Hogg for his hospitality while I was visiting the Center for Cosmology and Particle Physics (CCPP) at New York University. I want to thank Rebecca Oppenheimer for letting me join one of the observing runs at Palomar Observatory in San Diego. I thank Rob Fergus for an enjoyable collaboration and for his hospitality. I also thank Ingo Steinwart for the invitation to give a talk at his group in Stuttgart. I thank all collaborators such as David Balduzzi, Kun Zhang, Francesco Dinuzzo, *etc*. I also want to thank fellow postdocs and Ph.D. students David Lopez-Paz, Gary Doran, Ilya Tolstikhin, and many more. I am grateful for their productive collaborations. Needless to say, my former supervisors, Yee Whye Teh, John Shawe-Taylor, and Sanparith Marukatat have all contributed in a way to this achievement.

I thank former and current members of Max Planck Institute for Intelligent Systems, especially those from empirical inference department, with whom I share good times such as movie nights and ski trips together. My life as a graduate student would have been boring without them. I want to thank members of Empirical Inference Journal Club (EIJC) for their contributions to make it an enjoyable and productive reading group. I want to thank Sabrina Rehbaum for helping me out with so many administrative stuffs and most importantly for being so kind to listen to me during my tough time. I thank Karin Bierig for being such a wonderful officemate throughout the time I spent in Tübingen.

Last but not least, I thank my family for being supportive and understanding no matter what decision I made. I feel incredibly lucky to have them and it is impossible to describe in words how much I am thankful for them.

Contributions

The majority of this dissertation results from the collaborations with several people and I would like to acknowledge their contributions explicitly. Major contributions stem from the following publications:

(Ch. 3) K. Muandet*, B. Sriperumburdur*, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 2015 (* contributed equally.)

(Ch. 3) 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

(Ch. 3) K. Muandet, B. Sriperumbudur, and B. Schölkopf. Kernel mean estimation via spectral filtering. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1–9. Curran Associates, Inc., 2014b

(Ch. 4) 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

(Ch. 5) 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

(Ch. 5) 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

Additionally, I have been involved in other projects during my Ph.D. study which only *partially* influence and inspire the theme of the thesis. These publications include

(1) K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning, W&CP 28 (3)*, pages 819–827. JMLR, 2013

(2) K. Zhang, B. Schölkopf, K. Muandet, Z. Wang, Z. Zhou, and C. Persello. *Single-Source Domain Adaptation with Target and Conditional Shift*, chapter 19, pages 427–456. Chapman & Hall/CRC Machine Learning & Pattern Recognition. Chapman and Hall/CRC, Boca Raton, USA, 2014

(3) 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

(4) D. Lopez-Paz, K. Muandet, and B. Recht. The randomized causation coefficient. *Journal of Machine Learning*, 2015a

(5) 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 (To Appear)*, 2015b

(6) B. Schölkopf, K. Muandet, K. Fukumizu, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 2015

Lastly, Chaper 2 is a shorten version of a longer review paper under preparation.

Contents

List of Figures

List of Tables

List of Symbols

$C(\mathcal{X})$	A space of all continuous functions on $\mathcal{X}$.
$C_0(\mathcal{X})$	A space of all continuous functions on $\mathcal{X}$ which vanish at infinity.
$C_b(\mathcal{X})$	A space of all bounded continuous functions on $\mathcal{X}$.
$L^1(\mathbb{R}^d)$	A space of Lebesgue integrable functions
$L^2(\mathbb{R}^d)$	A space of square integrable functions
X	A random variable taking value in $\mathcal{X}$
$\varphi_{\mathbb{P}}$	A characteristic function of distribution $\mathbb{P}$
$\mathbf{C}_{XX}$	A covariance operator on X
$\mathbf{C}_{XY}$	A cross-covariance operator from X to Y
$\mathbf{C}_{XY\|Z}$	A conditional cross-covariance operator of X and Y given Z
$\mathcal{F}$	A function space
$\mathscr{H}$	An RKHS of functions from $\mathcal{X}$ to $\mathbb{R}$
$\mathscr{F}$	An RKHS of functions from $\mathcal{Y}$ to $\mathbb{R}$
$\mathbf{K}$	A Gram matrix of kernel k
$\mathbf{L}$	A Gram matrix of kernel l
$\mathbf{T}_k$	An integral operator associated with kernel k
$\mathfrak{R}_n(\mathcal{F})$	The Rademacher complexity of the function class $\mathcal{F}$ based on n i.i.d. sample
ϕ	A feature map from an input space $\mathcal{X}$ to a high-dimensional feature space $\mathscr{H}$.
$\mathbb{P}$	A probability measure over some input space $\mathcal{X}$
$\mathscr{P}_+^1(\mathcal{X})$	A set of all probability measures defined on $\mathcal{X}$.
$\mathscr{P}_b(\mathcal{X})$	A set of all finite Borel measures defined on $\mathcal{X}$.
$\mathrm{HS}(\mathscr{F}, \mathscr{H})$	A Hilbert space of Hilbert-Schmidt operators mapping from $\mathscr{F}$ to $\mathscr{H}$
φ	A feature map from an input space $\mathcal{Y}$ to a high-dimensional feature space $\mathscr{F}$.
k	A positive definite kernel function on $\mathcal{X}$
l	A positive definite kernel function on $\mathcal{Y}$
x	A scalar value

Chapter 1

Introduction

I begin by giving a motivation of the thesis, its overview, and a brief outline of the subsequent chapters.

1.1 Motivations

Machine learning (ML) has played an important role in computer science and artificial intelligence as a mean to understand how to build a machine that is capable of *learning*, and in which situations it may succeed or fail. The ultimate goal is to build an "intelligent" machine that can learn from past experience, just like human naturally do. This endeavour has already led to many successful applications of ML across different fields, ranging from astronomy and high-energy physics to robotics and causal inference. In my opinion, a key to this success lies in its multi-disciplinary nature that brings together collaborations from statisticians, neuroscientists, psychologists, cognitive scientists, and many more.

Despite the success, we are still far from understanding what an intelligent machine is. I have always been fascinated by what can be achieved through technology. The technological revolution has made our lives different from our ancestors. Better living, reliable health-care, and scientific discoveries are just tips of the iceberg. The capability of computers in performing complex tasks such as the chess-playing robots whose ability exceeds that of the human world champion and the IBM Watson that outperforms human competitors at Jeopardy has increased exponentially. But, whether or not these machines are truly intelligent remains obscure. Understanding the meaning of intelligence has a great implication on what the intelligent systems can or cannot accomplish, their impact on our life, and the danger they may pose to our future. I believe one of the key ingredients to this understanding lie in their *ability to learn* and *make future prediction* about the world.

Empirical risk minimization (ERM) is one of the most prevalent frameworks for studying the statistical learning from empirical data (Vapnik 1992). Ultimately, we are interested in finding the functional relationship between two random variables X and Y based only on the empirical data. That is, given the independent and identically distributed (i.i.d.) random pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ from some unknown distribution $\mathbb{P}(X, Y)$ where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, the ERM finds a function $f : \mathcal{X} \to \mathcal{Y}$ which minimizes

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)), \quad f \in \mathcal{F} \tag{1.1}$$

for some function class $\mathcal{F}$. For instance, visual object recognition is one of the most important abilities we possess. In this case, $\mathbf{x}_i$ may represent images of car and y_i labels $\mathbf{x}_i$ by the type of car, *e.g.*, $\mathcal{Y} = \{\texttt{sedan}, \texttt{truck}, \texttt{sportcar}, \texttt{minivan}, \texttt{etc}\}$. From a collection of examples $(\mathbf{x}_i, y_i)$, we want to find f that when applied to any image of car, returns its correct type. Since in practice we do not have access to $\mathbb{P}(X, Y)$, the empirical risk (1.1) is used as a surrogate to

its population counterpart given by

$$R(f) = \int \ell(y, f(\mathbf{x}))\, \mathrm{d}\mathbb{P}(\mathbf{x}, y), \quad f \in \mathcal{F}. \tag{1.2}$$

The function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denotes a problem-specific loss function. For examples, if $\mathcal{Y} = \{0, 1\}$, we have a classification problem and a natural choice of ℓ is a 0-1 loss $\ell(y_i, f(\mathbf{x}_i)) = \mathbb{1}_{f(\mathbf{x}_i) \neq y_i}$, whereas if $\mathcal{Y} = \mathbb{R}$, we have a regression problem and the common loss function is a square loss $\ell(y_i, f(\mathbf{x}_i)) = (f(\mathbf{x}_i) - y_i)^2$.

In the past decades, several efforts have been devoted to a quest for sufficient and necessary conditions under which certain problems are learnable using the ERM. This usually translates into showing that a *uniform convergence bound* holds, *i.e.*,

$$\mathbb{P}^n \left\{ \sup_{f \in \mathcal{F}} \left| \widehat{R}(f) - R(f) \right| > \varepsilon \right\} \leq g(\varepsilon, n, \mathcal{F})$$

where $g(\varepsilon, n, \mathcal{F})$ represents a function that depends on ε, n, and $\mathcal{F}$, and vanishes as $n \to \infty$. This ensures that for any distribution $\mathbb{P}(X, Y)$, there exists a finite number of training examples n for which the learner can generalize well to the unseen test data given that both training and test data are generated i.i.d. from the same distribution and the complexity of the function class, *e.g.*, Rademacher complexity and VC dimension, is bounded. Although no assumption is generally made, prior knowledge about $\mathbb{P}(X, Y)$ may be used to improve learning. See, *e.g.*, Boucheron et al. (2005) for review. Moreover, another important line of research is exploratory data analysis such as principal component analysis (PCA) in which one is interested in extracting important properties of the underlying distribution $\mathbb{P}(X)$ from empirical data $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Unlike traditional setting, the primary objects of interest in this thesis are *probability distributions* $\mathbb{P}_i(\mathbf{x})$ over some input space $\mathcal{X}$ rather than data points $\mathbf{x}_i$ themselves. The ultimate goal is then to generalize and develop learning algorithms that operate directly on a space of probability distributions. Interestingly, from a measure-theoretic point of view, many classical settings can be viewed as learning from distributions, *i.e.*, when the data points $\mathbf{x}_i$ are replaced by the *Dirac measures* $\delta_{\mathbf{x}_i}$ which puts mass only at points $\mathbf{x}_i$ (cf. Figure 2.2). By enriching this perspective, the thesis investigates the feature representation of probability distribution, its empirical estimators, and general frameworks for learning on distributions based on such a representation.

1.1.1 Why Learning on Probability Distributions?

There are, in fact, many reasons why learning on probability distributions is important.

Firstly, it can be very useful in domain adaptation and transfer learning (see, *e.g.*, Ben-David et al. (2010), Pan et al. (2011), Pan and Yang (2010), Blanchard et al. (2011a), Muandet et al. (2013) and references therein). Several attempts have been made in generalizing the ERM to a scenario where the training and test data come from different distributions. To learn successfully in such a scenario, the algorithms need to understand how the distributions governing data generating processes change across time or domains. Moreover, the training data may be obtained from distinct and heterogeneous distributions and the knowledge of the distribution of the test data may not be available during the training time.

Secondly, probability distributions are good at modeling noisy/uncertain observations. Emerging technology allows us to collect a tremendous amount of data, which are usually noisy. Specialized technique is needed to deal with such data. For example, gene expression data are often measured with high uncertainty. Replication, which can be costly, is required (Yang and Speed 2002) to reduce such uncertainty. Similarly, the astronomical data are always subjected

to uncertainty due to evolving nature of the objects and atmospheric disturbance. To reduce this uncertainty, the measurement is often made several times to obtain the average values (Kirkpatrick et al. 2011, Bovy et al. 2011, Ross et al. 2012).

Moreover, in the era of "big data", it is imperative for the machine learning algorithms to be able to extract *high-level* information contained in such data. Most of classical algorithms only make use of information from individual data points, and often neglect their interactions. In group anomaly detection, for instance, we are interested in the anomalous events that occur in the aggregate levels (Chandola et al. 2009, Póczos et al. 2011, Xiong et al. 2011b;a, Muandet and Schölkopf 2013, Guevara et al. 2014). That is, the behaviour of the group may exhibit anomalous characteristic whereas none of the points in the group is anomalous (*e.g.*, high-energy physics). On the other hand, we may be interested in reducing the amount of data, while preserving most of the information that is necessary for successful learning. For example, we can summarize a set of data points by its average which throws away lots of information. Representing a set of data points by the distribution can capture most of the information while reducing the amount of computation required. The summary also help concealing sensitive information about individual sample, *i.e.*, privacy-preserving (Dwork 2008).

We can interpret many problems in statistics as learning problems on *empirical* probability distributions. For example, a "statistical estimator" is essentially a function from an empirical distribution to values of certain statistics such as parameter values, independence, conditional independence, and causal relation (Lopez-Paz et al. 2015b). Statisticians often consider the "plug-in" estimators whose form are known in advance (Lehmann and Casella 1998). In contrast, if training examples are available, *learning* such estimators allows one to impose weaker assumptions about the underlying data-generating process and may lead to "better" estimators. In many research areas, one is also interested in generalizing *domain-general knowledge* which is domain-invariant as opposed to the *domain-specific knowledge* which is specific to input domain. Examples include theory of causality in cognitive science and psychology (Goodman et al. 2011).

Most importantly, probability distributions constitute more complex concept and relation intelligent entities may encounter in reality, and by studying learning problems on them I hope to gain insights into the limitations of the current intelligent systems, and how to improve them.

1.1.2 Why Kernel Mean Representation?

Previous approaches based on kernel density estimation (Póczos et al. 2013, Oliva et al. 2014), divergence measure (Póczos et al. 2011), generative model (Jebara et al. 2004b, Xiong et al. 2011a), information geometry (Amari 2010), for example, have been applied successfully for learning and statistical inference from probability distributions. In contrast, this thesis focuses on the *kernel mean representation*. There are multiple reasons why this representation is attractive for learning framework on distributions.

First of all, kernel mean representation is very simple. It is fully characterized by a transformation

$$\mathbb{P} \longmapsto \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x},\cdot)] =: \boldsymbol{\mu}_{\mathbb{P}}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel function. As we can see, $\boldsymbol{\mu}_{\mathbb{P}}$ is simply a mean vector in feature space associated with the kernel k. As a result, we do not need to deal with distributions explicitly as many operations on $\mathbb{P}$ can be translated into operations on $\mu_{\mathbb{P}}$. The kernel mean $\mu_{\mathbb{P}}$ can be estimated consistently from the empirical data with provable guarantee.

Secondly, a certain class of kernel functions known as *characteristic kernels* ensures that the kernel mean representation captures all necessary information about the distribution (Fukumizu et al. 2004, Sriperumbudur et al. 2008; 2010). In other words, the map $\boldsymbol{\mu} : \mathbb{P} \mapsto \boldsymbol{\mu}_{\mathbb{P}}$ is injective

which implies that $\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|_{\mathscr{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. As a result, we can use the kernel mean representation to define a metric over a space of probability distributions (Sriperumbudur et al. 2010). A diverse choice of k also gives this representation more flexibility. Most machine learning algorithms can be extended to a space of probability distributions by choosing appropriate kernels (or approximation) of this representation (Gómez-Chova et al. 2010, Muandet et al. 2012, Guevara et al. 2014).

Next, basic operations on distributions can be performed by means of the inner product in the feature space. For example, we have $\mathbb{E}_{\mathbb{P}}[f(\mathbf{x})] = \langle f, \boldsymbol{\mu}_{\mathbb{P}} \rangle_{\mathscr{H}}$ for all $f \in \mathscr{H}$. Likewise, $\mathbb{E}_{Y|\mathbf{x}}[g(Y) \,|\, X = \mathbf{x}] = \langle g, \mathcal{U}_{Y|\mathbf{x}} \rangle_{\mathscr{F}}$ for all $g \in \mathscr{F}$ where $\mathcal{U}_{Y|\mathbf{x}}$ denotes the kernel mean embedding of the conditional distribution $\mathbb{P}(Y|X = \mathbf{x})$. Consequently, the kernel mean representation permits a probabilistic inference in a non-parametric fashion, *e.g.*, kernel belief propagation (Song et al. 2011a), kernel Monte Carlo filter (Kanagawa et al. 2013), and kernel Bayes' rule (Fukumizu et al. 2011).

In some applications such as testing for homogeneity from finite sample, the kernel mean representation allows one to bypass an intermediate density estimation, which is known to be difficult in high-dimensional setting (Wasserman 2006; Section 6.5). Moreover, the applications of kernel mean embedding can be extended straightforwardly to non-vectorial data such as graphs, strings, and semi-groups (Gärtner 2003). Most of the previous approaches only work in standard Euclidean space.

1.2 Thesis Overview and Contribution

The major contributions of this thesis can be summarized as follows:

- Overall, the thesis introduces learning frameworks when the inputs are not just points, but probability distributions. The use of kernel mean embedding as a representation for distribution allows us to generalize many of the classical algorithms and establishes interesting relationships with existing frameworks. The thesis also investigates the kernel mean estimation problem.

- The thesis gives a comprehensive review on both theory and practical applications of Hilbert space embedding of probability distributions in the past years. To the best of my knowledge, this is the first comprehensive review of research in this area.

- One of the most fundamental questions is how to estimate the kernel mean effectively and efficiently from the sample, which is an essential step in the applications of kernel mean embedding. The thesis investigates this question and shows that the standard kernel mean estimator, *i.e.*,

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

  can be improved by the *linear* shrinkage estimator of the form $\hat{\boldsymbol{\mu}}_{\alpha} := \alpha f^* + (1 - \alpha)\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ for some $\alpha \in [0, 1]$ and $f^* \in \mathscr{H}$. Hence, we propose a new family of estimators called *kernel mean shrinkage estimator* (KMSE) and provide several theoretical guarantees. By taking the geometrical properties of RKHS into account, the thesis provide *non-linear* extensions by mean of spectral filtering algorithms which are quite popular in the theory of inverse problem and regularization. The proposed idea can also be used to estimate other quantities such as covariance operators.

- The thesis provides a generalization of the ERM framework to a space of distributions. That is, we observe i.i.d. sample $(\mathbb{P}_1, y_1), \ldots, (\mathbb{P}_n, y_n)$ rather than $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

We show that the resulting framework amounts to constructing a kernel-based learning framework over a set of distributions when each of them is represented by the kernel mean embedding. The proposed framework allows one to generalize several well-known algorithms such as kernel ridge regression and Gaussian processes to a space of probability distributions. In particular, the thesis provides an extension of well-known support vector machine (SVM) to a space of probability distributions which we call a *support measure machine* (SMM) with theoretical insights and encouraging empirical results. In addition, the thesis provides discussions regarding connections to classical learning algorithms, possible extensions, and potential future directions.

- The proposed framework can also be applied in an unsupervised setting, especially for exploratory data analysis. First, the thesis provides an analysis of the feature representation of distributions and illustrate this by performing PCA on distributions. Next, it presents the algorithm for group anomaly detection called *one-class support measure machine* (OCSMM) and provides an analysis on the connection to variable kernel density estimation (VKDE). Lastly, the thesis demonstrates the proposed framework on the domain adaptation/generalization via the *domain-invariant component analysis* (DICA) algorithm with learning-theoretic bound.

- Last but not least, I want to point out that learning from distributions has potential applications in statistics. Many problems in statistics such as hypothesis testing involve finding a function of the empirical distribution to a certain set of outputs called *statistic*, *e.g.*, $\{-1, +1\}$ indicating whether or not to reject the null hypothesis. Conventional approach is to use *plug-in* estimators. On a contrary, if training data is available, we may *learn* such an estimator automatically from the data using the proposed frameworks. Preliminary results have demonstrated the effectiveness of this approach in real-world applications, *e.g.*, see Szabó et al. (2015), Lopez-Paz et al. (2015b).

1.3 Outline of the Thesis

Figure 1.1 depicts a high-level outline of the thesis whose details can be described as follows.

Chapter 2: This chapter provides a brief literature review on the area of kernel methods and a comprehensive review on kernel mean embedding of marginal and conditional distributions and their applications. It also provides the discussions regarding the relationships between kernel mean embeddings and other methods.

Chapter 3: This chapter addresses the kernel mean estimation problem and shows that the standard empirical estimator of kernel mean can be improved by the shrinkage estimators. A novel class of estimators called *kernel mean shrinkage estimators* (KMSEs) is proposed. Several theoretical analyses including consistency and convergence rate of estimators are also provided. Lastly, it provides extensive experimental results as evidence of the improvement of KMSEs over standard kernel mean estimator.

Chapter 4: Owing to the kernel mean embedding and its estimators, this chapter presents a supervised learning framework on probability distributions. It first discusses the *distributional risk minimization* framework and present the representer theorem for probability distributions. Next, the positive definite kernel functions for distributions based on the kernel mean embeddings are proposed including a *support measure machine* (SMM) which is a generalization of

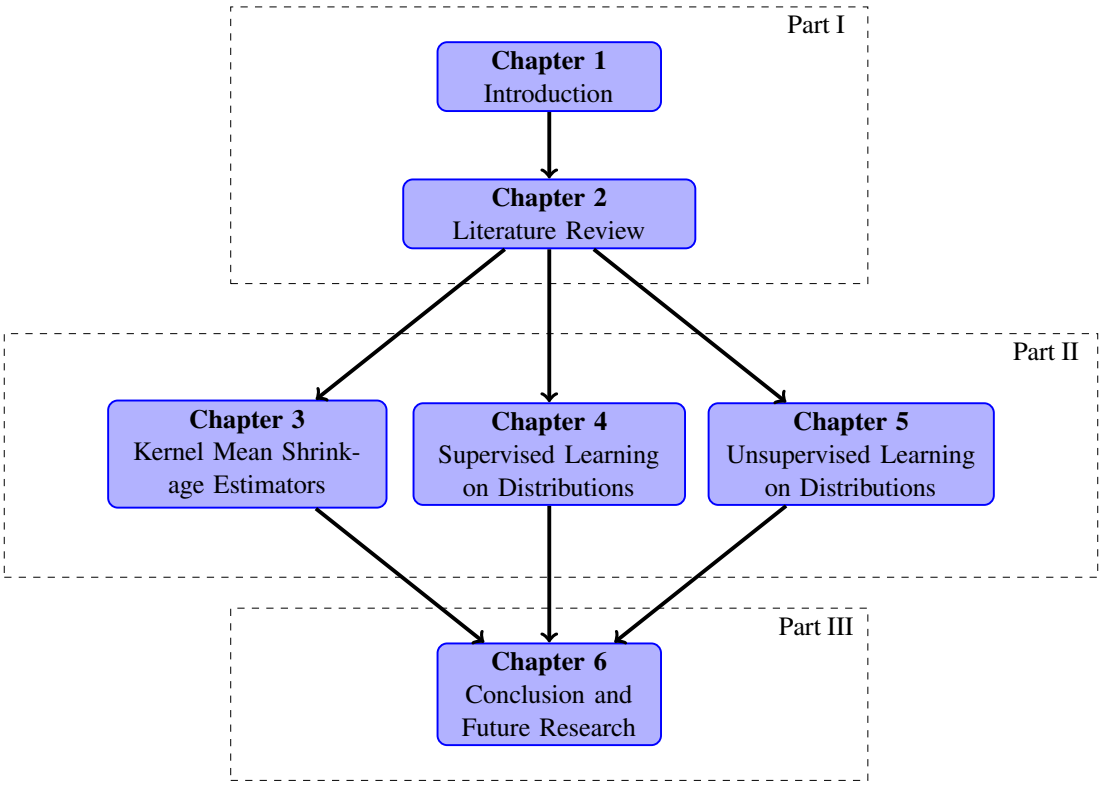


Figure 1.1: The outline of the thesis.

well-known support vector machine (SVM) to probability measures. I also discuss its connection to classical algorithms such as Parzen window classifiers. Both theoretical analysis and empirical results are also provided.

Chapter 5: This chapter demonstrates the learning framework on distributions in an unsupervised setting. First, an analysis of the proposed feature representation and empirical illustration via PCA on distributions are provided. Then, the thesis presents two applications, namely, group anomaly detection and domain adaptation/generalization.

Chapter 6: This chapter concludes the thesis and gives some suggestions for future research.

END OF CHAPTER 1

Chapter 2

Literature Review

2.1 Definitions & Notations

Table 2.1 summarizes the basic notations used throughout the thesis. I use capital letters to denote random variables and lowercase letters to denote instantiations of random variables, *e.g.*, X and x. I use a bold typeface to indicate vector and matrix (or operator) quantities, *e.g.*, $\mathbf{x}$ and $\mathbf{X}$. When describing the data set, I denote the total number of data points by n, and the total number of feature dimensions by d. The feature vector for the data point i is denoted by $\mathbf{x}_i$, and individual feature values are denoted by x_{ij}.

The primary object of interest in this thesis is probability distribution. For a topological input space $\mathcal{X}$, I denote by $\mathbb{P}$ a probability measure over such a space where a Borel σ-algebra is generated by the topology. I use $\varphi_{\mathbb{P}}$ to denote a characteristic function of $\mathbb{P}$. Let $\mathscr{P}$ be a space of all probability measures $\mathbb{P}$. For a random variable X taking value in $\mathcal{X}$, I denote the associated probability distribution by $\mathbb{P}(X)$ and $\mathbb{P}_X$ interchangeably. Given a pair of random variables X and Y, I decompose $\mathscr{P}$ into $\mathscr{P}_X$, which consists of the marginal distribution $\mathbb{P}(X)$, and $\mathscr{P}_{Y|X}$, which consists of posteriors $\mathbb{P}(Y|X)$.

For a topological space $\mathcal{X}$, $C(\mathcal{X})$ (*resp.* $C_b(\mathcal{X})$) denotes the space of all continuous (*resp.* bounded continuous) functions on $\mathcal{X}$. For a locally compact Hausdorff space $\mathcal{X}$, $f \in C(\mathcal{X})$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \geq \epsilon\}$ is compact. I denote the class of all continuous functions on $\mathcal{X}$ which vanish at infinity by $C_0(\mathcal{X})$. Denote by $\mathscr{P}_b(\mathcal{X})$ (*resp.* $\mathscr{P}^1_+(\mathcal{X})$), the set of all finite Borel (*resp.* probability) measures defined on $\mathcal{X}$.

2.2 Kernel Methods in Machine Learning

In this section, I introduce the kernel methods and the concept of reproducing kernel Hilbert space (RKHS) which form the backbone of this thesis.

2.2.1 A Kernel Trick

A solution to 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 1933b) can be expressed entirely in terms of inner product $\langle \mathbf{x}, \mathbf{x}' \rangle$, which is basically a similarity measure between $\mathbf{x}$ and $\mathbf{x}'$. However, a linear function class induced by this inner product is too restrictive for many real-world problems. Hence, 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:

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

Table 2.1: Basic notations used throughout the thesis

Symbol	Description
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$	non-empty sets (input spaces)
$X, Y, Z, \ldots$	random variables taking values in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$
$x, y, z, \ldots$	instantiations of random variables $X, Y, Z, \ldots$
$\mathbf{v}, v_i$	a vector and its ith element
$\mathbf{M}$	a matrix
$\mathbb{P}$	a probability distribution
$\hat{\mathbb{P}}$	an empirical distribution
$\varphi_{\mathbb{P}}$	a characteristic function of $\mathbb{P}$
$k(\mathbf{x}, \mathbf{x}')$	a real-valued positive definite kernel function on $\mathcal{X} \times \mathcal{X}$
$l(\mathbf{y}, \mathbf{y}')$	a real-valued positive definite kernel function on $\mathcal{Y} \times \mathcal{Y}$
$\phi(\mathbf{x}), \varphi(\mathbf{y})$	a feature map associated to the kernel k and l, respectively
$\mathscr{H}, \mathscr{F}$	an RKHS associated to the kernel k and l, respectively
$\mathbf{C}_{XX}$	a covariance operator on X
$\mathbf{C}_{XY}$	a cross-covariance operator from X to Y
$\mathbf{C}_{XY\|Z}$	a conditional cross-covariance operator of X and Y given Z

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

$$k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle. \tag{2.2}$$

I 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 $\mathbf{x}'$. Consequently, we can obtain a non-linear extensions of the linear algorithms simply by substituting $\langle \mathbf{x}, \mathbf{x}' \rangle$ with $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$. It is important to 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 the feature space $\mathcal{F}$ corresponds to the non-linear counterpart in the input space.

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

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

In other words, the new similarity measure is just the square of the dot product in $\mathcal{X}$. This result also 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 that resulting algorithm will 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 computational 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, we can use $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^2$ directly. This is an essential aspect of kernel methods, often referred to as a *kernel trick* in machine learning community.

It turns out that there exists a general class of k which guarantee that there exists some $\phi : \mathcal{X} \to \mathcal{F}$ for which $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$ as soon as k is *positive definite* (cf. Definition 2.1). Since the inner product $\langle \cdot, \cdot \rangle$ is positive definite, it follows from (2.2) that k is positive definite for any choice of explicit feature map ϕ.

Definition 2.1. *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel if it is symmetric,* i.e.*, $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})$, and 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 $n \in \mathbb{N}$ and choice of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Indeed a kernel function in the sense of Definition 2.1 associates to a space of functions called reproducing kernel Hilbert space (RKHS) $\mathscr{H}$, hence the name *reproducing kernel* (Aronszajn 1950). From this perspective, whenever we use the kernel k, we often think of a *canonical feature map*

$$k : \mathcal{X} \to \mathscr{H} \subset \mathbb{R}^{\mathcal{X}} \tag{2.5}$$

$$\mathbf{x} \mapsto k(\mathbf{x}, \cdot) \tag{2.6}$$

where $\mathbb{R}^{\mathcal{X}}$ denotes the vector space of functions from $\mathcal{X}$ to $\mathbb{R}$. An inner product in $\mathscr{H}$ satisfies the *reproducing property*

$$k(\mathbf{x}, \mathbf{x}') = \langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle. \tag{2.7}$$

Further detail of RKHS will be provided in Section 2.2.2. Note that although we do not need to know $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ explicitly, it is possible to derive $\phi(\cdot)$ directly from the kernel k (see, *e.g.*, Schölkopf and Smola (2001) for concrete examples).

The kernel trick not only results in more powerful learning algorithms, but also allows domain experts to come up with domain-specific kernel functions which can be verified easily. This leads to a number of kernel functions in various application domains (Genton 2002). In machine learning, commonly used kernels include the Gaussian and Laplacian kernels

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\right), \qquad k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2}{\sigma}\right), \tag{2.8}$$

where $\sigma > 0$ is a bandwidth parameter. Compared to the Gaussian kernel, the Laplacian kernel is less sensitive to changes in bandwidth parameter. These kernels belong to a class of kernel functions called a radial basis function (RBF) kernel. Both kernels are *translation invariant* which form an important class of kernel functions with essential properties, see, *e.g.*, Theorem 2.2.[1]

The kernel trick applies not only to real-valued random variables, but also extend to multivariate random variables, structured data, functional data, and other domains on which positive definite kernels may be defined. A review of several classes of kernel functions can be found in Genton (2002). Hofmann et al. (2008) also provides a general review of kernel methods in machine learning.

Another characterization of symmetric positive definite kernel k is the Mercer's theorem (Mercer 1909).

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

Theorem 2.1 (Mercer's theorem)**.** *Suppose k is a continuous positive definite kernel on a compact set $\mathcal{X}$, and the integral operator $\mathbf{T}_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ defined by*

$$(\mathbf{T}_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{2.9}$$

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

$$\int_{\mathcal{X}} k(\mathbf{u}, \mathbf{v}) f(\mathbf{u}) f(\mathbf{v}) \, \mathrm{d}\mathbf{u} \, \mathrm{d}\mathbf{v} \geq 0. \tag{2.10}$$

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

$$k(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{u}) \psi_i(\mathbf{v}) \tag{2.11}$$

where the convergence is absolute and uniform.

The condition (2.10) is known as a Mercer's condition and the kernel functions that satisfy this condition is often referred to as Mercer's kernels. It is important to 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 (2.2), the converse is not necessarily true. Since we are interested in the feature map ϕ, throughout this thesis, we consider the positive definite kernels that satisfy (2.2). Moreover, there is an intrinsic connection between integral operator $\mathbf{T}_k$, covariance operator $\mathbf{C}_{XX}$, and Gram matrix $\mathbf{K}$ (Rosasco et al. 2010) (see also Section 2.2.4).

Steinwart and Scovel (2012) studied the Mercer's theorem in general domains in which compactness assumption on $\mathcal{X}$ may not be satisfied. There is also a connection between Mercer's theorem in functional analysis and Karhunen-Loève theorem in the theory of stochastic processes (Rogers and Williams 2000a;b).

When the kernel k is translation invariant, *i.e.*, $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, we can characterize the kernel by Bochner's theorem (Bochner 1933).

Theorem 2.2 (Bochner's theorem)**.** *A 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*

$$\varphi(\mathbf{x} - \mathbf{x}') = \int e^{\sqrt{-1}\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')} \, \mathrm{d}\Lambda(\boldsymbol{\omega}). \tag{2.12}$$

In other words, Bochner's theorem states that k is the inverse Fourier transform of Λ and the translation-invariant kernels are the class of kernel functions that have non-negative Fourier transform.

By virtue of Theorem 2.2, we can interpret the kernel $k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$ 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 $k(\mathbf{x} - \mathbf{x}') = e^{-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2)}$ is a Gaussian distribution of the form $(2\pi/\sigma^2)^{-d/2} e^{-\sigma^2 \|\boldsymbol{\omega}\|^2/2} \, \mathrm{d}\boldsymbol{\omega}$. For Laplacian kernel $k(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x} - \mathbf{x}'\|/\sigma}$, the corresponding measure is a Cauchy distribution, *i.e.*, $\Lambda(\boldsymbol{\omega}) = \prod_d \frac{\sigma}{\pi(1 + \omega_d^2)}$.

As we will see later, Bochner's theorem also allows us to characterize the kernel mean embedding. Similarly, the measure Λ determines which frequency component of the characteristic function of $\mathbb{P}$ occurs in the embedding $\boldsymbol{\mu}_{\mathbb{P}}$. Hence, it follows from the uniqueness of the characteristic function that if the support of Λ is the entire $\mathbb{R}^d$, $\boldsymbol{\mu}_{\mathbb{P}}$ will uniquely determines $\mathbb{P}$ (Sriperumbudur et al. 2008; 2010; 2011a). In the context of this thesis, I prefer to think about Λ as a filter that selects certain properties when computing the similarity measure between probability distributions $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ w.r.t. a certain class of distributions $\mathscr{P}$ (more below).

Another promising application of Bochner's theorem is a finite approximation of kernel function. The feature map ϕ of many kernel functions such as the Gaussian kernel is infinite dimensional. In which case, the construction of the Gram matrix $\mathbf{K}$ where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is required. Therefore, most kernel-based learning algorithms scale at least quadratically with the sample size, which makes them prohibitive for large-scale problems. Rahimi and Recht (2007) proposes to approximate the translation invariant kernel k by replacing the integral in (2.12) with a finite sum based on a Monte Carlo sample $\boldsymbol{\omega} \sim \Lambda$. The Johnson-Lindenstrauss Lemma (Dasgupta and Gupta 2003, Blum 2005) ensures that this transformation will preserve similarity between data points. See also Kar and Karnick (2012), Le et al. (2013), Pham and Pagh (2013) and references therein for a generalization of this idea. Another common method to approximate $\mathbf{K}$ is a low-rank approximation, see, *e.g.*, Bach (2013) and references therein.

2.2.2 Reproducing Kernel Hilbert Space

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

Definition 2.2. *A Hilbert space is a real (or complex) inner product space that is also a complete metric space w.r.t. the distance function induced by the inner product.*

Well-known examples of Hilbert spaces include standard Euclidean space $\mathbb{R}^d$ with $\langle \mathbf{x}, \mathbf{y} \rangle$ the vector dot product of $\mathbf{x}$ and $\mathbf{y}$, a space of square summable sequences ℓ^2 of $\mathbf{x} = (x_1, x_2, \ldots)$ with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$ such that the series $\sum_{n=1}^{\infty} |z_n|^2$ converges, and the space of square-integrable functions $L_2[a, b]$ with inner product $\langle f, g \rangle = \int_a^b f(x) g(x) \, \mathrm{d}x$. Hilbert spaces with their norm given by the inner product are examples of *Banach spaces* (Ledoux and Talagrand 1991). 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 thesis will deal mostly with the embedding of distributions in the Hilbert space. Sriperumbudur et al. (2011b) has already extended the idea to a more general Banach space.

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

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

$$|\mathbf{F}_{\mathbf{x}}[f]| = |f(\mathbf{x})| \leq C \|f\|_{\mathscr{H}}, \qquad \forall f \in \mathscr{H}. \tag{2.13}$$

Intuitively speaking, functions in the RKHS are smooth in the sense of (2.13). This smoothness property ensures that the solution in RKHS obtained from learning algorithms will be well-behaved, *i.e.*, small $\|f - g\|_{\mathscr{H}}$ implies that $f(x)$ and $g(x)$ are close a.e.. 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 Hilbert spaces. The space of square-integrable functions $L_2[a, b]$ does not have this property.

That is, it is very easy to find a function in $L_2[a, b]$ that attains zero risk on the training data, *i.e.*, overfitting.

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

Theorem 2.3 (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*

$$\mathbf{A}f = \langle f, g_{\mathbf{A}} \rangle_{\mathscr{H}}, \qquad \forall f \in \mathscr{H}. \tag{2.14}$$

The Riesz representation theorem will be used to prove a sufficient condition for the existence of the kernel mean embedding in the Hilbert space (see Lemma 2.7). 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 2.4 states this result, which is called a *reproducing property*.

Proposition 2.4 (reproducing property)**.** *For each $\mathbf{x} \in \mathcal{X}$, there exists a function $k_{\mathbf{x}} \in \mathscr{H}$ such that*

$$\mathbf{F}_{\mathbf{x}}[f] = \langle k_{\mathbf{x}}, f \rangle_{\mathscr{H}} = f(\mathbf{x}). \tag{2.15}$$

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

$$k(\mathbf{x}, \mathbf{y}) = k_{\mathbf{y}}(\mathbf{x}) = \langle k_{\mathbf{x}}, k_{\mathbf{y}} \rangle_{\mathscr{H}} = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathscr{H}}, \tag{2.16}$$

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 2.3, 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, and vice versa.*

The sufficient and necessary conditions for a function $k(\cdot, \cdot)$ to be a reproducing kernel are given in Definition 2.1. Detailed exposition of RKHS can be found in Schölkopf and Smola (2001), Berlinet and Thomas-Agnan (2004), for example.

2.2.3 Learning with Kernels

In this section I review some well-known algorithms, namely principal component analysis, support vector machine, ridge regression, and Gaussian process together with their kernelized counterparts.

Kernel Principal Component Analysis (KPCA)

Principal component analysis (PCA) is an essential tool for modern data analysis (Hotelling 1933a, Jolliffe 1986). The PCA provides a powerful mathematical tool to unravel interesting, sometimes hidden, structures that underlies a complex data set. The goal of PCA is to find a meaningful basis that "best" explains a data set in terms of the variance. That is, given a data

set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$, PCA looks for a principal component $\mathbf{v}_1$ in $\mathbb{R}^d$ such that the variance of the projection

$$\frac{1}{n}\sum_{i=1}^{n}\left(\langle \mathbf{x}_i, \mathbf{v}_1\rangle - \sum_{j=1}^{n}\langle \mathbf{x}_j, \mathbf{v}_1\rangle\right)^2$$

is maximized. It is often assumed that the data set $\mathcal{D}$ is centered such that $\sum_{i=1}^{n}\mathbf{x}_i = \mathbf{0}$. PCA assumes that all basis vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ are orthonormal, *i.e.*, $\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0$, $i \neq j$ and $\|\mathbf{v}_i\|_2 = 1$. In other words, the projection matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)^\top$ is an orthonormal matrix.

Let define a $n \times d$ matrix $\mathbf{X}$ whose rows correspond to data points and columns correspond to features (or variables). Then, the covariance matrix $\mathbf{C}$ can be expressed in terms of $\mathbf{X}$ as

$$\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top. \tag{2.17}$$

It is not difficult to show that the orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ coincides with the first p eigenvectors of $\mathbf{C}$ with largest eigenvalues. The eigenvalues specify the amount of variance captured by the corresponding eigenvectors. Consequently, the PCA finds the projection $\mathbf{V}$ by solving the following eigendecomposition problem

$$\mathbf{C}\mathbf{V} = \Lambda\mathbf{V} \tag{2.18}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix consisting of corresponding eigenvalues.

The standard PCA algorithm only gives a linear projection which can be very restrictive in many applications. As described above, the most straightforward extension of PCA to deal with non-linear relationship is to replace (2.17) with the covariance matrix in the feature space, *i.e.*, each column of $\mathbf{X}$ now consists of the feature map $\phi(\mathbf{x}_i)$. However, for infinite dimensional feature space, solving the eigenvalue problem (2.18) is no longer possible in practice.

Alternatively, one can resort to the well-known trick that conventional PCA can be reformulated in such a way that the data vectors appear only in the form of the scalar product (Schölkopf et al. 1998). That is, we decompose the dot product matrix $\mathbf{X}^\top\mathbf{X}$ and left multiply by the data matrix: $\mathbf{X}^\top\mathbf{X}\mathbf{U} = \Lambda\mathbf{U} \Leftrightarrow (\mathbf{X}\mathbf{X}^\top)(\mathbf{X}\mathbf{U}) = \Lambda(\mathbf{X}\mathbf{U})$. As a result, the PCA can be performed using dot product matrix instead of covariance matrix. Let $\mathbf{K}$ be a Gram matrix such that $\mathbf{K}_{ij} = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$. Substituting (2.17) with the feature map $\phi(\mathbf{x}_i)$ in (2.18), then all solutions $\mathbf{v}_k$ lie in the span of $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)\}$. That is, there exist coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ such that

$$\mathbf{v}_k = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i). \tag{2.19}$$

Consequently, we can find eigenvector $\mathbf{v}_k$ by solving the eigenvalue problem

$$n\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha} \tag{2.20}$$

for nonzero eigenvalues. Additionally, we normalize the solution $\boldsymbol{\alpha}$ belonging to nonzero eigenvalues by requiring that the corresponding vectors in $\mathscr{H}$ be normalized, *i.e.*, $(\mathbf{v}_k \cdot \mathbf{v}_k) = 1$. This translates into

$$1 = \sum_{i,j=1}^{m}\alpha_i\alpha_j\mathbf{K}_{ij} = \boldsymbol{\alpha}\cdot\mathbf{K}\boldsymbol{\alpha} = \lambda_k(\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}). \tag{2.21}$$

Given a new data point $\mathbf{x}$, the projected value of $\mathbf{x}$ onto the component $\mathbf{v}_k$ can be computed as $\sum_{i=1}^{n}\alpha_i\langle\phi(\mathbf{x}_i), \phi(\mathbf{x})\rangle$. The time complexity of eigenvalue problem (2.20) depends only on n, rather than the dimensionality d of the feature space. For large n, standard low-rank approximations for $\mathbf{K}$ can be applied to reduce computational complexity.

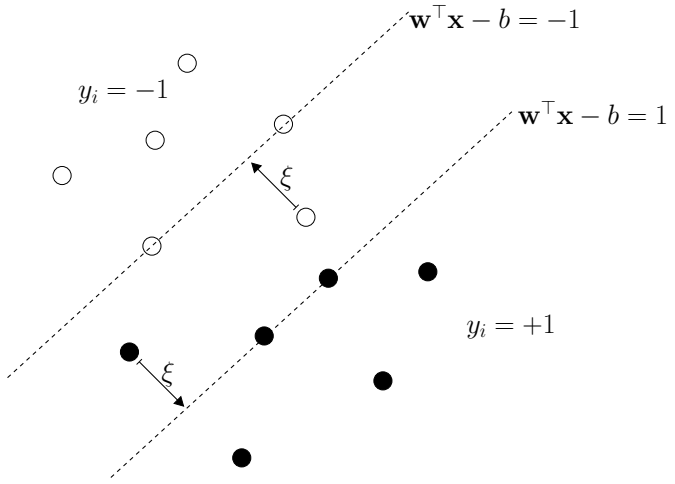


Figure 2.1: An illustration of the separating hyperplanes of the soft-margin SVM.

Support Vector Machines (SVM)

Support vector machine (SVM) is one of the most successful algorithms for classification. Given a training data

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n.$$

The objective of SVM is to find the maximum-margin hyperplane that separates the points having $y_i = +1$ from those having $y_i = -1$.

In the following, I only give the detail of a soft-margin version of the SVM because it is widely used in many applications. Other variants of SVM formulation can be found in Schölkopf and Smola (2001). If the data in $\mathcal{D}$ are linearly separable[2], the idea of linear SVM is to select two hyperplanes in a way that they separate the training data and there are no points between them. In general, there could be an infinitely many of such hyperplanes. We call the distance between these two hyperplane "the margin". Among all possible hyperplanes, the SVM selects the ones to maximize this margin. Since these two hyperplanes can be described by the following equations:

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} - b &= 1 \\ \mathbf{w} \cdot \mathbf{x} - b &= -1, \end{aligned}$$

it is straightforward to show that the margin is simply $2/\|\mathbf{w}\|$. Hence, maximizing margin is equivalent to minimizing $\|\mathbf{w}\|$ w.r.t. the constraints that the training data are correctly classified.

Unfortunately, the assumption of linear separability does not hold for many real-world data. As a result, it is impossible to find such a linear hyperplane that separates the data perfectly. To relax this assumption and allow some errors to be made, the idea of soft-margin SVM is to introduce non-negative slack variables, ξ_i, which measure the degree of mis-classification of the classifier on the data point $\mathbf{x}_i$. The objective function of the soft-margin SVM then involves a trade-off between a large margin and a small error penalty. The optimization problem of SVM

[2]Two point sets are said to be linearly separable in d-dimensional space if they can be separated by a hyperplane. Mathematically, let X and X' be two sets of points in d-dimensional space. Then, X and X' are linearly separable if there exists real number $w_1, w_2, \ldots, w_d, k$ such that $\sum_{i=1}^d w_i x_i \geq k$ for every point $\mathbf{x} \in X$ and $\sum_{i=1}^d w_i x'_i < k$ for every point $\mathbf{x}' \in X'$.

can be formulated as follow:

$$\begin{aligned} \underset{\mathbf{w},\boldsymbol{\xi},b}{\text{minimize}} \quad & \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \right\} \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ 1 \leq i \leq n, \end{aligned}$$

where the parameter C controls a trade-off between a large margin and a small error penalty. Figure 2.1 illustrates the separating hyperplane of the soft-margin SVM. Using Lagrange multipliers, we can solve the problem above via the following saddle-point problem:

$$\min_{\mathbf{w},\boldsymbol{\xi},b}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i - b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \right\}$$

where $\boldsymbol{\alpha} \geq 0$ and $\boldsymbol{\beta} \geq 0$. Consequently, we obtain the dual form of soft-margin SVM

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \\ & \sum_{i=1}^{n}\alpha_i y_i = 0. \end{aligned}$$

From the dual form of SVM, one can replace $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ directly with a non-linear similarity measure such as kernel function $k(\mathbf{x}, \mathbf{x}')$ to get a non-linear classifier.

Using the Karush–Kuhn–Tucker condition, the solution $\mathbf{w}$ can be expressed as a linear combination of the training vectors

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i k(\mathbf{x}_i, \cdot).$$

The function value evaluated on a new sample $\mathbf{x}^*$ can be computed by taking the inner product in $\mathscr{H}$ between the weight vector $\mathbf{w}$ and the data feature map $\phi(\mathbf{x}^*)$, *i.e.*,

$$f(\mathbf{x}^*) = \langle \mathbf{w}, \phi(\mathbf{x}^*) \rangle = \sum_{i=1}^{n}\alpha_i y_i k(\mathbf{x}_i, \mathbf{x}^*).$$

Then, the classification of $\mathbf{x}^*$ can be achieved by considering the sign of $f(\mathbf{x}^*)$, *i.e.*, $y^* = \operatorname{sign}[f(\mathbf{x}^*)]$ where $\operatorname{sign}[c] = 1$ if $c > 0$ and -1 otherwise. Solving for an SVM solution amounts to a quadratic program which can be solved efficiently using specialized implementations.

Kernel Ridge Regression (KRR) and Gaussian Process (GP)

Ridge regression and its kernelized counterpart is arguably one of the most elementary algorithm for regression problem. Given a data set of labeled examples $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, the goal of regression is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ which approximates well the conditional expectation $\mathbb{E}[Y|X = \cdot]$ where the expectation is taken over the conditional distribution $\mathbb{P}(Y|X = \cdot)$. In practice, we consider the function f which minimize the squared loss function

$$L(f) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2. \tag{2.22}$$

First, let consider a class of linear functions $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ for some parameter vector $\mathbf{w} \in \mathbb{R}^d$, in which case we can re-write (2.22) in terms of $\mathbf{w}$ as

$$L^*(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2. \tag{2.23}$$

Here, $L^*(\mathbf{w})$ denotes the dual loss function. Taking a derivative of $L^*(\mathbf{w})$ w.r.t. $\mathbf{w}$ and setting it to zero yield a close-form solution for $\mathbf{w}$, *i.e.*,

$$\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}, \tag{2.24}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$. This algorithm is commonly known as a least square (LS) algorithm. We can see from (2.24) that a problem may arise if $\mathbf{X}\mathbf{X}^\top$ does not have full rank. In other words, the problem is underdetermined and solving for $\mathbf{w}$ becomes ill-posed, namely, a solution may not exist, be unique, or does not change continuously with the initial condition.

To overcome this problem, one can resort to the regularized version of (2.23),

$$L^*_\lambda(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{2.25}$$

where λ is a positive regularization parameter. Solving for $\mathbf{w}$ yields a well-known ridge regression (RR) algorithm:

$$\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}. \tag{2.26}$$

Clearly, $(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})$ has full rank. Again, the LS and RR only consider a class of linear functions. To extend ridge regression to approximate non-linear functions, we can directly replace $\mathbf{x}_i$ with feature map $\phi(\mathbf{x}_i)$ in (2.25) and (2.26). In this case, the dimensionality of $\mathbf{w}$ can be much higher if not infinite. Alternatively, denoting $\Phi = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))^\top$ we may resort to the following identity

$$\mathbf{w} = (\Phi\Phi^\top + \lambda\mathbf{I}_d)^{-1}\Phi\mathbf{y} = \Phi(\Phi^\top\Phi + \lambda\mathbf{I}_n)^{-1}\mathbf{y} = \Phi(\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y} = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i)$$

where we define $\boldsymbol{\alpha} := (\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y}$. Hence, given a test point $\mathbf{x}_*$, we can evaluate the function value by $f(\mathbf{x}_*) = \langle \mathbf{w}, \phi(\mathbf{x}_*)\rangle = \sum_{i=1}^{n}\alpha_i k(\mathbf{x}_i, \mathbf{x}_*)$.

I close this section by reviewing Gaussian processes (GPs), which can be viewed as a Bayesian counterpart of least square and ridge regression algorithms. GPs extend multivariate Gaussian distributions to infinite dimensional vectors, *i.e.*, functions. Similar to least square regression, the GPs can be obtained via *Bayesian linear regression*. That is, assuming noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, the linear regression model is

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \quad y = f_{\mathbf{w}}(\mathbf{x}) + \varepsilon. \tag{2.27}$$

Assuming a zero-mean Gaussian prior over parameters $\mathbf{w}$ with covariance $\boldsymbol{\Sigma}_p$ and applying Bayes' theorem yield a posterior distribution over $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{w}\,;\, \frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{A}^{-1}\right), \quad \mathbf{A} = \boldsymbol{\Sigma}_p^{-1} + \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^\top. \tag{2.28}$$

GP allows for a full posterior over a function class. Hence, the predictive distribution at the test point $\mathbf{x}_*$ can be obtained by marginalizing out the parameter $\mathbf{w}$ as

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) \;\; = \;\; \int f_{\mathbf{w}}(\mathbf{x}_*)p(\mathbf{w}|\mathbf{y}, \mathbf{X})\,\mathrm{d}\mathbf{w} \tag{2.29}$$

$$= \mathcal{N}\left(y_*\,;\, \frac{1}{\sigma^2}\mathbf{x}_*^\top \mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^\top \mathbf{A}^{-1}\mathbf{x}_*\right). \tag{2.30}$$

Like in linear least square, we may increase the expressiveness of the Bayesian linear regression by considering a function of the form $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x})$. Hence, the mean and the covariance of predictive distribution become

$$\mathbf{m} = \Phi(\mathbf{x}_*)^\top \mathbf{\Sigma}_p \Phi(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{2.31}$$

$$\mathbf{C} = \Phi(\mathbf{x}_*)^\top \mathbf{\Sigma}_p \Phi(\mathbf{x}_*) - \Phi(\mathbf{x}_*)^\top \mathbf{\Sigma}_p \Phi(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\Phi^\top \mathbf{\Sigma}_p \Phi(\mathbf{x}_*) \tag{2.32}$$

where we used the same trick as in the case of KRR to derive the mean and matrix inversion lemma to derive the covariance. See Rasmussen and Williams (2005; Chapter 2) for more detail. Notice that we have defined $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \mathbf{\Sigma}_p \phi(\mathbf{x}')$. Because the mean and covariance of the predictive distribution can be written solely in terms of inner product $\phi(\mathbf{x})^\top \mathbf{\Sigma}_p \phi(\mathbf{x}')$, this algorithm also lends itself to the kernel trick.

The GP prior may be defined directly over the space of functions. That is, we assume that f is drawn from the GP prior such that for all n and all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} + \sigma^2\mathbf{I})$ where the entry $\mathbf{\Sigma}_{ij}$ specifies the covariance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ which is usually specified by the kernel function, *i.e.*, $\mathbf{\Sigma}_{ij} = \mathrm{Cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j)$. Given a set of n training data point $\{\mathbf{x}_i, y_i\}_{i=1}^n$, joint distribution of $\mathbf{y}$ and y_* is

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_*^\top \\ \mathbf{k}_* & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right) \tag{2.33}$$

Since Gaussian distribution is close under marginalization, we have $p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_*)$ as in (2.31) and (2.32). Note that σ^2 plays a similar role as regularization parameter λ in kernel ridge regression.

There is a correspondence between weight-space view and function-space view of GP. For any set of basis functions $\phi(\mathbf{x})$, the corresponding covariance function is $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \mathbf{\Sigma}_p \phi(\mathbf{x}')$. Conversely, for every covariance function k, there is a possibly infinite expansion in terms of basis functions, *i.e.*, $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^\infty \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}')$, due to Mercer's theorem (cf. Theorem 2.1).

2.2.4 Cross-Covariance and Hilbert-Schmidt Operators

The covariance, cross-covariance, and Hilbert-Schmidt operators on RKHS 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 RKHS. 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 random variable taking values on $\mathcal{X} \times \mathcal{Y}$ and $(\mathscr{H}, k)$ and $(\mathscr{F}, l)$ be RKHS with measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. We assume the integrability

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

which ensures that $\mathscr{H} \subset L^2(\mathbb{P}_X)$ and $\mathscr{F} \subset L^2(\mathbb{P}_Y)$. The *cross-covariance operator* $\mathbf{C}_{YX} : \mathscr{H} \to \mathscr{F}$ can be defined as

$$\mathbf{C}_{YX} := \mathbb{E}_{YX}[\varphi(Y) \otimes \phi(X)] - \boldsymbol{\mu}_Y \otimes \boldsymbol{\mu}_X \tag{2.34}$$

$$= \boldsymbol{\mu}_{\mathbb{P}_{YX}} - \boldsymbol{\mu}_{\mathbb{P}_Y \otimes \mathbb{P}_X}. \tag{2.35}$$

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

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

for all $f \in \mathscr{H}$ and $g \in \mathscr{F}$. These two equivalent definitions stem from the relations between the covariance operator and mean element of the joint measure $\mathbb{P}_{XY}$ (Baker 1973). If $X = Y$, we call $\mathbf{C}_{XX}$ the *covariance operator*, which is self-adjoint and positive.

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 $\mathbf{C}_{YX}$ can be obtained as

$$\begin{aligned} \widehat{\mathbf{C}}_{YX} &= \frac{1}{n} \sum_{i=1}^{n} \{l(\mathbf{y}_i, \cdot) - \hat{\boldsymbol{\mu}}_Y\} \otimes \{k(\mathbf{x}_i, \cdot) - \hat{\boldsymbol{\mu}}_X\} \\ &= \frac{1}{n} \Phi \mathbf{H} \Psi^\top. \end{aligned} \tag{2.36}$$

where $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n$ is the centering matrix with $\mathbf{1}_n$ an $n \times n$ matrix of ones, and $\Phi = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))^\top$, $\Psi = (\varphi(\mathbf{y}_1), \ldots, \varphi(\mathbf{y}_n))^\top$. The empirical covariance operator $\widehat{\mathbf{C}}_{XX}$ can be obtained in a similar way, *i.e.*, $\widehat{\mathbf{C}}_{XX} = \frac{1}{n}\Phi\mathbf{H}\Phi^\top$.

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 2.6. *There exists a unique bounded operator $\mathbf{V}_{YX} : \mathscr{H} \to \mathscr{F}$, $\|\mathbf{V}\| \leq 1$, such that*

$$\mathbf{C}_{YX} = \mathbf{C}_{YY}^{1/2} \mathbf{V}_{YX} \mathbf{C}_{XX}^{1/2}, \tag{2.37}$$

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

The operator $\mathbf{V}_{YX}$ is often referred to as the *normalized cross-covariance operator* and has been used as a basis for conditional dependence measure (Fukumizu et al. 2007; 2008). Compared to $\mathbf{C}_{YX}$, $\mathbf{V}_{YX}$ captures the same information about the dependence of X and Y, but with less influence of the marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$.

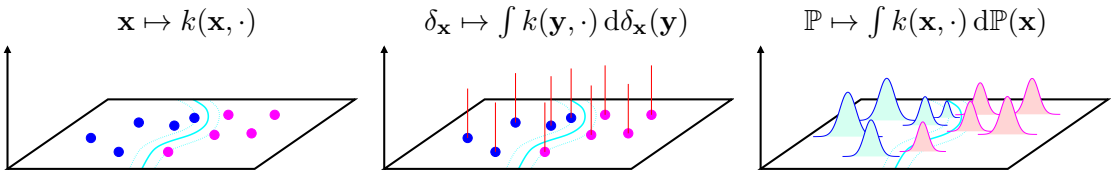
The covariance operator serves as a basic building block in classical kernel-based methods such as kernel PCA (Schölkopf et al. 1998, Zwald et al. 2004), kernel Fisher discriminant, and kernel CCA (Fukumizu et al. 2007). More recent applications of covariance operator include non-linear independence and conditional independence measures (Gretton et al. 2005b, Zhang et al. 2008; 2011, Doran et al. 2014). It is known that the covariance operator and integral operator defined in Theorem 2.1 are very much related (Hein and Bousquet 2004, Rosasco et al. 2010).

Hilbert-Schmidt Operators. 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, where the index set I and J need not be countable. A Hilbert-Schmidt operator is a bounded operator $\mathbf{A} : \mathscr{F} \to \mathscr{H}$ whose Hilbert-Schmidt norm

$$\|\mathbf{A}\|_{\mathrm{HS}}^2 = \sum_{j \in J} \|\mathbf{A} f_j\|_{\mathscr{H}}^2 = \sum_{i \in I} \sum_{j \in J} |\langle \mathbf{A} f_j, h_i \rangle_{\mathscr{H}}|^2 \tag{2.38}$$

is finite. The Hilbert-Schmidt operators mapping from $\mathscr{F}$ to $\mathscr{H}$ form a Hilbert space $\mathrm{HS}(\mathscr{F}, \mathscr{H})$ with inner product $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle \mathbf{A} f_j, \mathbf{B} f_j \rangle_{\mathscr{H}}$. The Hilbert space of Hilbert-Schmidt operators is beyond the scope of this thesis; see, *e.g.*, Zwald et al. (2004) for further detail.

The cross-covariance operator $\mathbf{C}_{YX}$ is in fact Hilbert-Schmidt. To see that, let first consider a *rank-one operator* defined as the tensor product $a \otimes b$ from $\mathscr{F}$ to $\mathscr{H}$ where $a \in \mathscr{F}$ and



(a) data point **(b)** Dirac measure **(c)** probability measure

Figure 2.2: 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 a class of probability measures.

$b \in \mathscr{H}$, *i.e.*, we have $(b \otimes a)f \mapsto \langle f, a \rangle_{\mathscr{F}} b$. It is not difficult to show using Parseval's identity that $\|a \otimes b\|^2_{\mathrm{HS}} = \|a\|^2_{\mathscr{F}} \|b\|^2_{\mathscr{H}} < \infty$. Thus, by definition this operator is Hilbert-Schmidt. Given a Hilbert-Schmidt operator $\mathbf{A} \in \mathrm{HS}(\mathscr{F}, \mathscr{H})$, we can write the inner product between $\mathbf{A}$ and $a \otimes b$ as

$$\langle \mathbf{A}, a \otimes b \rangle_{\mathrm{HS}} = \langle a, \mathbf{A} b \rangle_{\mathscr{H}} \tag{2.39}$$

If $\mathbf{A}$ is also a rank-one operator $u \otimes v$ where $u \in \mathscr{F}$ and $v \in \mathscr{H}$, we have $\langle u \otimes v, a \otimes b \rangle_{\mathrm{HS}} = \langle u, a \rangle_{\mathscr{F}} \langle v, b \rangle_{\mathscr{H}}$. Hence, we can see that the cross-covariance operator $\mathbf{C}_{YX}$ is the unique element in $\mathrm{HS}(\mathscr{H}, \mathscr{F})$ satisfying

$$\langle \mathbf{C}_{YX}, \mathbf{A} \rangle_{\mathrm{HS}} = \mathbb{E}_{XY} \left[\langle \phi(X) \otimes \varphi(Y), \mathbf{A} \rangle_{\mathrm{HS}} \right]. \tag{2.40}$$

If follows from Jensen's inequality and Cauchy-Schwarz inequality that

$$\begin{aligned} \left| \mathbb{E}_{XY} \left[\langle \phi(X) \otimes \varphi(Y), \mathbf{A} \rangle_{\mathrm{HS}} \right] \right| &\leq \mathbb{E}_{XY} \left| \left[\langle \phi(X) \otimes \varphi(Y), \mathbf{A} \rangle_{\mathrm{HS}} \right] \right| \\ &\leq \mathbb{E}_{XY} \left[\|\phi(X) \otimes \varphi(Y)\|_{\mathrm{HS}} \right] \|\mathbf{A}\|_{\mathrm{HS}}. \end{aligned}$$

Hence, by Riesz representation theorem, the cross-covariance operator exists if and only if $\mathbb{E}_{XY} \left[\|\phi(X) \otimes \varphi(Y)\|_{\mathrm{HS}} \right] < \infty$, which is equivalent to ensuring $\mathbb{E}_{XY} [\sqrt{k(X, X) l(Y, Y)}] < \infty$. Setting $\mathbf{A} = f \otimes g$ where $f \in \mathscr{H}$ and $g \in \mathscr{F}$ yields the result in (2.34).

Recently, the Hilbert-Schmidt operators have received much attention in machine learning community. For instance, Gretton et al. (2005a) uses a Hilbert-Schmidt norm of $\mathbf{C}_{YX}$ as a measure of statistical dependence between random variables X and Y (see also Chwialkowski and Gretton (2014) and Chwialkowski et al. (2014) for an extension to random processes). Quang et al. (2014) proposes a *Log-Hilbert-Schmidt metric* between positive definite operators on a Hilbert space, which is applied in particular to compute distance between covariance operators in RKHS.

2.3 Kernel Mean Embedding of Marginal Distributions

This section presents the idea of Hilbert-space embedding of distributions by generalizing the standard point of view of the kernel feature map of random sample to Dirac measures. Then, I show how reproducing kernels can be used to represent probability measures in feature space. I summarize this generalization in Figure 2.2.

2.3.1 From Data Points to Probability Measures

We can generalize the concept of high-dimensional feature map of data points $\mathbf{x} \in \mathcal{X}$ to measures on $(\mathcal{X}, \mathcal{A})$ 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

$$\delta_{\mathbf{x}}(A) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{if } \mathbf{x} \notin A, \end{cases} \tag{2.41}$$

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

$$\int f(\mathbf{t}) \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) = f(\mathbf{x}). \tag{2.42}$$

When f belongs to the Hilbert space $\mathscr{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, we can rewrite (2.42) using the reproducing property of $\mathscr{H}$ as

$$\int f(\mathbf{t}) \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) = \int \langle f, k(\mathbf{t}, \cdot) \rangle \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) = \left\langle f, \int k(\mathbf{t}, \cdot) \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}) \right\rangle = \langle f, k(\mathbf{x}, \cdot) \rangle. \tag{2.43}$$

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:

$$f \longmapsto \int f(\mathbf{t}) \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{t}), \tag{2.44}$$

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 $\langle f, k(\mathbf{x}, \cdot) \rangle$ give the same result $f(\mathbf{x})$, *i.e.*, value of f at the point $\mathbf{x}$, the former gives a measure-theoretic point of view of the latter (see also Figure 2.2). As a result, we can naturally define a feature map from a space of Dirac measures to $\mathscr{H}$ as

$$\delta_{\mathbf{x}} \longmapsto \int_{\mathcal{X}} k(\mathbf{y}, \cdot) \, \mathrm{d}\delta_{\mathbf{x}}(\mathbf{y}). \tag{2.45}$$

Intuitively speaking, 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 (2.43) this extension is not quite useful in practice because both algorithms are in fact equivalent. It is therefore more interesting to consider a non-trivial probability measure.

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 may consider a linear combination

$$\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \tag{2.46}$$

of Dirac measures putting the mass a_i at the point $\mathbf{x}_i$. This is known as a *signed* measure which constitutes a class of measures with finite support. A measure of the form (2.46) 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 (2.46) 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).

Likewise, for any measurable function f we have

$$\int f \,\mathrm{d}\left(\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}\right) = \sum_{i=1}^{n} a_i \int f \,\mathrm{d}\delta_{\mathbf{x}_i} = \sum_{i=1}^{n} a_i f(\mathbf{x}_i). \tag{2.47}$$

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

$$\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \longmapsto \sum_{i=1}^{n} a_i k(\mathbf{x}_i, \cdot) \tag{2.48}$$

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 $\boldsymbol{\mu} = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ denotes the discrete measure, we have for any f in $\mathscr{H}$

$$\left\langle f, \sum_{i=1}^{n} a_i k(\mathbf{x}_i, \cdot) \right\rangle = \sum_{i=1}^{n} a_i f(\mathbf{x}_i) = \int f \,\mathrm{d}\boldsymbol{\mu}. \tag{2.49}$$

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. It is easy to show that any function in $\mathscr{H}$ is measurable whenever k is measurable (see Schölkopf and Smola (2001) and Berlinet and Thomas-Agnan (2004) for technical details).

Consequently, we define the representer in $\mathscr{H}$ of the measure $\mathbb{P}$ through the mapping

$$\boldsymbol{\mu} : \mathscr{P}_+^1(\mathcal{X}) \longrightarrow \mathscr{H}, \quad \mathbb{P} \longmapsto \int k(\mathbf{x}, \cdot) \,\mathrm{d}\mathbb{P}(\mathbf{x}) \tag{2.50}$$

which will be denoted by $\boldsymbol{\mu}_{\mathbb{P}}$. This is essentially the kernel mean embedding we use throughout the thesis. The set $\mathscr{P}_+^1(\mathcal{X})$ contains signed measures $\mathbb{P}$ for which the embedding $\boldsymbol{\mu}_{\mathbb{P}}$ exists and belongs to $\mathscr{H}$. The conditions under which this is the case will be discussed later. Theoretical properties of $\boldsymbol{\mu}_{\mathbb{P}}$ will be described in Section 2.3.2.

Definition 2.4 (kernel mean embedding (Berlinet and Thomas-Agnan 2004, Smola et al. 2007))**.** *Suppose that a space $\mathscr{P}_+^1(\mathcal{X})$ consists of all Borel probability measures $\mathbb{P}$ on some input space $\mathcal{X}$. A kernel mean embedding of probability measures in $\mathscr{P}_+^1(\mathcal{X})$ into an RKHS $\mathscr{H}$ endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by a mapping*

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

where the integral used is a Bochner integral.

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

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \cdot). \tag{2.51}$$

Clearly, $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ is an unbiased estimator of $\boldsymbol{\mu}_{\mathbb{P}}$, and by the law of large number, $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ converges to $\boldsymbol{\mu}_{\mathbb{P}}$ as $n \to \infty$. Sriperumbudur et al. (2012) provides a thorough discussion on several properties of this estimator. In Chapter 3, I will discuss how to improve the estimation of $\boldsymbol{\mu}_{\mathbb{P}}$ by means of shrinkage estimator.

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, I restrict our attention to a space of marginal distributions $\mathbb{P}(X)$, and will provide an extension to a space of conditional distribution $\mathbb{P}(Y|X)$ in Section 2.4.

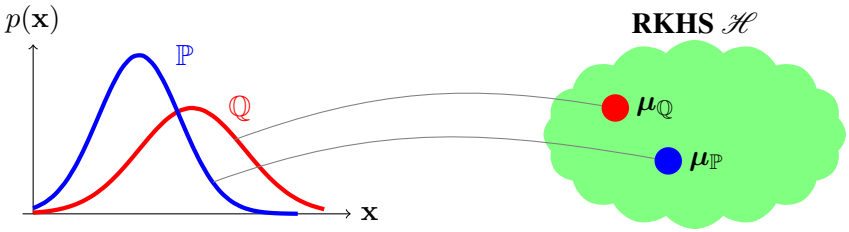


Figure 2.3: Embedding of marginal distributions: each distribution is mapped into an RKHS via an expectation operation. It corresponds to a mean element in the RKHS.

2.3.2 Theoretical Properties

Next, I provide some theoretical properties of the kernel mean embedding. The following result establishes sufficient conditions that guarantee the existence of $\boldsymbol{\mu}_{\mathbb{P}}$.

Lemma 2.7 (Smola et al. (2007))**.** *There exists* $\boldsymbol{\mu}_{\mathbb{P}} \in \mathscr{H}$ *if* $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[\sqrt{k(\mathbf{x},\mathbf{x})}] < \infty$.

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

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

where we use Jensen's inequality in $(*)$. Hence, by Riesz representation theorem (see, *e.g.*, Theorem 2.3), there exists a $\boldsymbol{\mu}_{\mathbb{P}} \in \mathscr{H}$ such that $\mathbf{L}_{\mathbb{P}} f = \langle f, \boldsymbol{\mu}_{\mathbb{P}}\rangle_{\mathscr{H}}$. ■

From the proof of Lemma 2.7, $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[f(\mathbf{x})] = \langle f, \boldsymbol{\mu}_{\mathbb{P}}\rangle_{\mathscr{H}}$ 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 $\boldsymbol{\mu}_{\mathbb{P}}$. 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). It can be extended to conditional distribution (see Section 2.4).

The following result, which appears in Lopez-Paz et al. (2015b), is a slight modification of Theorem 27 from Song (2008) which establishes the convergence of the empirical mean embedding $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ to the embedding of its population counterpart $\boldsymbol{\mu}_{\mathbb{P}}$ in RKHS norm:[3]

Theorem 2.8. *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*

$$\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\|_{\mathscr{H}} \leq 2\sqrt{\frac{\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x},\mathbf{x})]}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}. \tag{2.52}$$

We can see that the convergence happens at a rate $O(n^{-1/2})$. Generally, if we do not make any prior assumption about $\mathbb{P}$, this rate is known to be optimal.

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 $\boldsymbol{\mu}_{\mathbb{P}}$ becomes just the first

[3]The similar result is also given in Gretton et al. (2012a).

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 I 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 2.1 (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*

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

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

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

where $\mathbf{m}_{\mathbb{P}}(i)$ denotes the ith moment of the distribution $\mathbb{P}$. That is, the embedding incorporate 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 2.2 (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*

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

which is essentially a moment-generating function of a random variable X with distribution $\mathbb{P}$.

Example 2.3 (characteristic function)**.** *Consider the 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. Let $\varphi_{\mathbb{P}}$ be a characteristic function of $\mathbb{P}$. Bochner's theorem (see Theorem 2.2) allows us to express the kernel mean embedding as*

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

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 $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} = \langle \varphi_{\mathbb{P}}, \varphi_{\mathbb{Q}} \rangle_{L^2(\mathbb{R}^d, \Lambda)}$ (see, e.g.*, Theorem 5.1).*

Some authors might also consider explicitly the mean embedding of the sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. I 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 $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$. Hence, the same line of reasoning can be applied.

2.3.3 Universal and Characteristic Kernels

The notion of *universal kernels* and *characteristic kernels* play crucial roles in kernel mean embedding. The kernel k is said to be universal in the sense of Steinwart (2002) if the corresponding RKHS $\mathscr{H}$ is dense in $C_b(\mathcal{X})$, a space of bounded continuous functions on $\mathcal{X}$. That is, for any $f \in C_b(\mathcal{X})$ there exists a function $g \in \mathscr{H}$ and $\varepsilon > 0$ such that $\|f - g\|_\infty < \varepsilon$. This implies that in principle any kernel-based learning algorithms with universal kernels can approximate any bounded continuous function f arbitrarily well. The universal kernel is essential for kernel mean embedding as it was shown that for a universal kernel k, $\|\boldsymbol{\mu}_\mathbb{P} - \boldsymbol{\mu}_\mathbb{Q}\|_\mathscr{H} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, *i.e.*, the map $\boldsymbol{\mu}$ is injective (see Gretton et al. (2012a; Theorem 8) and Cortes et al. (2008)). In other words, there is no information loss when mapping the distribution into the Hilbert space. Examples of universal kernels on a compact domain are Gaussian and Laplace kernels. On the discrete domain, the kernel $k(\mathbf{x}, \mathbf{x}') = \mathbb{1}_{\{\mathbf{x}=\mathbf{x}'\}}$ is universal.

The characteristic kernel is defined as follows.

Definition 2.5. *The kernel k is said to be characteristic if the map $\boldsymbol{\mu}$ is injective. The RKHS $\mathscr{H}$ is said to be characteristic if its reproducing kernel is characteristic.*

The notion of characteristic kernel was first introduced in Fukumizu et al. (2008) being the kernels that satisfy Definition 2.5. Fukumizu et al. (2008) also 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). In particular, Sriperumbudur et al. (2008) shows that translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if the support of its Fourier transform is the entire $\mathbb{R}^d$. It follows immediately from Definition 2.5 that all universal kernels are characteristic, *e.g.*, Gaussian and Laplacian kernels, but not vice versa. That is, Definition 2.5 is not a sufficient condition for universal kernel. For the connection between universal and characteristic kernels, see Sriperumbudur et al. (2011a).

The notion of characteristic kernel is crucial in certain applications such as two-sample testing 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 distributions (Muandet et al. 2012, Muandet and Schölkopf 2013, Oliva et al. 2014, Szabó et al. 2015). In these applications, the notion of universal kernel is more important (Christmann and Steinwart 2010; Example 1). It is sometimes more favourable to interpret kernel k as a *weight function* which determines which frequency component occurs in the embedding (see Example 2.3). A shape of the kernel k in the Fourier domain can be more informative in these cases.

2.3.4 Maximum Mean Discrepancy and Its Applications

The kernel mean embedding has been used to define a metric for probability distributions which is important for many 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

$$\gamma[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{F}} \left\{ \int f(\mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) - \int f(\mathbf{y}) \, \mathrm{d}\mathbb{Q}(\mathbf{y}) \right\} \tag{2.53}$$

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 2.9. $\gamma[C_b(\mathcal{X}), \mathbb{P}, \mathbb{Q}] = 0$ *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 semi-norm 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,

$$\begin{aligned}
\text{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] &= \sup_{\|f\|_{\mathscr{H}}\leq 1} \left\{ \int f(\mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}) - \int f(\mathbf{y})\, \mathrm{d}\mathbb{Q}(\mathbf{y}) \right\} \\
&= \sup_{\|f\|_{\mathscr{H}}\leq 1} \left\{ \langle f, \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}) \rangle - \langle f, \int k(\mathbf{y}, \cdot)\, \mathrm{d}\mathbb{Q}(\mathbf{y}) \rangle \right\} \\
&= \sup_{\|f\|_{\mathscr{H}}\leq 1} \{ \langle f, \boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}} \rangle \} \\
&= \|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|^2_{\mathscr{H}} \qquad (2.54)
\end{aligned}$$

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

$$\text{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] = \mathbb{E}_{X,\tilde{X}}[k(X, \tilde{X})] - 2\mathbb{E}_{X,Y}[k(X, Y)] + \mathbb{E}_{Y,\tilde{Y}}[k(Y, \tilde{Y})] \qquad (2.55)$$

where $X, \tilde{X} \sim \mathbb{P}$ and $Y, \tilde{Y} \sim \mathbb{Q}$. It follows that $\text{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] = 0$ if and only if $\mathscr{H}$ is characteristic.

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

$$\widehat{\text{MMD}}_b[\mathscr{H}, \mathbf{X}, \mathbf{Y}] := \sup_{\|f\|_{\mathscr{H}}\leq 1} \left(\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y}_j) \right). \qquad (2.56)$$

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

$$\widehat{\text{MMD}}_u[\mathscr{H}, \mathbf{X}, \mathbf{Y}] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(\mathbf{y}_i, \mathbf{y}_j)$$

$$-\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(\mathbf{x}_i,\mathbf{y}_j). \tag{2.57}$$

Note that (2.57) is an unbiased estimate which is a sum of two U-statistics and a sample average (Serfling 1981; Chapter 5). The biased counterpart $\widehat{\mathrm{MMD}}_b[\mathscr{H},\mathbf{X},\mathbf{Y}]$ can be obtained using V-statistics. The convergence of empirical MMD has been established in Gretton et al. (2012a; Theorem 7).

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* $H_0 : \|\boldsymbol{\mu}(X) - \boldsymbol{\mu}(Y)\|_{\mathscr{H}} = 0$ against the *alternative hypothesis* $H_1 : \|\boldsymbol{\mu}(X) - \boldsymbol{\mu}(Y)\|_{\mathscr{H}} \neq 0$. 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 literature (Anderson et al. 1994, Biau and Györfi 2005, Nguyen et al. 2007). First, the MMD test is distribution free.[4] 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.

The MMD can be computed in quadratic time $O(n^2d)$, 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 (2.57), *i.e.*, drawing pairs from $\mathbf{X}$ and $\mathbf{Y}$ without replacement. This method reduces the time complexity of MMD from $O(n^2d)$ 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. Ji Zhao (2014) proposes an efficient test called *FastMMD* which employs the random Fourier feature to transform the MMD test with translation invariant kernel. The time complexity also reduces to $O(nd)$. For spherically invariant kernel, 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

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

$\hat{\mathbb{P}}$ 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*: $\sigma^2 = \text{median}\{\|\mathbf{x}_i - \mathbf{x}_j\|^2 : i, j = 1, \ldots, n\}$ (Schölkopf and Smola 2001). 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. 2007; 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. 2009a, Pan et al. 2011) and domain adaptation (Blanchard et al. 2011a, Muandet et al. 2013), selection bias correction (Huang et al. 2007), herding (Chen et al. 2010, Huszar and Duvenaud 2012), and Markov chain Monte Carlo (Sejdinovic et al. 2014), *etc*.

2.3.5 Recovering Information from Mean Embeddings

In this section we discuss two closely related problems, namely, distributional pre-image problem[5] (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 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.

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).

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

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 $\hat{\boldsymbol{\mu}}_X$ 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 $\boldsymbol{\mu}_{\mathbb{P}_{\boldsymbol{\theta}}}$ be its mean embedding in $\mathscr{H}$. One can find $\mathbb{P}_{\boldsymbol{\theta}}$ by the following minimization problem

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \|\hat{\boldsymbol{\mu}}_X - \boldsymbol{\mu}_{\mathbb{P}_{\boldsymbol{\theta}}}\|^2_{\mathscr{H}} \tag{2.58}$$

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 (2.58) reduces to the classical pre-image problem. The pre-image $\mathbf{x}$ can be viewed as a point estimate of the underlying distribution.

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 $\hat{\boldsymbol{\mu}}_X = \sum_{i=1}^n \beta_i \phi(\mathbf{x}_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. In this case, the optimization problem (2.58) reduces to

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{Q}\boldsymbol{\pi} + \boldsymbol{\pi}^\top \mathbf{R}\boldsymbol{\pi}\,, \tag{2.59}$$

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 (2.59) 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 $\boldsymbol{\mu} = \sum_{j=1}^m \beta_j \phi(\mathbf{z}_j)$ 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 $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$, and that the density of $\mathbb{P}$ can be estimated from $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ 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). This result is a generalization of the known result for functions in an RKHS.

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

$$\mathcal{E}_T^2 := \left\| \boldsymbol{\mu}_{\mathbb{P}} - \frac{1}{T}\sum_{t=1}^{T} \phi(\mathbf{x}_t) \right\|_{\mathscr{H}}^2 = \|\boldsymbol{\mu}_{\mathbb{P}} - \hat{\boldsymbol{\mu}}_T\|_{\mathscr{H}}^2 , \tag{2.60}$$

where $\hat{\boldsymbol{\mu}}_T$ 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 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 (2.60) 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.

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.

2.3.6 Approximating the Kernel Mean Embedding

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 sub-sample, 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

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathscr{H}} \approx \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{y}) \tag{2.61}$$

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 shift-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

$$\tilde{\boldsymbol{\mu}}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}(\mathbf{x}_i) \in \mathbb{R}^m. \tag{2.62}$$

Since $\mathbf{z}(\mathbf{x}_i) \in \mathbb{R}^m$ for all i, so does $\tilde{\boldsymbol{\mu}}_{\mathbb{P}}$. 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 (2.62) 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. (2015b; Lemma 1). Also, the random features are limited to only a certain class of kernel functions.

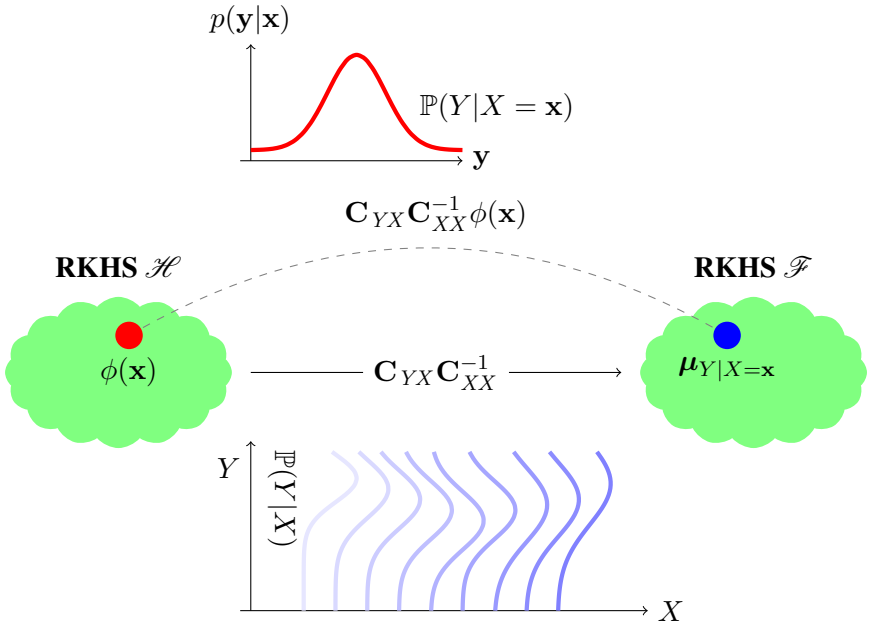


Figure 2.4: From marginal distribution to conditional distribution: Unlike the embeddings discussed in the previous chapter, the embedding of conditional distribution $\mathbb{P}(Y|X)$ is not a single element in the RKHS. Instead, it may be viewed as a family of Hilbert space embeddings of the conditional 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 §2.4.4 that there is a natural interpretation in a vector-valued regression framework.

2.4 Kernel Mean Embedding of Conditional Distributions

In the previous chapter, I discuss the embedding of marginal distributions in RKHS and gives comprehensive reviews on various applications. Throughout this section I will extend the language of kernel mean embedding developed in the previous section 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 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, van Hoof et al. 2015), 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 2.4 gives a schematic illustration of conditional mean embedding.

2.4.1 From Marginal to Conditional

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, I 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 cloudiness 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 cloudiness and rain. Following the above definitions, one may subsequently ask given that it is cloudy outside, *i.e.*, $X =$ 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 $X = x$ occurs. In other words, conditional probabilities allow us to reason about causality.[6]

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

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}. \tag{2.63}$$

As we can see in the first equation of (2.63), 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 (2.63) provide the powerful language for statistical inference in statistics and machine learning.

Conditional mean embedding. 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 that they satisfy

$$\mathcal{U}_{Y|\mathbf{x}} = \mathbb{E}_{Y|\mathbf{x}}[\varphi(Y)|X = \mathbf{x}] = \mathcal{U}_{Y|X} k(\mathbf{x}, \cdot) \tag{2.64}$$

$$\mathbb{E}_{Y|\mathbf{x}}[g(Y)|X = \mathbf{x}] = \langle g, \mathcal{U}_{Y|\mathbf{x}} \rangle_{\mathcal{G}}, \quad \forall g \in \mathscr{H}_Y. \tag{2.65}$$

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 (2.64) says that the conditional mean embedding of $\mathbb{P}(Y|X = \mathbf{x})$ should corresponds 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 2.4). Condition (2.65) 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 explicits form of $\mathcal{U}_{Y|X}$ and $\mathcal{U}_{Y|\mathbf{x}}$.

Definition 2.6 (Song et al. (2009; 2013))**.** *Let $\mathbf{C}_{XX} : \mathscr{H} \to \mathscr{H}$ and $\mathbf{C}_{YX} : \mathscr{H} \to \mathscr{F}$ be covariance operator on X and cross-covariance operator from X to Y, respective. Then, the conditional mean embedding $\mathcal{U}_{Y|X}$ and $\mathcal{U}_{Y|\mathbf{x}}$ are defined as*

$$\mathcal{U}_{Y|X} := \mathbf{C}_{YX}\mathbf{C}_{XX}^{-1} \tag{2.66}$$

$$\mathcal{U}_{Y|\mathbf{x}} := \mathbf{C}_{YX}\mathbf{C}_{XX}^{-1} k(\mathbf{x}, \cdot). \tag{2.67}$$

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

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 2.6 satisfies both (2.64) and (2.65). This result follows from Fukumizu et al. (2004; Theorem 2). One should also keep in mind that, unlike the marginal mean embedding, the operator $\mathbf{C}_{YX}\mathbf{C}_{XX}^{-1}$ 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 $\mathbb{E}_{Y|X}[g(Y)|X]$ is an element of $\mathscr{H}$ may not hold in general (Fukumizu et al. 2004, Song et al. 2009).[7]

Since the joint distribution $\mathbb{P}(X, Y)$ is unknown in practice, we cannot compute $\mathbf{C}_{XX}$ and $\mathbf{C}_{YX}$ 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

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

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 $\widehat{\mathbf{C}}_{XX}$ is a compact operator, we need a regularizer $\lambda\mathcal{I}$ for the inverse of $\widehat{\mathbf{C}}_{XX}$ to be well-posed. Another possibility is to employ the spectral filtering algorithms, *i.e.*, $\hat{\boldsymbol{\mu}} = \Phi g_\lambda(\mathbf{K})\mathbf{k}_\mathbf{x}$ 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 2.10 gives a formal characterization on the empirical estimator of conditional mean embedding.

Theorem 2.10 (Song et al. (2009))**.** *The conditional mean embedding $\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}}$ can be estimated as*

$$\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}} = \Phi(\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{k}_\mathbf{x}. \tag{2.68}$$

Interestingly, we may write (2.68) as $\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}} = \Phi\boldsymbol{\beta} = \sum_{i=1}^m \beta_i\varphi(\mathbf{y}_i)$ where $\boldsymbol{\beta} := (\mathbf{K} + n\lambda\mathbf{I}_n)^{-1}\mathbf{k}_\mathbf{x} \in \mathbb{R}^n$. That is, it can be written 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 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 $O_p(n^{-1/2})$.

2.4.2 Basic Operations on Kernel Mean Embedding

In this section I 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

[7] If X and Y are discrete random variables and the kernels are characteristic, $\mathbb{E}_{Y|X}[g(Y)|X] \in \mathscr{H}$.

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 distribution.

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:

$$\text{Sum rule:} \quad \mathbb{P}(X) = \sum_{Y} \mathbb{P}(X,Y) \tag{2.69}$$

$$\text{Product rule:} \quad \mathbb{P}(X,Y) = \mathbb{P}(Y|X)\mathbb{P}(X) \tag{2.70}$$

Combining (2.69) and (2.70) 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 (2.69) 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, I will show how these two 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).

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

$$\boldsymbol{\mu}_X = \mathbb{E}_Y[\mathcal{U}_{X|Y}\varphi(Y)] = \mathcal{U}_{X|Y}\mathbb{E}_Y[\varphi(Y)] = \mathcal{U}_{X|Y}\boldsymbol{\mu}_Y. \tag{2.71}$$

Product rule. Consider a tensor product of the joint feature map $\phi(X) \otimes \varphi(Y)$. We can then factor $\boldsymbol{\mu}_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \varphi(Y)]$ according to the law of 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 $\boldsymbol{\mu}_X^{\otimes} := \mathbb{E}_X[\phi(X) \otimes \phi(X)]$ and $\boldsymbol{\mu}_Y^{\otimes} := \mathbb{E}_Y[\varphi(Y) \otimes \varphi(Y)]$. Then, we can write the product rule in terms of kernel mean embeddings as

$$\boldsymbol{\mu}_{XY} = \mathcal{U}_{X|Y}\boldsymbol{\mu}_Y^{\otimes} = \mathcal{U}_{Y|X}\boldsymbol{\mu}_X^{\otimes}. \tag{2.72}$$

On the one hand, both (2.71) and (2.72) do not require any parametric assumption on the underlying distributions. On the other hand, these operations can practically be both statistically difficult and computationally costly in some applications (more below).

Bayes' rule. An extension of Bayes' rule called *kernel Bayes' rule* has been proposed in Fukumizu et al. (2013) which provides a mathematical tool to obtain an embedding $\boldsymbol{\mu}_{\mathbb{P}(\theta|\mathbf{X})}$ of posterior $\mathbb{P}(\theta|\mathbf{X})$ from the embeddings of prior $\Pi(\theta)$ and the likelihood $\mathbb{P}(\mathbf{X}|\theta)$. See Fukumizu et al. (2013) for technical details.

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

2.4.3 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 distribution 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 2.4.2 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 do 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 space (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 2.3.5 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 $\mathbb{P}(s_1, \ldots, s_T, o_1, \ldots, o_T)$ where s_t is the hidden state at timestep t and 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 $\mathbb{P}(o_1, s_1) \prod_t \mathbb{P}(o_t|s_t)\mathbb{P}(s_t|s_{t-1})$. Thus, the system is characterized by two important models, namely, a *transition model* $\mathbb{P}(s_t|s_{t-1})$ which describes the evolution of the system and the *observation model* $\mathbb{P}(o_t|s_t)$ 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.*, $\mathbb{P}(s_{t+1}|h_{t+1})$ where $h_t = (o_1, \ldots, o_t)$. The distribution $\mathbb{P}(s_{t+1}|h_{t+1})$ can be obtained in two steps. First, we *update* the distribution by $\mathbb{P}(s_{t+1}|h_t) = \mathbb{E}_{s_t|h_t}[\mathbb{P}(s_{t+1}|s_t)|h_t]$. Then, we *condition* the distribution on a new observation o_{t+1} using Bayes rule to obtain $\mathbb{P}(s_{t+1}|h_t o_{t+1}) \propto \mathbb{P}(o_{t+1}|s_{t+1})\mathbb{P}(s_{t+1}|h_t)$. 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 need 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)). Empirically, although requiring labeled sequence of observations to perform filtering, 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* $\mathcal{O}_{ij} = \mathbb{P}(h_{t+1} = i|h_t = j)\mathbb{P}(X_t = x_t|h_t = j)$). 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 in 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.

Lastly, 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.

2.4.4 Regression Perspectives

As illustrated in Figure 2.4, 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 I review below.

Recall that the conditional mean embedding is defined via $\mathbb{E}[g(Y)|X = \mathbf{x}] = \langle g, \hat{\boldsymbol{\mu}}_{Y|\mathbf{x}} \rangle_{\mathscr{F}}$. That is, for every $\mathbf{x} \in \mathcal{X}$, $\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}}$ is a function on $\mathcal{Y}$ and thereby defines a mapping from $\mathcal{X}$ to $\mathscr{F}$. Furthermore, the empirical estimator in (2.68) can be expressed as $\hat{\boldsymbol{\mu}}_{Y|\mathbf{x}} = \Phi(\mathbf{K} +$

$n\lambda \mathbf{I}_n)^{-1}\mathbf{k}_{\mathbf{x}}$, which already suggests that the conditional mean embedding is the solution to an underlying regression problem. Given a sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n) \in \mathcal{X} \times \mathscr{F}$, a vector-valued regression problem can be formulated as

$$\widehat{\mathcal{E}}_\lambda(f) = \sum_{i=1}^{n} \|\mathbf{y}_i - f(\mathbf{x}_i)\|_{\mathscr{F}}^2 + \lambda \|f\|_{\mathscr{H}_\Gamma}^2 \tag{2.73}$$

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 $\hat{\boldsymbol{\mu}}_{Y|X}$ can be obtained as a minimizer of the optimization of the form (2.73).[8]

Following the analysis of Grünewälder et al. (2012), a natural optimization problem for the conditional mean embedding is to find a function $\boldsymbol{\mu} : \mathcal{X} \to \mathscr{F}$ that minimizes the following objective:

$$\mathcal{E}[\boldsymbol{\mu}] = \sup_{\|g\|_{\mathscr{F}} \leq 1} \mathbb{E}_X \left[(\mathbb{E}_Y[g(Y)|X] - \langle g, \boldsymbol{\mu}(X) \rangle_{\mathscr{F}})^2 \right] \tag{2.74}$$

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

$$\mathcal{E}_s[\boldsymbol{\mu}] = \mathbb{E}_{(X,Y)} \left[\|l(Y, \cdot) - \boldsymbol{\mu}(X)\|_{\mathscr{F}}^2 \right], \tag{2.75}$$

which can then be replaced by its empirical counterpart

$$\widehat{\mathcal{E}}_s[\boldsymbol{\mu}] = \sum_{i=1}^{n} \|l(\mathbf{y}_i, \cdot) - \boldsymbol{\mu}(\mathbf{x}_i)\|_{\mathscr{F}}^2 + \lambda \|\boldsymbol{\mu}\|_{\mathscr{H}_\Gamma}^2. \tag{2.76}$$

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 $\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} \Gamma_{\mathbf{x}_i} c_i$ 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 $\hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} \Gamma_{\mathbf{x}_i} (\mathbf{K} + \lambda \mathbf{I})^{-1} l(\mathbf{y}_j, \cdot)$ which is exactly the embedding in (2.68). 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 $O(\log(n)/n)$ compared to the state-of-the-art rates of $O(n^{-1/4})$ of Song et al. (2009). However, it is important to note that the analysis is done under the assumption that the RKHS $\mathscr{F}$ is finite dimensional.

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 (2.76) (see, *e.g.*, Muandet et al.

[8] 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.

(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).

2.5 Relationships between Mean Embedding and Other Methods

I conclude this chapter by discussing the relationships between kernel mean embedding and other methods across different disciplines.

Kernel mean has played a fundamental role in most kernel algorithms since the beginning of the field itself. Classical algorithms for classification and anomaly detection employed a mean function in the RKHS as their building block. 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 the class-conditional means $\hat{\boldsymbol{\mu}}_{\{y=+1\}} := \frac{1}{n}\sum_{y=+1}\phi(\mathbf{x}_i)$ and $\hat{\boldsymbol{\mu}}_{\{y=-1\}} := \frac{1}{m}\sum_{y=-1}\phi(\mathbf{x}_i)$. 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 $\hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)$ 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 $\hat{\boldsymbol{\mu}}_{\{y=+1\}}$ and $\hat{\boldsymbol{\mu}}_{\{y=-1\}}$ 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).

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; 2013) 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

$$D_E(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{XY}\|X-Y\| - \mathbb{E}_{XX'}\|X-X'\| - \mathbb{E}_{YY'}\|Y-Y'\|, \tag{2.77}$$

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 the 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.[9] For the energy distance, the equivalence holds if the energy distances are computed with semi-metric of negative type (Sejdinovic et al. 2012; Theorem 11). Fundamentally, the finding is that the kernel-based and distance-based methods are equivalent if we allow "distance" ρ that may not satisfy the triangle inequality. However, since distance kernels are continuous but unbounded functions, one need to restrict the class of distributions for which kernel embeddings are well-defined, *i.e.*, to ensure that $\mathbb{E}_X k(X,X) < \infty$.

Harmeling et al. (2013) establishes a link between Fourier optics and kernel mean embedding from computer vision viewpoint. A simple imaging system can be described by the so-called *incoherent imaging equation*

$$q(\mathbf{u}) = \int f(\mathbf{u}-\boldsymbol{\xi})p(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \tag{2.78}$$

[9] 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)).

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 (2.78), 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).

The kernel mean embedding can also be understood probabilistically. Let consider the following example.[10] 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

$$\mathbb{E}_{\mathbb{P}(\mathbf{x})\mathrm{GP}(f)}[f(\mathbf{x})f(\cdot)] = \iint f(\mathbf{x})f(\cdot)\mathrm{GP}(f)p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}f = \int k(\mathbf{x},\cdot)p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \boldsymbol{\mu}_{\mathbb{P}}.$$

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 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.

2.6 Discussions

To conclude, the extensive literature review clearly indicates that kernel mean embedding of distributions has made a tremendous impact in a wide range of disciplines including statistics, control, causality, and computer vision. Moreover, it also demonstrates the potential of kernel methods in giving rise to modern applications in machine learning community. Despite these successes, there remain several open questions, some of which will be addressed in the following chapters. Along the way, I will also provide suggestions for future research.

END OF CHAPTER 2

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

Chapter 3

Kernel Mean Shrinkage Estimators

It is apparent that most practical applications of kernel mean embedding must rely on its empirical estimate $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ instead of $\boldsymbol{\mu}_{\mathbb{P}}$. This chapter provides a thorough analysis of kernel mean estimation problem, and proposes a novel class of estimators called *kernel mean shrinkage estimator (KMSE)* which improve upon the standard empirical average.

3.1 Introduction

Recall that a kernel mean is defined w.r.t. a probability distribution $\mathbb{P}$ over a measurable space $\mathcal{X}$ by

$$\boldsymbol{\mu}_{\mathbb{P}} := \int_{\mathcal{X}} k(\mathbf{x}, \cdot) \ \mathrm{d}\mathbb{P}(\mathbf{x}) \in \mathscr{H}, \tag{3.1}$$

where $\boldsymbol{\mu}_{\mathbb{P}}$ is a Bochner integral and $\mathscr{H}$ is a reproducing kernel Hilbert space (RKHS) endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is measurable. Relying on an i.i.d sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from $\mathbb{P}$, an estimate of the true kernel mean is the empirical average

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \cdot) \,. \tag{3.2}$$

We refer to this estimator as a *kernel mean estimator* (KME). The contribution of this chapter is to show that there exist estimators that can improve upon this standard estimator.

I will show that the empirical estimator in (3.2) is, in a certain sense, not optimal, *i.e.*, there exist "better" estimators (more below), and then propose simple estimators that outperform the empirical estimator. While it is reasonable to argue that $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ is the "best" possible estimator of $\boldsymbol{\mu}_{\mathbb{P}}$ if nothing is known about $\mathbb{P}$ (in fact $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ is minimax in the sense of van der Vaart (1998; Theorem 25.21, Example 25.24)), we show that "better" estimators of $\boldsymbol{\mu}_{\mathbb{P}}$ can be constructed if mild assumptions are made on $\mathbb{P}$. This work is to some extent inspired by Stein's seminal work in 1955, which showed that the maximum likelihood estimator (MLE) of the mean, $\boldsymbol{\theta}$ of a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I})$ is "inadmissible" (Stein 1955)—*i.e.*, there exists a better estimator—though it is minimax optimal. In particular, Stein showed that there exists an estimator that always achieves smaller total mean squared error regardless of the true $\boldsymbol{\theta} \in \mathbb{R}^d$, when $d \geq 3$. Perhaps the best known estimator of such kind is James-Stein estimator (James and Stein 1961).

3.2 Estimation of the Mean of Multivariate Normal Distribution

I will first discuss the basic problem of estimating the mean vector of a multivariate normal distribution (Stein 1955). I will also give an explicit form of *James-Stein estimator* (James and Stein 1961), which serves as a motivation for our kernel mean estimators.

3.2.1 Basic Setup

Our goal is to estimate from an observation $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top \in \mathbb{R}^d$ of a d-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with known covariance matrix $\mathbf{C} = \sigma^2\mathbf{I}$, under the loss function

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \mathbf{A}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}), \tag{3.3}$$

where $\mathbf{A}$ is a positive definite matrix and $\hat{\boldsymbol{\mu}}$ denotes an estimate of $\boldsymbol{\mu}$ obtained from the observation $\mathbf{x}$. Throughout I will focus only on the case of $\mathbf{A} = \mathbf{I}$ such that the loss function becomes a square loss $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. We measure the goodness of an estimator by the associated risk function

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) := \mathbb{E}[L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})] = \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2, \tag{3.4}$$

which in this case is a *mean square error (MSE)*. An estimator $\hat{\boldsymbol{\mu}}$ is said to be *inadmissible* if there exists an estimator $\tilde{\boldsymbol{\mu}}$ such that $\mathbb{E}\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ for all $\boldsymbol{\mu}$ and there exists at least one $\boldsymbol{\mu}$ for which the strict inequality holds. In which case, $\tilde{\boldsymbol{\mu}}$ is said to be *better* than $\hat{\boldsymbol{\mu}}$. The estimator $\hat{\boldsymbol{\mu}}$ is *admissible* if there exists no better estimator.

Without any prior knowledge about $\boldsymbol{\mu}$, the most natural estimator of $\boldsymbol{\mu}$ is the observation itself, *i.e.*,

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}}(\mathbf{x}) = \mathbf{x}. \tag{3.5}$$

It is known that $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ is a maximum likelihood, minimum variance unbiased, invariant, and minimax estimator for $\boldsymbol{\mu}$ (Lehmann and Casella 1998). Nevertheless, as will be shown below, this estimator is inadmissible in the sense that there exist other estimators whose risk is everywhere smaller than the risk of $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$. It is not difficult to show that the risk of $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ is constant, *i.e.*, $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_{\mathrm{ML}}) = d\sigma^2$, regardless of $\boldsymbol{\mu}$.

3.2.2 James-Stein Estimator

James and Stein (1961) proposed an estimator of the mean of a multivariate normal distribution that is better in mean squared error than the sample mean. The estimator is given by

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}}(\mathbf{x}) = \left(1 - \frac{(d-2)\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x}. \tag{3.6}$$

To show that James-Stein estimator is uniformly better than the standard maximum likelihood estimator, I first provide a renowned Stein's lemma (see Appendix C.1 for the proof).

Lemma 3.1 (Stein's lemma)**.** *Let X be a standard normally distributed random variable and $g : \mathbb{R} \to \mathbb{R}$ be an absolutely continuous function such that $\mathbb{E}|g'(X)| < \infty$. Then,*

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)]. \tag{3.7}$$

By virtue of Lemma 3.1, we are now in a position to show that $\hat{\boldsymbol{\mu}}_{\mathrm{JS}}(\mathbf{x})$ is better than $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}(\mathbf{x})$.

Theorem 3.2. *The James-Stein estimator $\hat{\boldsymbol{\mu}}_{JS}$ dominates the ML estimator $\hat{\boldsymbol{\mu}}_{ML}$ everywhere in terms of MSE.* i.e.*, for all $\boldsymbol{\mu} \in \mathbb{R}^d$, $d > 2$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{JS} - \boldsymbol{\mu}\|^2 \leq \mathbb{E}\|\hat{\boldsymbol{\mu}}_{ML} - \boldsymbol{\mu}\|^2, \tag{3.8}$$

where the strict inequality holds for at least one $\boldsymbol{\mu}$.

Proof of Theorem 3.2. Let $\hat{\boldsymbol{\mu}}_g := \hat{\boldsymbol{\mu}} + g(\hat{\boldsymbol{\mu}})$ for some smooth function g. Then, we have

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}\|^2] &= \mathbb{E}[\|\hat{\boldsymbol{\mu}} + g(\hat{\boldsymbol{\mu}}) - \boldsymbol{\mu}\|^2] \\
&= \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] + 2\mathbb{E}[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top g(\hat{\boldsymbol{\mu}})] + \mathbb{E}[\|g(\hat{\boldsymbol{\mu}})\|^2] \\
&\overset{(*)}{=} \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] + 2\sigma^2\mathbb{E}[\mathrm{div} g(\hat{\boldsymbol{\mu}})] + \mathbb{E}[\|g(\hat{\boldsymbol{\mu}})\|^2] \\
&= \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] + \mathbb{E}\left[\|g(\hat{\boldsymbol{\mu}})\|^2 + 2\sigma^2 \mathrm{div} g(\hat{\boldsymbol{\mu}})\right] \qquad (3.9)
\end{aligned}$$

where we use Lemma 3.1 to obtain $(*)$. This is known as Stein's identity. The second term of the rhs of (3.9) no longer depends on $\boldsymbol{\mu}$.

Recall that for James-Stein estimator, we have

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}} = \left(1 - \frac{c\sigma^2}{\|\hat{\boldsymbol{\mu}}\|^2}\right)\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}} - \frac{c\sigma^2}{\|\hat{\boldsymbol{\mu}}\|^2}\hat{\boldsymbol{\mu}}, \qquad \text{and} \qquad g(\hat{\boldsymbol{\mu}}) = -\frac{c\sigma^2}{\|\hat{\boldsymbol{\mu}}\|^2}\hat{\boldsymbol{\mu}}\,.$$

This gives

$$\|g(\hat{\boldsymbol{\mu}})\|^2 = \frac{c^2\sigma^4}{\|\hat{\boldsymbol{\mu}}\|^2}, \qquad \text{and} \qquad \mathrm{div} g(\hat{\boldsymbol{\mu}}) = -\frac{cd\sigma^2 + 2c\sigma^2}{\|\hat{\boldsymbol{\mu}}\|^2}\,.$$

Putting everything together gives

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{\mu}}_{\mathrm{JS}} - \boldsymbol{\mu}\|^2] &= \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] + \mathbb{E}\left[\|g(\hat{\boldsymbol{\mu}})\|^2 + 2\sigma^2 \mathrm{div} g(\hat{\boldsymbol{\mu}})\right] \\
&= d\sigma^2 + \mathbb{E}\left[\frac{c\sigma^4(c - 2(d-2))}{\|\hat{\boldsymbol{\mu}}\|^2}\right]
\end{aligned}$$

Consequently, for any $c \in (0, 2(d-2))$, it follows that

$$\mathbb{E}[\|\hat{\boldsymbol{\mu}}_{\mathrm{JS}} - \boldsymbol{\mu}\|^2] \leq \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]\,.$$

This concludes the proof. ■

Interestingly, the James-Stein estimator is itself inadmissible, and there exists a wide class of estimators that outperform the MLE, see, *e.g.*, Berger (1976). Ultimately, Stein's result suggests that one can construct estimators better than the usual empirical estimator if the relevant parameters are estimated jointly and if the definition of risk ultimately looks at all of these parameters (or coordinates) together. This finding is quite remarkable as it is counter-intuitive as to why joint estimation should yield better estimators when all parameters are mutually independent (Efron and Morris 1977). Although the Stein phenomenon has been extensively studied in the statistics community, it has not received much attention in the machine learning community.

The James-Stein estimator is a special case of a larger class of estimators known as *shrinkage estimators* (Gruber 1998). In its most general form, the shrinkage estimator is a combination of a model with low bias and high variance, and a model with high bias but low variance. For example, one might consider the following estimator:

$$\hat{\boldsymbol{\mu}}_{\mathrm{shrink}} \triangleq \lambda\tilde{\boldsymbol{\mu}} + (1-\lambda)\hat{\boldsymbol{\mu}}_{\mathrm{ML}},$$

where $\lambda \in [0,1]$, $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ denotes the usual maximum likelihood estimate of $\boldsymbol{\mu}$, and $\tilde{\boldsymbol{\mu}}$ is an arbitrary point in the input space. In the case of James-Stein estimator, we have $\tilde{\boldsymbol{\mu}} = 0$. Our proposal of shrinkage estimator to estimate $\boldsymbol{\mu}_{\mathbb{P}}$ will rely on the same principle. However, our work differs fundamentally from the Stein's seminal works and those along this line in two aspects. First, our setting is "non-parametric" in the sense that we do not assume any parametric form for the distribution, whereas most of traditional works focus on some specific distributions, *e.g.*, the

Gaussian distribution. The non-parametric setting is very important in most applications of kernel means because it allows us to perform statistical inference without making any assumption on the parametric form of the true distribution $\mathbb{P}$. Second, our setting involves a "non-linear feature map" into a high-dimensional space. For example, if we use the Gaussian RBF kernel, the mean function $\boldsymbol{\mu}_{\mathbb{P}}$ lives in an infinite-dimensional space. As a result, higher moments of the distribution come into play and therefore one cannot adopt Stein's setting straightforwardly as it involves only the first moment. A direct generalization of James-Stein estimator to infinite-dimensional Hilbert space has been considered, for example, in Berger and Wolpert (1983), Mandelbaum and Shepp (1987), Privault and Réveillac (2008). In those works, the parameter to be estimated is assumed to be the mean of a Gaussian measure on the Hilbert space from which samples are drawn. In contrast, our setting involves samples that are drawn from $\mathbb{P}$ defined on an arbitrary measurable space, and not from a Gaussian measure defined on a Hilbert space.

3.3 Improving Kernel Mean Estimation via Shrinkage

3.3.1 Our Setup

We assume throughout the paper that we observe a sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ of size n drawn independently and identically (i.i.d.) from some unknown distribution $\mathbb{P}$ over a measurable space $\mathcal{X}$. Denote by $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ the true kernel mean (3.1) and its empirical estimate (3.2) respectively. We remove the subscript for ease of notation, but we will use $\boldsymbol{\mu}_{\mathbb{P}}$ (resp. $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$) and $\boldsymbol{\mu}$ (resp. $\hat{\boldsymbol{\mu}}$) interchangeably. We measure the quality of an estimator $\tilde{\boldsymbol{\mu}} \in \mathscr{H}$ of $\boldsymbol{\mu}$ by the risk function, $R : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$, $R(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}) = \mathbb{E}\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^2_{\mathscr{H}}$, where $\mathbb{E}$ denotes the expectation over the choice of random sample of size n drawn i.i.d. from the distribution $\mathbb{P}$. When $\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}$, for the ease of notation, we will use Δ to denote $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$, which can be rewritten as

$$\Delta = \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2_{\mathscr{H}} = \frac{1}{n}\left(\mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}})\right), \tag{3.10}$$

where $\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}[k(\mathbf{x}, \tilde{\mathbf{x}})] \triangleq \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[\mathbb{E}_{\tilde{\mathbf{x}}\sim\mathbb{P}}[k(\mathbf{x}, \tilde{\mathbf{x}})]]$ with $\mathbf{x}$ and $\tilde{\mathbf{x}}$ being independent copies.

Instead of $\hat{\boldsymbol{\mu}}$, in this chapter we propose and investigate the following kernel mean estimator

$$\hat{\boldsymbol{\mu}}_\alpha \triangleq \alpha f^* + (1 - \alpha)\hat{\boldsymbol{\mu}} \tag{3.11}$$

where $\alpha \geq 0$ and f^* is a fixed, but arbitrary function in $\mathscr{H}$. Basically, it is a shrinkage estimator that shrinks the empirical estimator toward a function f^* by an amount specified by α. The choice of f^* can be arbitrary, but we will assume that f^* is chosen independent of the sample. If $\alpha = 0$, the estimator $\hat{\boldsymbol{\mu}}_\alpha$ reduces to the empirical estimator $\hat{\boldsymbol{\mu}}$. We denote by Δ_α the risk of the shrinkage estimator in (3.11), *i.e.*, $\Delta_\alpha \triangleq R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha)$.

The following theorem asserts that the shrinkage estimator $\hat{\boldsymbol{\mu}}_\alpha$ achieves smaller risk than that of the empirical estimator $\hat{\boldsymbol{\mu}}$ given an appropriate choice of α, regardless of the function f^*.

Theorem 3.3. *For all distributions $\mathbb{P}$ and kernel k satisfying $\int k(\mathbf{x}, \mathbf{x})\, \mathrm{d}\mathbb{P}(\mathbf{x}) < \infty$, $\Delta_\alpha < \Delta$ if and only if*

$$\alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}}\right). \tag{3.12}$$

In particular, $\arg\min_{\alpha\in\mathbb{R}}(\Delta_\alpha - \Delta)$ *is unique and is given by* $\alpha_* \triangleq \frac{\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}}$.

Proof of Theorem 3.3. Note that

$$\Delta_\alpha = \mathbb{E}\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|^2_{\mathscr{H}} = \|\mathbb{E}[\hat{\boldsymbol{\mu}}_\alpha] - \boldsymbol{\mu}\|^2_{\mathscr{H}} + \mathbb{E}\,\|\hat{\boldsymbol{\mu}}_\alpha - \mathbb{E}\hat{\boldsymbol{\mu}}_\alpha\|^2_{\mathscr{H}} = \|\mathrm{Bias}(\hat{\boldsymbol{\mu}}_\alpha)\|^2_{\mathscr{H}} + \mathrm{Var}(\hat{\boldsymbol{\mu}}_\alpha),$$

where

$$\mathrm{Bias}(\hat{\boldsymbol{\mu}}_\alpha) = \mathbb{E}[\hat{\boldsymbol{\mu}}_\alpha] - \boldsymbol{\mu} = \mathbb{E}[\alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}] - \boldsymbol{\mu} = \alpha(f^* - \boldsymbol{\mu})$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\mu}}_\alpha) = (1-\alpha)^2 \mathbb{E}\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right\|_{\mathscr{H}}^2 = (1-\alpha)^2 \Delta.$$

Therefore,

$$\Delta_\alpha = \alpha^2 \left\|f^* - \boldsymbol{\mu}\right\|_{\mathscr{H}}^2 + (1-\alpha)^2 \Delta, \tag{3.13}$$

i.e., $\Delta_\alpha - \Delta = \alpha^2 \left[\Delta + \|f^* - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] - 2\alpha\Delta$. This is clearly negative if and only if (3.12) holds and is uniquely minimized at $\alpha_* \triangleq \frac{\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|_{\mathscr{H}}^2}$. ■

Remark 3.1. *The following observations follow immediately from Theorem 3.3:*

(i) The shrinkage estimator always improves upon the standard one regardless of the direction of shrinkage, as specified by the choice of f^*. In other words, there exists a wide class of kernel mean estimators that achieve smaller risk than the standard one.

(ii) The range of α depends on the choice of f^*. The further f^* is from $\boldsymbol{\mu}$, the smaller the range of α becomes. Thus, the shrinkage gets smaller if f^* is chosen such that it is far from the true kernel mean. This effect is akin to James-Stein estimator.

(iii) From (3.12)*, since* $0 < \alpha < 2$, i.e., $0 < (1-\alpha)^2 < 1$*, it follows that* $\mathrm{Var}(\hat{\boldsymbol{\mu}}_\alpha) < \mathrm{Var}(\hat{\boldsymbol{\mu}}) = \Delta$, i.e.*, the shrinkage estimator always improves upon the empirical estimator in terms of the variance. Further improvement can be gained by reducing the bias by incorporating the prior knowledge about the location of $\boldsymbol{\mu}$ via f^*. This implies that we can potentially gain "twice" by adopting the shrinkage estimator: by reducing variance of the estimator and by incorporating prior knowledge in choosing f^* such that it is close to the true kernel mean.*

While Theorem 3.3 shows $\hat{\boldsymbol{\mu}}$ to be inadmissible by providing a family of estimators that are better than $\hat{\boldsymbol{\mu}}$, the result is not useful as all these estimators require the knowledge of $\boldsymbol{\mu}$ (which is the parameter of interest) through the range of α given in (3.12). In Section 3.3.2, we investigate Theorem 3.3 and show that $\hat{\boldsymbol{\mu}}_\alpha$ can be constructed under some weak assumptions on $\mathbb{P}$, without requiring the knowledge of $\boldsymbol{\mu}$.

From (3.12), the existence of positive α is guaranteed if and only if the risk of the empirical estimator is non-zero. Under some assumptions on k, the following result shows that $\Delta = 0$ if and only if the distribution $\mathbb{P}$ is a Dirac distribution, *i.e.*, the distribution $\mathbb{P}$ is a point mass. This result ensures, in many non-trivial cases, a non-empty range of α for which $\Delta_\alpha - \Delta < 0$.

Proposition 3.4. *Let $k(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x}-\mathbf{y}),\ \mathbf{x},\mathbf{y} \in \mathbb{R}^d$ be a characteristic kernel where $\psi \in C_b(\mathbb{R}^d)$ is positive definite. Then $\Delta = 0$ if and only if $\mathbb{P} = \delta_{\mathbf{x}}$ for some $\mathbf{x} \in \mathbb{R}^d$.*

Proof of Proposition 3.4. ($\Rightarrow$) If $\mathbb{P} = \delta_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{X}$, then $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} = k(\cdot,\mathbf{x})$ and thus $\Delta = 0$.
($\Leftarrow$) Suppose $\Delta = 0$. It follows from (3.10) that $\iint (k(\mathbf{x},\mathbf{x}) - k(\mathbf{x},\mathbf{y}))\,\mathrm{d}\mathbb{P}(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{y}) = 0$. Since k is translation invariant, this reduces to

$$\iint (\psi(0) - \psi(\mathbf{x}-\mathbf{y}))\,\mathrm{d}\mathbb{P}(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{y}) = 0.$$

By invoking Bochner's theorem, which states that ψ is the Fourier transform of a non-negative finite Borel measure Λ, *i.e.*, $\psi(\mathbf{x}) = \int e^{-i\mathbf{x}^\top \boldsymbol{\omega}}\,\mathrm{d}\Lambda(\boldsymbol{\omega}),\ \mathbf{x} \in \mathbb{R}^d$, we obtain (see (16) in the proof of Proposition 5 in Sriperumbudur et al. (2011a))

$$\iint \psi(\mathbf{x}-\mathbf{y})\,\mathrm{d}\mathbb{P}(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{y}) = \int |\varphi_{\mathbb{P}}(\boldsymbol{\omega})|^2\,\mathrm{d}\Lambda(\boldsymbol{\omega}),$$

thereby yielding

$$\int (|\varphi_{\mathbb{P}}(\boldsymbol{\omega})|^2 - 1)\, \mathrm{d}\Lambda(\boldsymbol{\omega}) = 0, \tag{3.14}$$

where $\varphi_{\mathbb{P}}$ is the characteristic function of $\mathbb{P}$. Note that $\varphi_{\mathbb{P}}$ is uniformly continuous and $|\varphi_{\mathbb{P}}| \leq 1$. Since k is characteristic, Theorem 9 in Sriperumbudur et al. (2010) implies that $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, using which in (3.14) yields $|\varphi_{\mathbb{P}}(\boldsymbol{\omega})| = 1$ for all $\boldsymbol{\omega} \in \mathbb{R}^d$. Since $\varphi_{\mathbb{P}}$ is positive definite on $\mathbb{R}^d$, it follows from Sasvári (2013; Lemma 1.5.1) that $\varphi_{\mathbb{P}}(\boldsymbol{\omega}) = e^{\sqrt{-1}\boldsymbol{\omega}^\top \mathbf{x}}$ for some $\mathbf{x} \in \mathbb{R}^d$ and thus $\mathbb{P} = \delta_{\mathbf{x}}$. ■

Positive-part Shrinkage Estimator

Similar to James-Stein estimator, we can show that the positive-part version of $\hat{\boldsymbol{\mu}}_\alpha$ also outperforms $\hat{\boldsymbol{\mu}}$, where the positive-part estimator is defined by

$$\hat{\boldsymbol{\mu}}_\alpha^+ := \alpha f^* + (1-\alpha)_+ \hat{\boldsymbol{\mu}} \tag{3.15}$$

with $(a)_+ := a$ if $a > 0$ and zero otherwise. (3.15) can be rewritten as

$$\hat{\boldsymbol{\mu}}_\alpha^+ = \begin{cases} \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}, & 0 \leq \alpha \leq 1 \\ \alpha f^* & 1 < \alpha < 2. \end{cases} \tag{3.16}$$

Let $\Delta_\alpha^+ \triangleq \mathbb{E}\|\hat{\boldsymbol{\mu}}_\alpha^+ - \boldsymbol{\mu}\|_{\mathscr{H}}^2$ be the risk of the positive-part estimator. Then, the following result shows that $\Delta_\alpha^+ \leq \Delta_\alpha$, given that α satisfies (3.12).

Proposition 3.5. *For any α satisfying* (3.12)*, we have that $\Delta_\alpha^+ \leq \Delta_\alpha < \Delta$.*

Proof of Proposition 3.5. According to (3.16), we decompose the proof into two parts. First, if $0 \leq \alpha \leq 1$, $\hat{\boldsymbol{\mu}}_\alpha$ and $\hat{\boldsymbol{\mu}}_\alpha^+$ behave exactly the same. Thus, $\Delta_\alpha^+ = \Delta_\alpha$. On the other hand, when $1 < \alpha < 2$, the bias-variance decomposition of these estimators yields

$$\Delta_\alpha = \alpha^2 \|f^* - \boldsymbol{\mu}\|_{\mathscr{H}}^2 + (1-\alpha)^2 \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \quad \text{and} \quad \Delta_\alpha^+ = \alpha^2 \|f^* - \boldsymbol{\mu}\|_{\mathscr{H}}^2.$$

It is easy to see that $\Delta_\alpha^+ < \Delta_\alpha$ when $1 < \alpha < 2$. This concludes the proof. ■

Proposition 3.5 implies that, when estimating α, it is better to restrict the value of α to be smaller than 1, although it can be greater than 1, as suggested by Theorem 3.3. The reason is that if $0 \leq \alpha \leq 1$, the bias is an increasing function of α, whereas the variance is a decreasing function of α. On the other hand, if $\alpha > 1$, both bias and variance become increasing functions of α. We will see later in Section 3.4 that $\hat{\boldsymbol{\mu}}_\alpha$ and $\hat{\boldsymbol{\mu}}_\alpha^+$ can be obtained naturally as a solution to a regularized regression problem.

3.3.2 Consequences of Theorem 3.3

As mentioned before, while Theorem 3.3 is interesting from the perspective of showing that the shrinkage estimator, $\hat{\boldsymbol{\mu}}_\alpha$ performs better—in the mean squared sense—than the empirical estimator, it unfortunately relies on the fact that $\boldsymbol{\mu}_{\mathbb{P}}$ (*i.e.*, the object of interest) is known, which makes $\hat{\boldsymbol{\mu}}_\alpha$ uninteresting. Instead of knowing $\boldsymbol{\mu}_{\mathbb{P}}$, which requires the knowledge of $\mathbb{P}$, in this section, we show that a shrinkage estimator can be constructed that performs better than the empirical estimator, uniformly over a class of probability distributions. To this end, we introduce the notion of an oracle upper bound.

Let $\mathscr{P}$ be a class of probability distributions $\mathbb{P}$ defined on a measurable space $\mathcal{X}$. We define an oracle upper bound as

$$U_{k,\mathscr{P}} \triangleq \inf_{\mathbb{P}\in\mathscr{P}} \frac{2\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}}.$$

It follows immediately from Theorem 3.3 and the definition of $U_{k,\mathscr{P}}$ that if $U_{k,\mathscr{P}} \neq 0$, then for any $\alpha \in (0, U_{k,\mathscr{P}})$, $\Delta_\alpha - \Delta < 0$ holds "uniformly" for all $\mathbb{P} \in \mathscr{P}$.

Note that by virtue of Proposition 3.4, the class $\mathscr{P}$ cannot contain the Dirac measure $\delta_{\mathbf{x}}$ (for any $\mathbf{x} \in \mathbb{R}^d$) if the kernel k is translation invariant and characteristic on $\mathbb{R}^d$. Below we give concrete examples of $\mathscr{P}$ for which $U_{k,\mathscr{P}} \neq 0$ so that the above uniformity statement holds. In particular, we will show (see Theorem 3.6) that for $\mathcal{X} = \mathbb{R}^d$, if a non-trivial bound on the L^2-norm of the characteristic function of $\mathbb{P}$ is known, it is possible to construct shrinkage estimators that are better (in mean squared error) than the empirical average. In such a case, unlike in Theorem 3.3, α does not depend on the individual distribution $\mathbb{P}$, but only on an upper bound associated with a class $\mathscr{P}$.

Theorem 3.6. *Let $k(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x}-\mathbf{y}) \neq 0$, $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ with $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ and ψ is a positive definite function. For a given constant $A \in (0,1)$, let $A_\psi := \frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}$ and*

$$\mathscr{P}_{k,A} \triangleq \left\{\mathbb{P} \in M^1_+(\mathbb{R}^d) : \|\phi_{\mathbb{P}}\|_{L^2} \leq \sqrt{A_\psi}\right\},$$

where $\phi_{\mathbb{P}}$ denotes the characteristic function of $\mathbb{P}$. Then for all $\mathbb{P} \in \mathscr{P}_{k,A}$, $\Delta_\alpha < \Delta$ if

$$\alpha \in \left(0, \frac{2(1-A)}{1 + (n-1)A + \frac{n\|f^*\|^2_{\mathscr{H}}}{\psi(0)} + \frac{2n\sqrt{A}\|f^*\|_{\mathscr{H}}}{\sqrt{\psi(0)}}}\right].$$

Proof of Theorem 3.6. By Theorem 3.3, we have that

$$\Delta_\alpha < \Delta, \ \forall\ \alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}}\right). \tag{3.17}$$

Consider

$$\begin{aligned}
\frac{\Delta}{\Delta + \|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}} &= \frac{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}}) + n\|f^* - \boldsymbol{\mu}\|^2_{\mathscr{H}}} \\
&= \frac{1 - \frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})}}{1 + (n-1)\frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})} + \frac{n\|f^*\|^2}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})} - \frac{2n\langle f^*, \boldsymbol{\mu}\rangle_{\mathscr{H}}}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})}} \\
&\geq \frac{1 - \frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})}}{1 + (n-1)\frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})} + \frac{n\|f^*\|^2}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})} + \frac{2n\|f^*\|\sqrt{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})}}.
\end{aligned} \tag{3.18}$$

Note that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}}) &= \int\int \psi(\mathbf{x}-\mathbf{y})\,\mathrm{d}\mathbb{P}(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{y}) \\
&\overset{(*)}{=} \int |\varphi_{\mathbb{P}}(\boldsymbol{\omega})|^2 \psi^\wedge(\boldsymbol{\omega})\ \mathrm{d}\boldsymbol{\omega} \\
&\leq \sup_{\boldsymbol{\omega}\in\mathbb{R}^d} \psi^\wedge(\boldsymbol{\omega}) \|\varphi_{\mathbb{P}}\|^2_{L_2} \leq (2\pi)^{-d/2}\|\psi\|_{L_1}\|\varphi_{\mathbb{P}}\|^2_{L_2},
\end{aligned} \tag{3.19}$$

where $\psi^\wedge$ is the Fourier transform of ψ and $(*)$ follows by invoking Bochner's theorem, which states that ψ is the Fourier transform of a non-negative finite Borel measure with density $(2\pi)^{-d/2}\psi^\wedge$, *i.e.*, $\psi(\mathbf{x}) = (2\pi)^{-d/2}\int e^{-i\mathbf{x}^\top\boldsymbol{\omega}}\psi^\wedge(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega},\ \mathbf{x}\in\mathbb{R}^d$ (see (16) in the proof of Proposition 5 in Sriperumbudur et al. (2011a)). As $\mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x}) = \psi(0)$, we have that

$$\frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})} \leq \frac{A\|\varphi_{\mathbb{P}}\|_{L^2}^2}{A_\psi}$$

and therefore for any $\mathbb{P}\in\mathscr{P}_{k,A}$, $\frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})} \leq A$. Using this in (3.18) and combining it with (3.17) yields the result. ■

Remark 3.2. *We provide some discussion regarding Theorem 3.6.*

(i) Theorem 3.6 shows that for any $\mathbb{P}\in\mathscr{P}_{k,A}$, it is possible to construct a shrinkage estimator that dominates the empirical estimator, i.e.*, the shrinkage estimator has a strictly smaller risk than that of the empirical estimator.*

(ii) Suppose that $\mathbb{P}$ has a density, denoted by p, w.r.t. the Lebesgue measure and $\varphi_{\mathbb{P}}\in L^2(\mathbb{R}^d)$. By Plancherel's theorem, $p\in L^2(\mathbb{R}^d)$ as $\|p\|_{L_2} = \|\varphi_{\mathbb{P}}\|_{L_2}$, which means that $\mathscr{P}_{k,A}$ includes distributions with square integrable densities (note that in general not every p is square integrable). Since $\|\varphi_{\mathbb{P}}\|_{L_2}^2 \leq \|\varphi_{\mathbb{P}}\|_{L_1}$, it is easy to check that

$$\left\{\mathbb{P}\in M_+^1(\mathbb{R}^d) : \|\phi_{\mathbb{P}}\|_{L^1} \leq \frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}\right\} \subset \mathscr{P}_{k,A},$$

which means the bounded densities belong to $\mathscr{P}_{k,A}$ as $\varphi_{\mathbb{P}}\in L^1(\mathbb{R}^d)$ implies that $\mathbb{P}$ has a density, $p\in C_0(\mathbb{R}^d)$. Moreover, it is easy to check that larger the value of A, larger is the class $\mathscr{P}_{k,A}$ and smaller is the range of α for which $\Delta_\alpha < \Delta$ and vice-versa.

In the following, we present some concrete examples to elucidate Theorem 3.6.

Example 3.1 (Gaussian kernel and Gaussian distribution)**.** *Define*

$$\mathscr{N} \triangleq \left\{\mathbb{P}\in M_+^1(\mathbb{R}^d)\,\middle|\,\mathrm{d}\mathbb{P}(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}}e^{-\frac{\|\mathbf{x}-\boldsymbol{\theta}\|_2^2}{2\sigma^2}}\,\mathrm{d}\mathbf{x},\ \boldsymbol{\theta}\in\mathbb{R}^d,\,\sigma>0\right\},$$

where $\psi(\mathbf{x}) = e^{-\|\mathbf{x}\|_2^2/2\tau^2}$, $\mathbf{x}\in\mathbb{R}^d$ and $\tau>0$. For $\mathbb{P}\in\mathscr{N}$, it is easy to verify that

$$\varphi_{\mathbb{P}}(\boldsymbol{\omega}) = e^{\sqrt{-1}\boldsymbol{\theta}^\top\boldsymbol{\omega}-\frac{1}{2}\sigma^2\|\boldsymbol{\omega}\|_2^2},\ \boldsymbol{\omega}\in\mathbb{R}^d \text{ and } \|\varphi_{\mathbb{P}}\|_{L_2}^2 = \int e^{-\sigma^2\|\boldsymbol{\omega}\|_2^2}\,\mathrm{d}\boldsymbol{\omega} = (\pi/\sigma^2)^{d/2}.$$

Also, $\|\psi\|_{L_1} = (2\pi\tau^2)^{d/2}$. Therefore, for $\mathscr{P}_{k,A} \triangleq \{\mathbb{P}\in\mathscr{N} : \sigma^2 \geq \pi\tau^2/A^{2/d}\}$, assuming $f^* = 0$, we obtain the result in Theorem 3.6, i.e.*, the result in Theorem 3.6 holds for all Gaussian distributions that are smoother (having larger variance) than that of the kernel.*

Example 3.2 (Linear kernel)**.** *Suppose $f^* = 0$ and $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{y}$. Let $\boldsymbol{\vartheta}$ and $\boldsymbol{\Sigma}$ represent the mean vector and covariance matrix of a distribution $\mathbb{P}$ defined on $\mathbb{R}^d$. Then it is easy to check that $\frac{\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})} = \frac{\|\boldsymbol{\vartheta}\|_2^2}{\mathrm{trace}(\boldsymbol{\Sigma})+\|\boldsymbol{\vartheta}\|_2^2}$ and therefore for a given $A\in(0,1)$, define*

$$\mathscr{P}_{k,A} \triangleq \left\{\mathbb{P}\in M_+^1(\mathbb{R}^d)\,\middle|\,\frac{\|\boldsymbol{\vartheta}\|_2^2}{\mathrm{trace}(\boldsymbol{\Sigma})} \leq \frac{A}{1-A}\right\}.$$

From (3.17) *and* (3.18)*, it is clear that for any* $\mathbb{P} \in \mathscr{P}_{k,A}$*,* $\Delta_\alpha < \Delta$ *if* $\alpha \in \left(0, \frac{2(1-A)}{1+(n-1)A}\right]$*. Note that this choice of kernel yields the setting similar to classical James-Stein estimation. In James-Stein estimation,* $\mathbb{P} \in \mathscr{N}$ *(see Example 3.1 for the definition of* $\mathscr{N}$*) and* $\boldsymbol{\vartheta}$ *is estimated as* $(1-\tilde{\alpha})\hat{\boldsymbol{\vartheta}}$*—which improves upon* $\hat{\boldsymbol{\vartheta}}$*—where* $\tilde{\alpha}$ *depends on the sample* $(\mathbf{x}_i)_{i=1}^n$ *and* $\hat{\boldsymbol{\vartheta}}$ *is the sample mean. In our case, for all* $\mathbb{P} \in \mathscr{P}_{k,A} = \left\{\mathbb{P} \in \mathscr{N} \,:\, \|\vartheta\|_2 \leq \sigma\sqrt{\frac{dA}{1-A}}\right\}$*,* $\Delta_\alpha < \Delta$ *if* $\alpha \in \left(0, \frac{2(1-A)}{1+(n-1)A}\right]$*. In addition, in contrast to the James-stein estimator which improves upon the empirical estimator (*i.e.*, sample mean) for only* $d \geq 3$*, we note here that the proposed estimator improves for any* d *as long as* $\mathbb{P} \in \mathscr{P}_{k,A}$*. On the other hand, the proposed estimator requires some knowledge about the distribution (particularly a bound on* $\|\boldsymbol{\vartheta}\|_2$*), which the James-Stein estimator does not (see Section 3.3.5 for more details).*

Example 3.3 (Exponential family)**.** *If the probability distribution* $\mathbb{P}$ *belongs to an exponential family, its probability density function can be expressed in the form*

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\Big(B(\mathbf{x}) + \boldsymbol{\theta}^\top T(\mathbf{x}) - Z(\boldsymbol{\theta})\Big)$$

whose squared L_2 *norm is given by*

$$\|p_{\boldsymbol{\theta}}\|_{L_2}^2 = \int p(\mathbf{x})^2 \,\mathrm{d}\mathbf{x}.$$

The above integral does not in general have a closed-form solution, unless $B(\mathbf{x})$ *and* $T(\mathbf{x})$ *obey certain conditions. Specifically, if* $2B(\mathbf{x}) = B(\eta\mathbf{x})$ *for some* η *and* T *is linear in* $\mathbf{x}$ *(Jebara et al. 2004a), we have*

$$\begin{aligned}
\int p(\mathbf{x})^2 \,\mathrm{d}\mathbf{x} &= \int \exp\Big(2B(\mathbf{x}) + 2\boldsymbol{\theta}^\top T(\mathbf{x}) - 2Z(\boldsymbol{\theta})\Big)\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{\eta}\int \exp\left(B(\eta\mathbf{x}) + \frac{2}{\eta}\boldsymbol{\theta}^\top T(\eta\mathbf{x}) - 2Z(\boldsymbol{\theta})\right)\mathrm{d}(\eta\mathbf{x}) \\
&= \frac{1}{\eta}\exp\left(-2Z(\boldsymbol{\theta})\right)\int \exp\left(B(\eta\mathbf{x}) + \frac{2}{\eta}\boldsymbol{\theta}^\top T(\eta\mathbf{x})\right)\mathrm{d}(\eta\mathbf{x}) \\
&= \frac{1}{\eta}\exp\left(-2Z(\boldsymbol{\theta})\right)\exp\left(Z\left(\frac{2\boldsymbol{\theta}}{\eta}\right)\right).
\end{aligned}$$

Thus, we obtain

$$\|p_{\boldsymbol{\theta}}\|_{L_2}^2 = \frac{1}{\eta}\exp\left(Z\left(\frac{2\boldsymbol{\theta}}{\eta}\right) - 2Z(\boldsymbol{\theta})\right).$$

Consequently, Theorem 3.6 holds if

$$\mathscr{P}_{k,A} = \left\{p_{\boldsymbol{\theta}} \,\middle|\, Z\left(\frac{2\boldsymbol{\theta}}{\eta}\right) - 2Z(\boldsymbol{\theta}) \leq \log\left(\eta\sqrt{\frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}}\right)\right\}.$$

Example 3.3 allows one to write the condition in Theorem 3.6 in term of the log partition function $Z(\boldsymbol{\theta})$. It is well known that much of the structure of exponential models can be derived from the log partition function, see, *e.g.*, Wainwright and Jordan (2008). Many probabilistic models such as Gaussian MRFs, Gaussian mixture model, and Latent Dirichlet Allocation can be expressed in exponential family form.

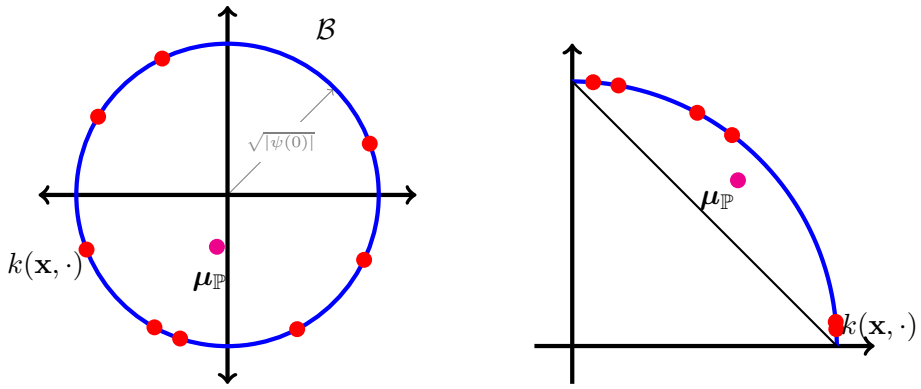


Figure 3.1: A 2D visualization of the ball of radius $\psi(0)$ in the RKHS. For stationary kernels, the feature map $\phi(\mathbf{x})$ always lie on this ball. As a result, all the kernel means $\mu_{\mathbb{P}}$ will lie inside the ball. Moreover, if $k(\mathbf{x},\mathbf{y}) > 0$ for all $\mathbf{x},\mathbf{y} \in \mathcal{X}$, all the feature maps $\phi(\mathbf{x})$ lie in the same quadrant. Thus, the kernel means $\mu_{\mathbb{P}}$ will always lie inside the ball segment.

3.3.3 Where to Shrink?

As discussed earlier, the choice of f^* may seem arbitrary, but in principle it should be chosen in such a way that it is close to the true $\mu_{\mathbb{P}}$. In general, without any assumption on the kernel, it follows from the strict convexity of $\|\cdot\|^2_{\mathscr{H}}$ and Jensen's inequality that

$$\mathbb{E}_{\mathbb{P}_n}\left[\|\hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}}\right] > \|\mathbb{E}_{\mathbb{P}_n}\left[\hat{\boldsymbol{\mu}}\right]\|^2_{\mathscr{H}} = \|\boldsymbol{\mu}\|^2_{\mathscr{H}},$$

which suggests that the true kernel mean μ will typically lie closer to the origin than the estimate $\hat{\boldsymbol{\mu}}$. One should therefore shrink the ordinary estimator toward the origin. Below I provide a geometrical illustration as to why the origin point may be reasonable for shrinkage in RKHS.

Stationary Kernels

An interesting property of this class of kernels is that for any $\mathbf{x} \in \mathcal{X}$, we have

$$k(\mathbf{x},\mathbf{x}) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle_{\mathscr{H}} = \psi(\mathbf{x}-\mathbf{x}) = \psi(0),$$

which implies that $\|\phi(\mathbf{x})\|^2_{\mathscr{H}} = \psi(0)$ for all $\mathbf{x} \in \mathcal{X}$. That is, all the feature maps $\phi(\mathbf{x})$ lie on the ball $\mathcal{B}$ of radius $\sqrt{|\psi(0)|}$ centered at the origin in the RKHS. Consequently, we have for any distribution $\mathbb{P}$ that

$$\|\mu_{\mathbb{P}}\|_{\mathscr{H}} = \left\| \int_{\mathcal{X}} \phi(\mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \right\|_{\mathscr{H}} \leq \int_{\mathcal{X}} \|\phi(\mathbf{x})\|_{\mathscr{H}} \, \mathrm{d}\mathbb{P}(\mathbf{x}) = \sqrt{|\psi(0)|}.$$

In other words, the kernel means $\mu_{\mathbb{P}}$ will always lie inside the ball $\mathcal{B}$. It will lie on the ball if and only if $\mathbb{P} = \delta_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{X}$. Furthermore, if $k(\mathbf{x},\mathbf{y}) > 0$ for all $\mathbf{x},\mathbf{y} \in \mathcal{X}$, all the feature maps $\phi(\mathbf{x})$ lie in the same quadrant. Thus, the kernel means $\boldsymbol{\mu}_{\mathbb{P}}$ will always lie inside the ball segment. These two cases are illustrated in Figure 3.1.

Non-stationary Kernels

Although it might be difficult in general to infer the choice of f^* from this class of kernels, there are particular families of non-stationary kernels that are closely related to the stationary kernels. For example, consider the *separable* non-stationary kernels $k(\mathbf{x},\mathbf{y}) = k_1(\mathbf{x})k_2(\mathbf{y})$ where k_1 and

k_2 are stationary kernels evaluated at the examples $\mathbf{x}$ and $\mathbf{y}$, respectively. Knowing the forms of functions k_1 and k_2 could facilitate in deriving the form of $\boldsymbol{\mu}_{\mathbb{P}}$. Another interesting class of non-stationary kernels is the class of *reducible* kernels. A kernel k is said to be reducible if there exists a bijective deformation ϕ such that $k(\mathbf{x}, \mathbf{y}) = \psi(\phi(\mathbf{x}) - \phi(\mathbf{y}))$ where ψ is a stationary kernel. As a result, the property of stationary kernel discussed earlier applies immediately to this class of kernel functions. For more detailed account on different classes of kernel functions, see *e.g.*, Genton (2002) and Rasmussen and Williams (2005; Chapter 4).

Motivated by the discussion above, I will focus on the case when $f^* = 0$ throughout this chapter and defer a more general class of f^* as an open problem for future works.

3.3.4 Data-Dependent Shrinkage Parameter

The discussion so far showed that the shrinkage estimator in (3.11) performs better than the empirical estimator if the data generating distribution satisfies a certain mild condition (see Theorem 3.6; Examples 3.1 and 3.2). However, since this condition is usually not checkable in practice, the shrinkage estimator lacks applicability. In this section, we present a completely data driven shrinkage estimator by estimating the shrinkage parameter α from data so that the estimator does not require any knowledge of the data generating distribution.

Since the maximal difference between Δ_α and Δ occurs at α_* (see the proof of Theorem 3.3), given an i.i.d. sample $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ from $\mathbb{P}$, we propose to estimate $\boldsymbol{\mu}$ using $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} = (1 - \tilde{\alpha})\hat{\boldsymbol{\mu}}$ (*i.e.*, assuming $f^* = 0$) where $\tilde{\alpha}$ is an estimator of $\alpha_* = \Delta/(\Delta + \|\boldsymbol{\mu}\|^2_{\mathscr{H}})$ given by

$$\tilde{\alpha} = \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}}}, \tag{3.20}$$

with $\hat{\Delta}$ and $\hat{\boldsymbol{\mu}}$ being the empirical versions of Δ and $\boldsymbol{\mu}$, respectively (see Theorem 3.7 for precise definitions). The following result shows that $\tilde{\alpha}$ is a $n\sqrt{n}$-consistent estimator of α_* and $\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}$ concentrates around $\|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}$. In addition, we show that

$$\Delta_{\alpha_*} \leq \Delta_{\tilde{\alpha}} \leq \Delta_{\alpha_*} + O(n^{-3/2}) \text{ as } n \to \infty,$$

which means the performance of $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ is similar to that of the best estimator (in mean squared sense) of the form $\hat{\boldsymbol{\mu}}_\alpha$. In what follows, we will call the estimator $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ an *empirical-bound kernel mean shrinkage estimator (B-KMSE)*.

Theorem 3.7. *Suppose $n \geq 2$ and $f^* = 0$. Let k be a continuous kernel on a separable topological space $\mathcal{X}$. Define*

$$\hat{\Delta} \triangleq \frac{\hat{\mathbb{E}}k(\mathbf{x}, \mathbf{x}) - \hat{\mathbb{E}}k(\mathbf{x}, \tilde{\mathbf{x}})}{n} \quad \textit{and} \quad \|\hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}} \triangleq \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j)$$

where $\hat{\mathbb{E}}k(\mathbf{x}, \mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i)$ and $\hat{\mathbb{E}}k(\mathbf{x}, \tilde{\mathbf{x}}) \triangleq \frac{1}{n(n-1)}\sum_{i \neq j}^n k(\mathbf{x}_i, \mathbf{x}_j)$ are the empirical estimators of $\mathbb{E}_{\mathbf{x}}k(\mathbf{x}, \mathbf{x})$ and $\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x}, \tilde{\mathbf{x}})$, respectively. Assume there exist finite constants $\kappa_1 > 0$, $\kappa_2 > 0$, $\sigma_1 > 0$ and $\sigma_2 > 0$ such that

$$\mathbb{E}\|k(\cdot, \mathbf{x}) - \boldsymbol{\mu}\|^m_{\mathscr{H}} \leq \frac{m!}{2}\sigma_1^2\kappa_1^{m-2}, \quad \forall\, m \geq 2. \tag{3.21}$$

and

$$\mathbb{E}|k(\mathbf{x}, \mathbf{x}) - \mathbb{E}_{\mathbf{x}}k(\mathbf{x}, \mathbf{x})|^m \leq \frac{m!}{2}\sigma_2^2\kappa_2^{m-2}, \quad \forall\, m \geq 2. \tag{3.22}$$

Then

$$|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2}) \text{ and } \left| \|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}} \right| = O_{\mathbb{P}}(n^{-3/2})$$

as $n \to \infty$. *In particular,*

$$\min_{\alpha} \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \leq \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \leq \min_{\alpha} \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 + O(n^{-3/2}) \tag{3.23}$$

as $n \to \infty$.

Before we prove Theorem 3.7, we present Bernstein's inequality in separable Hilbert spaces, quoted from Yurinsky (1995; Theorem 3.3.4), which will be used to prove Theorem 3.7.

Theorem 3.8 (Bernstein's inequality)**.** *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space,* H *be a separable Hilbert space,* $B > 0$ *and* $\theta > 0$. *Furthermore, let* $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n : \Omega \to H$ *be zero mean independent random variables satisfying*

$$\sum_{i=1}^{n} \mathbb{E}\|\boldsymbol{\xi}_i\|_H^m \leq \frac{m!}{2}\theta^2 B^{m-2}. \tag{3.24}$$

Then for any $\tau > 0$,

$$P^n \left\{ (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) : \left\| \sum_{i=1}^{n} \boldsymbol{\xi}_i \right\|_H \geq 2B\tau + \sqrt{2\theta^2\tau} \right\} \leq 2e^{-\tau}.$$

Proof of Theorem 3.7. Consider

$$\begin{aligned}
\tilde{\alpha} - \alpha_* &= \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2} - \frac{\Delta}{\Delta + \|\mu\|_{\mathscr{H}}^2} = \frac{\hat{\Delta}\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \Delta\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}{(\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2)(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2)} \\
&= \frac{\hat{\Delta}(\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2)}{(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2)(\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2)} + \frac{(\hat{\Delta} - \Delta)\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}{(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2)(\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2)} \\
&= \frac{\tilde{\alpha}(\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2)}{(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2)} + \frac{(\hat{\Delta} - \Delta)(1 - \tilde{\alpha})}{(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2)}.
\end{aligned}$$

Rearranging $\tilde{\alpha}$, we obtain

$$\tilde{\alpha} - \alpha_* = \frac{\alpha_*(\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2) + (1 - \alpha_*)(\hat{\Delta} - \Delta)}{\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}.$$

Therefore,

$$|\tilde{\alpha} - \alpha_*| \leq \frac{\alpha_*|\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2| + (1 + \alpha_*)|\hat{\Delta} - \Delta|}{(\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2) - (\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2) + (\hat{\Delta} - \Delta)}, \tag{3.25}$$

where it is easy to verify that

$$|\hat{\Delta} - \Delta| \leq \frac{|\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}}) - \hat{\mathbb{E}} k(\mathbf{x}, \tilde{\mathbf{x}})|}{n} + \frac{|\hat{\mathbb{E}} k(\mathbf{x}, \mathbf{x}) - \mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{x})|}{n}. \tag{3.26}$$

In the following we obtain bounds on $|\hat{\mathbb{E}} k(\mathbf{x}, \mathbf{x}) - \mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{x})|$, $|\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}}) - \hat{\mathbb{E}} k(\mathbf{x}, \tilde{\mathbf{x}})|$ and $|\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2|$ when the kernel satisfies (3.21) and (3.22).

Bound on $|\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})|$: Since k is a continuous kernel on a separable topological space $\mathcal{X}$, it follows from Lemma 4.33 of Steinwart and Christmann (2008) that $\mathscr{H}$ is separable. By defining $\xi_i \triangleq k(\mathbf{x}_i,\mathbf{x}_i) - \mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})$, it follows from (3.22) that $\theta = \sqrt{n}\sigma_2$ and $B = \kappa_2$ and so by Theorem 3.8, for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$|\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})| \leq \sqrt{\frac{2\sigma_2^2\tau}{n}} + \frac{2\kappa_2\tau}{n}. \tag{3.27}$$

Bound on $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}$: By defining $\xi_i \triangleq k(\cdot,\mathbf{x}_i) - \boldsymbol{\mu}$ and using (3.21), we have $\theta = \sqrt{n}\sigma_1$ and $B = \kappa_1$. Therefore, by Theorem 3.8, for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} \leq \sqrt{\frac{2\sigma_1^2\tau}{n}} + \frac{2\kappa_1\tau}{n}. \tag{3.28}$$

Bound on $|\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - \|\boldsymbol{\mu}\|_{\mathscr{H}}^2|$: Since

$$\left|\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - \|\boldsymbol{\mu}\|_{\mathscr{H}}^2\right| \leq (\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}} + \|\boldsymbol{\mu}\|_{\mathscr{H}})\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} \leq (\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + 2\|\boldsymbol{\mu}\|_{\mathscr{H}})\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}},$$

it follows from (3.28) that for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\left|\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - \|\boldsymbol{\mu}\|_{\mathscr{H}}^2\right| \leq D_1\sqrt{\frac{\tau}{n}} + D_2\left(\frac{\tau}{n}\right) + D_3\left(\frac{\tau}{n}\right)^{3/2} + D_4\left(\frac{\tau}{n}\right)^2, \tag{3.29}$$

where $(D_i)_{i=1}^4$ are positive constants that depend only on σ_1^2, κ and $\|\boldsymbol{\mu}\|_{\mathscr{H}}$, and not on n and τ.

Bound on $|\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})|$: Since

$$\begin{aligned}
&\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}}) \\
&= \frac{n^2(\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - \|\boldsymbol{\mu}\|_{\mathscr{H}}^2) + n(\mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x}) - \hat{\mathbb{E}}k(\mathbf{x},\mathbf{x})) + n(\|\boldsymbol{\mu}\|_{\mathscr{H}}^2 - \mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x}))}{n(n-1)},
\end{aligned} \tag{3.30}$$

it follows from (3.27) and (3.29) that for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$\begin{aligned}
|\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})| \quad &\leq \quad F_1\sqrt{\frac{\tau}{n}} + F_2\left(\frac{\tau}{n}\right) + F_3\left(\frac{\tau}{n}\right)^{3/2} + F_4\left(\frac{\tau}{n}\right)^2 + \frac{F_5}{n} \\
&\leq \quad F_1'\sqrt{\frac{1+\tau}{n}} + F_2'\left(\frac{1+\tau}{n}\right) + F_3'\left(\frac{1+\tau}{n}\right)^{3/2} \\
&\quad + F_4'\left(\frac{1+\tau}{n}\right)^2,
\end{aligned} \tag{3.31}$$

where $(F_i)_{i=1}^5$ and $(F_i')_{i=1}^4$ are positive constants that do not depend on n and τ.

Bound on $|\tilde{\alpha} - \alpha_*|$: Using (3.27) and (3.31) in (3.26), for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$|\hat{\Delta} - \Delta| \leq \frac{F_1''}{n}\sqrt{\frac{1+\tau}{n}} + \frac{F_2''}{n}\left(\frac{1+\tau}{n}\right) + \frac{F_3''}{n}\left(\frac{1+\tau}{n}\right)^{3/2} + \frac{F_4''}{n}\left(\frac{1+\tau}{n}\right)^2,$$

using which in (3.25) along with (3.29), we obtain that for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$|\tilde{\alpha} - \alpha_*| \leq \frac{\sum_{i=1}^4 \left(G_{i1}\alpha_* + \frac{G_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \tag{3.32}$$

where $\theta_n \triangleq \Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2$ and $(G_{i1})_{i=1}^4$, $(G_{i2})_{i=1}^4$ are positive constants that do not depend on n and τ. Since $\alpha_* = \frac{\Delta}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} = \frac{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) + (n-1)\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})} = O(n^{-1})$ and $\theta_n = \frac{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) + (n-1)\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{n} = O(1)$ as $n \to \infty$, it follows from (3.32) that $|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

Bound on $|\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}|$: Using (3.28) and (3.32) in

$$|\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| \leq \|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \hat{\boldsymbol{\mu}}_{\alpha_*}\|_{\mathscr{H}} \leq |\tilde{\alpha} - \alpha_*| \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + |\tilde{\alpha} - \alpha_*| \|\boldsymbol{\mu}\|_{\mathscr{H}},$$

for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$, we have

$$|\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| \quad \leq \quad \frac{\sum_{i=1}^6 \left(G'_{i1}\alpha_* + \frac{G'_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \tag{3.33}$$

where $(G'_{i1})_{i=1}^6$ and $(G'_{i2})_{i=1}^6$ are positive constants that do not depend on n and τ. From (3.33), it is easy to see that $|\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

Bound on $\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2$: Since

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 &\leq \left(\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} + \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}\right) |\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| \\
&\leq 2\left(\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}} + \|\boldsymbol{\mu}\|_{\mathscr{H}}\right) |\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| \\
&\leq 2\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + 2\|\boldsymbol{\mu}\|_{\mathscr{H}}\right) |\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}|,
\end{aligned}$$

for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 &\leq \frac{\sum_{i=1}^8 \left(G''_{i1}\alpha_* + \frac{G''_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \\
&\leq \frac{\sum_{i=1}^8 \left(G''_{i1}\alpha_* + \frac{G''_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1}{n}\right)^{i/2}\right|}, \\
&\leq \begin{cases} \frac{\gamma_n}{\phi_n}\sqrt{\frac{1+\tau}{n}}, & 0 < \tau \leq n-1 \\ \frac{\gamma_n}{\phi_n}\left(\frac{1+\tau}{n}\right)^4, & \tau \geq n-1 \end{cases},
\end{aligned}$$

where $\gamma_n \triangleq H_1\alpha_* + \frac{H_2}{n}(1+\alpha_*)$, $\phi_n \triangleq \left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1}{n}\right)^{i/2}\right|$ and $(H_i)_{i=1}^2$ are positive constants that do not depend on n and τ. In other words,

$$\mathbb{P}\left(\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 > \epsilon\right) \leq \begin{cases} 4\exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^2\right), & \frac{\gamma_n}{\phi_n\sqrt{n}} \leq \epsilon \leq \frac{\gamma_n}{\phi_n} \\ 4\exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^{1/4}\right), & \epsilon \geq \frac{\gamma_n}{\phi_n} \end{cases}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 &= \int_0^\infty \mathbb{P}\left(\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 > \epsilon\right) \mathrm{d}\epsilon \\
&\leq \frac{\gamma_n}{\phi_n\sqrt{n}} + 4\int_{\frac{\gamma_n}{\phi_n\sqrt{n}}}^{\frac{\gamma_n}{\phi_n}} \exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^2\right) \mathrm{d}\epsilon \\
&\qquad + 4\int_{\frac{\gamma_n}{\phi_n}}^{\infty} \exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^{1/4}\right) \mathrm{d}\epsilon \\
&= \frac{\gamma_n}{\phi_n\sqrt{n}} + \frac{2\gamma_n}{\phi_n\sqrt{n}}\int_0^{n-1} \frac{e^{-t}}{\sqrt{t+1}}\,\mathrm{d}t \\
&\quad + \frac{16e\gamma_n}{n^4\phi_n}\int_n^\infty t^3 e^{-t}\,\mathrm{d}t.
\end{aligned}$$

Since $\int_0^{n-1} \frac{e^{-t}}{\sqrt{t+1}}\,\mathrm{d}t \leq \int_0^\infty e^{-t}\,\mathrm{d}t = 1$ and $\int_n^\infty t^3 e^{-t}\,\mathrm{d}t \leq \int_0^\infty t^3 e^{-t}\,\mathrm{d}t = 6$, we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 - \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \leq \frac{3\gamma_n}{\phi_n\sqrt{n}} + \frac{96e\gamma_n}{n^4\phi_n}.$$

The claim in (3.23) follows by noting that $\gamma_n = O(n^{-1})$ and $\phi_n = O(1)$ as $n \to \infty$. ■

Remark 3.3. *Based on Theorem 3.7, we make the following observations.*

(i) $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\mu}$*. This follows from*

$$\begin{aligned}
\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} &\leq \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}} + O_{\mathbb{P}}(n^{-3/2}) \\
&\leq (1-\alpha_*)\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + \alpha_*\|\boldsymbol{\mu}\|_{\mathscr{H}} + O_{\mathbb{P}}(n^{-3/2})
\end{aligned}$$

with $\alpha_* = \frac{\Delta}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} = \frac{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})}{\mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) + (n-1)\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})} = O(n^{-1})$ *as* $n \to \infty$*. Using (3.28), we obtain* $\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}} = O_{\mathbb{P}}(n^{-1/2})$ *as* $n \to \infty$*, which implies that* $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\mu}$*.*

(ii) (3.23) shows that $\Delta_{\tilde{\alpha}} \leq \Delta_{\alpha_*} + O(n^{-3/2})$ *where* $\Delta_{\alpha_*} < \Delta$ *(see Theorem 3.3) and therefore for any* $\mathbb{P}$ *satisfying (3.21) and (3.22),* $\Delta_{\tilde{\alpha}} < \Delta + O(n^{-3/2})$ *as* $n \to \infty$*.*

(iii) Suppose the kernel is bounded, i.e.*,* $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} |k(\mathbf{x},\mathbf{y})| \leq \kappa < \infty$*. Then it is easy to verify that (3.21) and (3.22) hold with* $\sigma_1 = \sqrt{\kappa}$*,* $\kappa_1 = 2\sqrt{\kappa}$*,* $\sigma_2 = \kappa$ *and* $\kappa_2 = 2\kappa$ *and therefore the claims in Theorem 3.7 hold for bounded kernels.*

(iv) For $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{y}$*, we have*

$$\mathbb{E}\|k(\cdot,\mathbf{x}) - \boldsymbol{\mu}\|_{\mathscr{H}}^m = \mathbb{E}\left(\|k(\cdot,\mathbf{x}) - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right)^{m/2} = \mathbb{E}\left(\|\mathbf{x} - \mathbb{E}_{\mathbf{x}}\mathbf{x}\|_2^2\right)^{m/2} = \mathbb{E}\|\mathbf{x} - \mathbb{E}_{\mathbf{x}}\mathbf{x}\|_2^m$$

and

$$\mathbb{E}|k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x}} k(\mathbf{x},\mathbf{x})|^m = \mathbb{E}|\|\mathbf{x}\|_2^2 - \mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|_2^2|^m.$$

The conditions in (3.21) and (3.22) hold for $\mathbb{P} \in \mathscr{N}$ *where* $\mathscr{N}$ *is defined in Example 3.1. With* $\mathbb{P} \in \mathscr{N}$ *and* $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{y}$*, the problem of estimating* $\boldsymbol{\mu}$ *reduces to estimating* $\boldsymbol{\theta}$*, for which we have presented a James-Stein-like estimator,* $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ *that satisfies the oracle inequality in (3.23).*

(v) While the moment conditions in (3.21) and (3.22) are obviously satisfied by bounded kernels, for unbounded kernels, these conditions are quite stringent as they require all the higher moments to exist. These conditions can be weakened and the proof of Theorem 3.7 can be carried out using Chebyshev inequality instead of Bernstein's inequality but at the cost of a slow rate in (3.23).

3.3.5 Connection to James-Stein Estimator

In this section, we explore the connection of our proposed estimator in (3.11) to the James-Stein estimator. Recall that Stein's setting deals with estimating the mean of the Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d)$, which can be viewed as a special case of kernel mean estimation when we restrict to the class of distributions $\mathscr{P} \triangleq \{\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d) \,|\, \boldsymbol{\theta} \in \mathbb{R}^d\}$ and a linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ (see Example 3.2) assuming $f^* = 0$. In this case, it is easy to verify that $\Delta = d\sigma^2/n$ and $\Delta_\alpha < \Delta$ for

$$\alpha \in \left(0, \frac{2d\sigma^2}{d\sigma^2 + n\|\boldsymbol{\theta}\|^2}\right).$$

Let us assume that $n = 1$, in which case, we obtain $\Delta_\alpha < \Delta$ for $\alpha \in \left(0, \frac{2d\sigma^2}{\mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|^2}\right)$ as $\mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|^2 = \|\boldsymbol{\theta}\|^2 + d\sigma^2$. Note that the choice of α is dependent on $\mathbb{P}$ through $\mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|^2$ which is not known in practice. To this end, we replace it with the empirical version $\|\mathbf{x}\|^2$ that depends only on the sample $\mathbf{x}$. For an arbitrary constant $c \in (0, 2d)$, the shrinkage estimator (assuming $f^* = 0$) can thus be written as

$$\hat{\boldsymbol{\mu}}_\alpha = (1 - \alpha)\hat{\boldsymbol{\mu}} = \left(1 - \frac{c\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x} = \mathbf{x} - \frac{c\sigma^2 \mathbf{x}}{\|\mathbf{x}\|^2}$$

which is exactly the James-Stein estimator. This particular way of estimating the shrinkage parameter α has an intriguing consequence, as shown in Stein's seminal works (Stein 1955, James and Stein 1961), that the shrinkage estimator $\hat{\boldsymbol{\mu}}_\alpha$ can be shown to dominate the maximum likelihood estimator $\hat{\boldsymbol{\mu}}$ *uniformly* over all $\boldsymbol{\theta}$.

While it is compelling to see that there is seemingly a fundamental principle underlying both these settings, this connection also reveals crucial difference between our approach and classical setting of Stein—notably, original James-Stein estimator improves upon the sample mean even when the empirical norm of x is in the denominator (see $\hat{\boldsymbol{\mu}}_\alpha$ above).

3.4 Regression Perspective

In Section 3.3, I have shown that James-Stein-like shrinkage estimator (see (3.11)) improves upon the empirical estimator in estimating the kernel mean. In this section, I provide a regression perspective to shrinkage estimation. The starting point of the connection between regression and shrinkage estimation is the observation that the kernel mean $\boldsymbol{\mu}_{\mathbb{P}}$ and its empirical estimate $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ can be obtained as minimizers of the following risk functionals,

$$\mathcal{E}(g) := \int_{\mathcal{X}} \|k(\cdot, \mathbf{x}) - g\|^2_{\mathscr{H}} \, \mathrm{d}\mathbb{P}(\mathbf{x}) \qquad \text{and} \qquad \widehat{\mathcal{E}}(g) := \frac{1}{n}\sum_{i=1}^{n} \|k(\cdot, \mathbf{x}_i) - g\|^2_{\mathscr{H}},$$

respectively (Kim and Scott 2012). Given these formulations, it is natural to ask if minimizing the regularized version of $\widehat{\mathcal{E}}(g)$ will give a "better" estimator. While this question is interesting, it has to be noted that in principle, there is really no need to consider a regularized formulation as the problem of minimizing $\widehat{\mathcal{E}}$ is not ill-posed, unlike in function estimation or regression problems. To investigate this question, we will consider the minimization of the following regularized empirical risk functional,

$$\widehat{\mathcal{E}}_\lambda(g) \triangleq \widehat{\mathcal{E}}(g) + \lambda\Omega(\|g\|_{\mathscr{H}}) = \frac{1}{n}\sum_{i=1}^{n} \|k(\cdot, \mathbf{x}_i) - g\|^2_{\mathscr{H}} + \lambda\Omega(\|g\|_{\mathscr{H}}), \tag{3.34}$$

where $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ denotes a monotonically increasing function and $\lambda > 0$ is the regularization parameter. By representer theorem (Schölkopf et al. 2001a), any function $g \in \mathscr{H}$ that is a minimizer of (3.34) lies in a subspace spanned by $\{k(\cdot, \mathbf{x}_1), \ldots, k(\cdot, \mathbf{x}_n)\}$, *i.e.*, $g = \sum_{j=1}^n \beta_j k(\cdot, \mathbf{x}_j)$ for some $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_n]^\top \in \mathbb{R}^n$. Hence, by setting $\Omega(\|g\|_{\mathscr{H}}) = \|g\|^2_{\mathscr{H}}$, we can rewrite (3.34) in terms of $\boldsymbol{\beta}$ as

$$\widehat{\mathcal{E}}(g) + \lambda\Omega(\|g\|_{\mathscr{H}}) = \boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{K}\mathbf{1}_n + \lambda\boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} + c, \tag{3.35}$$

where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, c is a constant that does not depend on $\boldsymbol{\beta}$, and $\mathbf{1}_n = [1/n, 1/n, \ldots, 1/n]^\top$. Differentiating (3.35) w.r.t. $\boldsymbol{\beta}$ and setting it to zero yields an optimal weight vector $\boldsymbol{\beta} = \left(\frac{1}{1+\lambda}\right)\mathbf{1}_n$ and so the minimizer of (3.34) is given by

$$\hat{\boldsymbol{\mu}}_\lambda = \frac{1}{1+\lambda}\hat{\boldsymbol{\mu}} = \left(1 - \frac{\lambda}{1+\lambda}\right)\hat{\boldsymbol{\mu}} \triangleq (1-\alpha)\hat{\boldsymbol{\mu}}, \tag{3.36}$$

which is nothing but the shrinkage estimator in (3.11) with $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$. This provides a nice relation between shrinkage estimation and regularized risk minimization, wherein the regularization helps in shrinking the estimator $\hat{\boldsymbol{\mu}}$ towards zero although it is not required from the point of view of ill-posedness. In particular, since $0 < 1 - \alpha < 1$, $\hat{\boldsymbol{\mu}}_\lambda$ corresponds to a *positive-part* estimator proposed in Section 3.3.1 when $f^* = 0$.

Note that $\hat{\boldsymbol{\mu}}_\lambda$ is a consistent estimator of $\boldsymbol{\mu}$ as $\lambda \to 0$ and $n \to \infty$, which follows from

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} \leq \frac{1}{1+\lambda}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + \frac{\lambda}{1+\lambda}\|\boldsymbol{\mu}\|_{\mathscr{H}} \leq O_{\mathbb{P}}(n^{-1/2}) + O(\lambda).$$

In particular $\lambda = \tau n^{-1/2}$ (for some constant $\tau > 0$) yields the slowest possible rate for $\lambda \to 0$ such that the best possible rate of $n^{-1/2}$ is obtained for $\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} \to 0$ as $n \to \infty$. In addition, following the idea in Theorem 3.6, it is easy to show that $\mathbb{E}\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|^2_{\mathscr{H}} < \Delta$ if $\tau \in \left(0, \frac{2\sqrt{n}\Delta}{\|\boldsymbol{\mu}\|^2_{\mathscr{H}} - \Delta}\right)$. Note that $\hat{\boldsymbol{\mu}}_\lambda$ is not useful in practice as λ is not known *a priori*. However, by choosing

$$\lambda = \frac{\hat{\Delta}}{\|\hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}}},$$

it is easy to verify (see Theorem 3.7 and Remark 3.3) that

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|^2_{\mathscr{H}} < \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2_{\mathscr{H}} + O(n^{-3/2}) \tag{3.37}$$

as $n \to \infty$. Owing to the connection of $\hat{\boldsymbol{\mu}}_\lambda$ to a regression problem, in the following, we present an alternate data-dependent choice of λ obtained from leave-one-out cross validation (LOOCV) that also satisfies (3.37), and we refer to the corresponding estimator as *regularized kernel mean shrinkage estimator (R-KMSE)*.

To this end, for a given shrinkage parameter λ, denote by $\hat{\boldsymbol{\mu}}_\lambda^{(-i)}$ as the kernel mean estimated from $\{\mathbf{x}_j\}_{j=1}^n \backslash \{\mathbf{x}_i\}$. We will measure the quality of $\hat{\boldsymbol{\mu}}_\lambda^{(-i)}$ by how well it approximates $k(\cdot, \mathbf{x}_i)$ with the overall quality being quantified by the cross-validation score,

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^n \left\| k(\cdot, \mathbf{x}_i) - \hat{\boldsymbol{\mu}}_\lambda^{(-i)} \right\|^2_{\mathscr{H}}. \tag{3.38}$$

The LOOCV formulation in (3.38) differs from the one used in regression, wherein instead of measuring the deviation of the prediction made by the function on the omitted observation, we measure the deviation between the feature map of the omitted observation and the function itself. The following result shows that the shrinkage parameter in $\hat{\boldsymbol{\mu}}_\lambda$ (see (3.36)) can be obtained analytically by minimizing (3.38) and requires $O(n^2)$ operations to compute.

Proposition 3.9. *Let $n \geq 2$, $\rho := \frac{1}{n^2}\sum_{i,j=1}^n k(\mathbf{x}_i, \mathbf{x}_j)$ and $\varrho := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i)$. Assuming $n\rho > \varrho$, the unique minimizer of $LOOCV(\lambda)$ is given by*

$$\lambda_r = \frac{n(\varrho - \rho)}{(n-1)(n\rho - \varrho)}. \tag{3.39}$$

Proof of Proposition 3.9. Define $\alpha \triangleq \frac{\lambda}{\lambda+1}$ and $\phi(\mathbf{x}_i) \triangleq k(\cdot, \mathbf{x}_i)$. Note that

$$\begin{aligned}
LOOCV(\lambda) &\triangleq \frac{1}{n}\sum_{i=1}^n \left\| \frac{(1-\alpha)}{n-1}\sum_{j\neq i}\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \right\|_{\mathscr{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left\| \frac{n(1-\alpha)}{n-1}\hat{\boldsymbol{\mu}} - \frac{1-\alpha}{n-1}\phi(\mathbf{x}_i) - \phi(\mathbf{x}_i) \right\|_{\mathscr{H}}^2 \\
&= \left\| \frac{n(1-\alpha)}{n-1}\hat{\boldsymbol{\mu}} \right\|_{\mathscr{H}}^2 - \frac{2}{n}\left\langle \sum_{i=1}^n \frac{n-\alpha}{n-1}\phi(\mathbf{x}_i), \frac{n(1-\alpha)}{n-1}\hat{\boldsymbol{\mu}} \right\rangle_{\mathscr{H}} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left\| \frac{n-\alpha}{n-1}\phi(\mathbf{x}_i) \right\|_{\mathscr{H}}^2 \\
&= \left(\frac{n^2(1-\alpha)^2}{(n-1)^2} - \frac{2n(n-\alpha)(1-\alpha)}{(n-1)^2} \right) \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 \\
&\quad + \frac{(n-\alpha)^2}{n(n-1)^2}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) \\
&= \frac{1}{(n-1)^2}\left\{ \alpha^2(n^2\rho - 2n\rho + \varrho) + 2n\alpha(\rho - \varrho) + n^2(\varrho - \rho) \right\} \\
&=: \frac{F(\alpha)}{(n-1)^2}.
\end{aligned}$$

Since $\frac{\mathrm{d}}{\mathrm{d}\lambda}LOOCV(\lambda) = (n-1)^{-2}\frac{\mathrm{d}}{\mathrm{d}\alpha}F(\alpha)\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda} = (n-1)^{-2}(1+\lambda)^{-2}\frac{d}{d\alpha}F(\alpha)$, equating it zero yields (3.39). It is easy to show that the second derivative of $LOOCV(\lambda)$ is positive implying that $LOOCV(\lambda)$ is strictly convex and so λ_r is unique. ■

It is instructive to compare

$$\alpha_r = \frac{\lambda_r}{\lambda_r + 1} = \frac{\varrho - \rho}{(n-2)\rho + \varrho/n} \tag{3.40}$$

to the one in (3.20), where the latter can be shown to be $\frac{\varrho-\rho}{\varrho+(n-2)\rho}$, by noting that $\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) = \varrho$ and $\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) = \frac{n\rho-\varrho}{n-1}$ (in Theorem 3.7, we employ the U-statistic estimator of $\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})$, whereas ρ in Proposition 3.9 can be seen as a V-statistic counterpart). This means α_r obtained from LOOCV will be relatively larger than the one obtained from (3.20). Like in Theorem 3.7, the requirement that $n \geq 2$ in Theorem 3.9 stems from the fact that at least two data points are needed to evaluate the LOOCV score. Note that $n\rho > \varrho$ if and only if $\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) > 0$, which is guaranteed if the kernel is positive valued. We refer to $\hat{\boldsymbol{\mu}}_{\lambda_r}$ as R-KMSE, whose $\sqrt{n}$-consistency is established by the following result, which also shows that $\hat{\boldsymbol{\mu}}_{\lambda_r}$ satisfies (3.37).

Theorem 3.10. *Let $n \geq 2$, $n\rho > \varrho$ where ρ and ϱ are defined in Proposition 3.9 and k satisfies the assumptions in Theorem 3.7. Then $\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}} = O_{\mathbb{P}}(n^{-1/2})$,*

$$\min_{\alpha} \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \leq \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \leq \min_{\alpha} \mathbb{E}\|\hat{\boldsymbol{\mu}}_{\alpha} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 + O(n^{-3/2}) \tag{3.41}$$

where $\hat{\boldsymbol{\mu}}_\alpha = (1-\alpha)\hat{\boldsymbol{\mu}}$ *and therefore*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 < \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 + O(n^{-3/2}) \tag{3.42}$$

as $n \to \infty$.

Proof of Theorem 3.10. Since $\hat{\boldsymbol{\mu}}_{\lambda_r} = \frac{\hat{\boldsymbol{\mu}}}{1+\lambda_r} = (1-\alpha_r)\hat{\boldsymbol{\mu}}$, we have $\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}} \leq \alpha_r\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}} + \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}$. Note that

$$\alpha_r = \frac{n(\varrho-\rho)}{n(n-2)\rho+\varrho} = \frac{n\hat{\Delta}}{\hat{\Delta}+(n-1)\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2} = \frac{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})}{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) + (n-2)\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})},$$

where $\hat{\Delta}$, $\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2$, $\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x})$ and $\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})$ are defined in Theorem 3.7. Consider $|\alpha_r - \alpha_*| \leq |\alpha_r - \tilde{\alpha}| + |\tilde{\alpha} - \alpha_*|$ where $\tilde{\alpha}$ is defined in (3.20). From Theorem 3.7, we have $|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$ and

$$\begin{aligned}
\alpha_r - \tilde{\alpha} &= \frac{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})}{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) + (n-2)\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})} - \frac{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})}{2\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) + (n-2)\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})} \\
&= \frac{\tilde{\alpha}\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x})}{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) + (n-2)\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})} = (\tilde{\alpha} - \alpha_*)\beta + \alpha_*\beta,
\end{aligned}$$

where $\beta \triangleq \frac{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x})}{\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x})+(n-2)\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}})}$. Therefore, $|\alpha_r - \tilde{\alpha}| \leq |\tilde{\alpha} - \alpha_*||\beta| + \alpha_*|\beta|$, where $\alpha_* = O(n^{-1})$ as $n \to \infty$, which follows from Remark 3.3(i). Since $|\hat{\mathbb{E}}k(\mathbf{x},\mathbf{x}) - \mathbb{E}_{\mathbf{x}}k(\mathbf{x},\mathbf{x})| = O_{\mathbb{P}}(n^{-1/2})$ and $|\hat{\mathbb{E}}k(\mathbf{x},\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}k(\mathbf{x},\tilde{\mathbf{x}})| = O_{\mathbb{P}}(n^{-1/2})$, which follow from (3.27) and (3.31) respectively, we have $|\beta| = O_{\mathbb{P}}(n^{-1})$ as $n \to \infty$. Combining the above, we have $|\alpha_r - \tilde{\alpha}| = O_{\mathbb{P}}(n^{-2})$, thereby yielding $|\alpha_r - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$. Proceeding as in Theorem 3.7, we have

$$|\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| \leq \|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}_{\alpha_*}\|_{\mathscr{H}} \leq |\alpha_r - \alpha_*|\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + |\alpha_r - \alpha_*|\|\boldsymbol{\mu}\|_{\mathscr{H}},$$

which from the above follows that $|\|\hat{\boldsymbol{\mu}}_{\lambda_r} - \boldsymbol{\mu}\|_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\alpha_*} - \boldsymbol{\mu}\|_{\mathscr{H}}| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$. By arguing as in Remark 3.3(i), it is easy to show that $\hat{\boldsymbol{\mu}}_{\lambda_r}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\mu}$. (3.41) follows by carrying out the analysis as in the proof of Theorem 3.7 verbatim by replacing $\tilde{\alpha}$ with α_r, while (3.42) follows by appealing to Remark 3.3(ii). ■

3.4.1 Shrinkage via Spectral Filtering

Consider the following regularized risk minimization problem

$$\arg\inf_{\mathbf{F}\in\mathscr{H}\otimes\mathscr{H}} \quad \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\,\|k(\mathbf{x},\cdot) - \mathbf{F}[k(\mathbf{x},\cdot)]\|_{\mathscr{H}}^2 + \lambda\|\mathbf{F}\|_{\text{HS}}^2, \tag{3.43}$$

where the minimization is carried over the space of Hilbert-Schmidt operators, $\mathbf{F}$ on $\mathscr{H}$ with $\|\mathbf{F}\|_{\text{HS}}$ being the Hilbert-Schmidt norm of $\mathbf{F}$. As an interpretation, we are finding a smooth operator $\mathbf{F}$ that maps $k(\mathbf{x},\cdot)$ to itself (see Grünewälder et al. (2013) for more details on this smooth operator framework). It is not difficult to show that the solution to (3.43) is given $\mathbf{F} = \mathbf{C}_{XX}(\mathbf{C}_{XX} + \lambda I)^{-1}$ where $\mathbf{C}_{XX} = \int k(\cdot,\mathbf{x}) \otimes k(\cdot,\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{x})$ is a covariance operator defined on $\mathscr{H}$ (see, *e.g.*, Grünewälder et al. (2012)). Consequently, let us define

$$\boldsymbol{\mu}_\lambda = \mathbf{F}\boldsymbol{\mu} = \mathbf{C}_{XX}(\mathbf{C}_{XX} + \lambda I)^{-1}\boldsymbol{\mu},$$

which is an approximation to $\boldsymbol{\mu}$ as it can be shown that $\|\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} \to 0$ as $\lambda \to 0$ (see the proof of Theorem 3.13). Given an i.i.d. sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbb{P}$, the empirical counterpart of (3.43) is given by

$$\arg\min_{\mathbf{F}\in\mathscr{H}\otimes\mathscr{H}} \quad \frac{1}{n}\sum_{i=1}^n \|k(\mathbf{x}_i,\cdot) - \mathbf{F}[k(\mathbf{x}_i,\cdot)]\|_{\mathscr{H}}^2 + \lambda\|\mathbf{F}\|_{\text{HS}}^2 \tag{3.44}$$

resulting in

$$\check{\boldsymbol{\mu}}_\lambda := \mathbf{F}\hat{\boldsymbol{\mu}} = \widehat{\mathbf{C}}_{XX}(\widehat{\mathbf{C}}_{XX} + \lambda\mathbf{I})^{-1}\hat{\boldsymbol{\mu}} \tag{3.45}$$

where $\widehat{\mathbf{C}}_{XX}$ is the empirical covariance operator on $\mathscr{H}$ given by

$$\widehat{\mathbf{C}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, \mathbf{x}_i) \otimes k(\cdot, \mathbf{x}_i).$$

Unlike $\hat{\boldsymbol{\mu}}_\lambda$ in (3.36), $\check{\boldsymbol{\mu}}_\lambda$ shrinks $\hat{\boldsymbol{\mu}}$ differently in each coordinate by taking the eigenspectrum of $\widehat{\mathbf{C}}_{XX}$ into account (see Proposition 3.11) and so we refer to it as the *spectral kernel mean shrinkage estimator (S-KMSE)*.

Proposition 3.11. *Let $\{(\gamma_i, \phi_i)\}_{i=1}^n$ be eigenvalue and eigenfunction pairs of $\widehat{\mathbf{C}}_{XX}$. Then*

$$\check{\boldsymbol{\mu}}_\lambda = \sum_{i=1}^{n} \frac{\gamma_i}{\gamma_i + \lambda}\langle\hat{\boldsymbol{\mu}}, \phi_i\rangle_{\mathscr{H}}\phi_i.$$

Proof of Proposition 3.11. Since $\widehat{\mathbf{C}}_{XX}$ is a compact self-adjoint operator on $\mathscr{H}$, by Hilbert-Schmidt theorem (Reed and Simon 1972; Theorems VI.16, VI.17), we have that $\widehat{\mathbf{C}}_{XX} = \sum_{i=1}^{n}\gamma_i\langle\phi_i, \cdot\rangle_{\mathscr{H}}\phi_i$. The result follows by using this in (3.45). ■

As shown in Proposition 3.11, the effect of S-KMSE is to reduce the contribution of high frequency components of $\hat{\boldsymbol{\mu}}$ (*i.e.*, contribution of $\hat{\boldsymbol{\mu}}$ along the directions corresponding to smaller γ_i) when $\hat{\boldsymbol{\mu}}$ is expanded in terms of the eigenfunctions of the empirical covariance operator, which are nothing but the kernel PCA basis (see Rasmussen and Williams (2005; Section 4.3)). This means, similar to R-KMSE, S-KMSE also shrinks $\hat{\boldsymbol{\mu}}$ towards zero, however, the difference being that while R-KMSE shrinks equally in all coordinates, S-KMSE controls the amount of shrinkage by the information contained in each coordinate. In particular, S-KMSE takes into account more information about the kernel by allowing for different amount of shrinkage in each coordinate direction according to the value of γ_i, wherein the shrinkage is small in the coordinates whose γ_i are large. Moreover, Proposition 3.11 reveals that the effect of shrinkage is akin to *spectral filtering* (Bauer et al. 2007)—which in our case corresponds to Tikhonov regularization—wherein S-KMSE filters out the high-frequency components of the spectral representation of the kernel mean.

The following result presents an alternate representation for $\check{\boldsymbol{\mu}}_\lambda$, using which we relate the smooth operator formulation in (3.44) to the regularization formulation in (3.34).

Proposition 3.12. *Define $\Phi \triangleq [k(\mathbf{x}_1, \cdot), \ldots, k(\mathbf{x}_n, \cdot)]$ and $\mathbf{1}_n \triangleq [1/n, \ldots, 1/n]^\top$. Then*

$$\check{\boldsymbol{\mu}}_\lambda = \widehat{\mathbf{C}}_{XX}(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\hat{\boldsymbol{\mu}} = \Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n,$$

where $\mathbf{K}$ is the Gram matrix, I is an identity operator on $\mathscr{H}$ and $\mathbf{I}$ is an $n \times n$ identity matrix.

Proof of Proposition 3.12. Consider

$$(\widehat{\mathbf{C}}_{XX} + \lambda I)\Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n = (\widehat{\mathbf{C}}_{XX}\Phi + \lambda\Phi)(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n. \tag{3.46}$$

Note that $\widehat{\mathbf{C}}_{XX}\Phi = [\widehat{\mathbf{C}}_{XX}k(\cdot, \mathbf{x}_1), \ldots, \widehat{\mathbf{C}}_{XX}k(\cdot, \mathbf{x}_n)]$ where for any $i \in \{1, \ldots, n\}$,

$$\widehat{\mathbf{C}}_{XX}k(\cdot, \mathbf{x}_i) = \frac{1}{n}\sum_{j=1}^{n} k(\cdot, \mathbf{x}_j)k(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n}\Phi\mathbf{k}_i^\top$$

with $\mathbf{k}_i$ being the i^{th} row of $\mathbf{K}$. Therefore $\widehat{\mathbf{C}}_{XX}\Phi = \frac{1}{n}\Phi\mathbf{K}$. Using this in (3.46), we have

$$\begin{aligned}(\widehat{\mathbf{C}}_{XX} + \lambda I)\Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n &= (n^{-1}\Phi\mathbf{K} + \lambda\Phi)(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n \\ &= \Phi(n^{-1}\mathbf{K} + \lambda\mathbf{I})(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n \\ &= \frac{1}{n}\Phi\mathbf{K}\mathbf{1}_n = \widehat{\mathbf{C}}_{XX}\Phi\mathbf{1}_n = \widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}}.\end{aligned}$$

Multiplying to the left on both sides by $(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}$, we obtain $\Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n = (\widehat{\mathbf{C}}_{XX}+\lambda I)^{-1}\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}}$ and the result follows by noting that $(\widehat{\mathbf{C}}_{XX}+\lambda I)^{-1}\widehat{\mathbf{C}}_{XX} = \widehat{\mathbf{C}}_{XX}(\widehat{\mathbf{C}}_{XX}+\lambda I)^{-1}$. ■

From Proposition 3.12, it is clear that

$$\check{\boldsymbol{\mu}}_\lambda = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\boldsymbol{\beta}_s)_j k(\cdot, \mathbf{x}_j) \tag{3.47}$$

where $\boldsymbol{\beta}_s \triangleq \sqrt{n}(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$. Given the form of $\check{\boldsymbol{\mu}}_\lambda$ in (3.47), it is easy to verify that $\boldsymbol{\beta}_s$ is the minimizer of (3.34) when $\widehat{\mathcal{E}}_\lambda$ is minimized over $\{g = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\boldsymbol{\beta})_j k(\cdot, \mathbf{x}_j) : \boldsymbol{\beta} \in \mathbb{R}^n\}$ with $\Omega(\|g\|_{\mathscr{H}}) \triangleq \|\boldsymbol{\beta}\|_2^2$.

The following result establishes the consistency of S-KMSE, $\check{\mu}_\lambda$. We provide a discussion about its convergence rate in Remark 3.4(ii).

Theorem 3.13. *Suppose $\mathcal{X}$ is a Polish space that is also locally compact Hausdorff. Let k be a continuous kernel on $\mathcal{X}$ that is c_0-universal,* i.e.*, $k(\cdot, \mathbf{x}) \in C_0(\mathcal{X})$, $\forall\, \mathbf{x} \in \mathcal{X}$ and*

$$\int\int k(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mu(\mathbf{x})\, \mathrm{d}\mu(\mathbf{y}) > 0,\ \forall\, \mu \in M_b(\mathcal{X})\backslash\{0\}.$$

Then $\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} \to 0$ as $\lambda\sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.

Proof of Theorem 3.13. By Proposition 3.12, we have $\check{\boldsymbol{\mu}}_\lambda = (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}}$. Define $\boldsymbol{\mu}_\lambda \triangleq (\mathbf{C}_{XX}+\lambda I)^{-1}\mathbf{C}_{XX}\boldsymbol{\mu}$. Let us consider the decomposition $\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu} = (\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\lambda) + (\boldsymbol{\mu}_\lambda - \boldsymbol{\mu})$ with

$$\begin{aligned}\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\lambda &= (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}(\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}} - \widehat{\mathbf{C}}_{XX}\boldsymbol{\mu}_\lambda - \lambda\boldsymbol{\mu}_\lambda) \\ &\overset{(*)}{=} (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}(\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}} - \widehat{\mathbf{C}}_{XX}\boldsymbol{\mu}_\lambda - \mathbf{C}_{XX}\boldsymbol{\mu} + \mathbf{C}_{XX}\boldsymbol{\mu}_\lambda) \\ &= (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}) \\ &\qquad + (\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}),\end{aligned}$$

where we used $\lambda\boldsymbol{\mu}_\lambda = \mathbf{C}_{XX}\boldsymbol{\mu} - \mathbf{C}_{XX}\boldsymbol{\mu}_\lambda$ in $(*)$. By defining $\mathcal{A}(\lambda) \triangleq \|\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}}$, we have

$$\begin{aligned}\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} &\leq \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|_{\mathscr{H}} + \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu})\|_{\mathscr{H}} \\ &\quad + \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu})\|_{\mathscr{H}} + \mathcal{A}(\lambda) \\ &\leq \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}\|\,(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}} + \mathcal{A}(\lambda)) \\ &\quad + \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}\|\mathcal{A}(\lambda) + \mathcal{A}(\lambda),\end{aligned} \tag{3.48}$$

where for any bounded linear operator $\mathbf{B}$, $\|\mathbf{B}\|$ denotes its operator norm. We now bound $\|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}\|$ as follows. It is easy to show that

$$(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX} = \Big(I - (\mathbf{C}_{XX} + \lambda I)^{-1}(\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX})\Big)^{-1}(\mathbf{C}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}$$

$$= \left(\sum_{j=0}^{\infty} \left((\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}) \right)^j \right) (\mathbf{C}_{XX} + \lambda I)^{-1} \mathbf{C}_{XX},$$

where the last line denotes the Neumann series and therefore

$$\begin{aligned} \|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1} \mathbf{C}_{XX}\| &\leq \sum_{j=0}^{\infty} \left\| (\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}) \right\|^j \|(\mathbf{C}_{XX} + \lambda I)^{-1} \mathbf{C}_{XX}\| \\ &\leq \sum_{j=0}^{\infty} \left\| (\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}) \right\|_{\text{HS}}^j, \end{aligned}$$

where we used $\|(\mathbf{C}_{XX} + \lambda I)^{-1} \mathbf{C}_{XX}\| \leq 1$ and the fact that $\mathbf{C}_{XX}$ and $\widehat{\mathbf{C}}_{XX}$ are Hilbert-Schmidt operators on $\mathscr{H}$ as $\|\mathbf{C}_{XX}\|_{\text{HS}} \leq \kappa < \infty$ and $\|\widehat{\mathbf{C}}_{XX}\|_{\text{HS}} \leq \kappa < \infty$ with κ being the bound on the kernel. Define $\eta : \mathcal{X} \to \text{HS}(\mathscr{H})$, $\eta(\mathbf{x}) = (\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \mathbf{\Sigma}_{\mathbf{x}})$, where $\text{HS}(\mathscr{H})$ is the space of Hilbert-Schmidt operators on $\mathscr{H}$ and $\mathbf{\Sigma}_{\mathbf{x}} := k(\cdot, \mathbf{x}) \otimes k(\cdot, \mathbf{x})$. Observe that $\mathbb{E} \frac{1}{n} \sum_{i=1}^n \eta(\mathbf{x}_i) = 0$. Also, for all $i \in \{1, \ldots, n\}$, $\|\eta(\mathbf{x}_i)\|_{\text{HS}} \leq \|(\mathbf{C}_{XX} + \lambda I)^{-1}\| \|\mathbf{C}_{XX} - \mathbf{\Sigma}_{\mathbf{x}}\|_{\text{HS}} \leq \frac{2\kappa}{\lambda}$ and $\mathbb{E}\|\eta(\mathbf{x}_i)\|_{\text{HS}}^2 \leq \frac{4\kappa^2}{\lambda^2}$. Therefore, by Bernstein's inequality (see Theorem 3.8), for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$ over the choice of $\{\mathbf{x}_i\}_{i=1}^n$,

$$\|(\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX})\|_{\text{HS}} \leq \frac{\kappa \sqrt{2\tau}}{\lambda \sqrt{n}} + \frac{2\kappa\tau}{\lambda n} \leq \frac{\kappa \sqrt{2\tau} (\sqrt{2\tau} + 1)}{\lambda \sqrt{n}}.$$

For $\lambda \geq \frac{\kappa \sqrt{8\tau} (\sqrt{2\tau} + 1)}{\sqrt{n}}$, we obtain that $\|(\mathbf{C}_{XX} + \lambda I)^{-1} (\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX})\|_{\text{HS}} \leq \frac{1}{2}$ and therefore $\|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1} \mathbf{C}_{XX}\| \leq 2$. Using this along with $\|(\widehat{\mathbf{C}}_{XX} + \lambda I)^{-1} \widehat{\mathbf{C}}_{XX}\| \leq 1$ and (3.28) in (3.48), we obtain that for any $\tau > 0$ and $\lambda \geq \frac{\kappa \sqrt{8\tau} (\sqrt{2\tau} + 1)}{\sqrt{n}}$, with probability at least $1 - 2e^{-\tau}$ over the choice of $\{\mathbf{x}_i\}_{i=1}^n$,

$$\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} \leq \frac{\sqrt{2\kappa\tau} + 4\tau\sqrt{\kappa}}{\sqrt{n}} + 4\mathcal{A}(\lambda). \tag{3.49}$$

We now analyze $\mathcal{A}(\lambda)$. To this end, we make two observations.

1. Since k is continuous and $\mathcal{X}$ is Polish, $\mathscr{H}$ is separable (Steinwart and Christmann 2008; Lemma 4.33).

2. Since k is c_0-universal, $\nu \mapsto \int k(\cdot, \mathbf{x}) \, \mathrm{d}\nu(\mathbf{x})$, $\nu \in M_b(\mathcal{X})$ is injective, which implies $\mathbf{C}_{XX} f = \int k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) = 0 \Rightarrow f = 0$, *i.e.*, $\mathbf{C}_{XX}$ has a trivial null-space, meaning that $\overline{\mathcal{R}(\mathbf{C}_{XX})} = \mathscr{H}$.

Based on these observations along with the fact that $\mathbf{C}_{XX}$ is compact (as it is Hilbert-Schmidt), it follows from Sriperumbudur et al. (2013; Proposition A.2) that $\mathcal{A}(\lambda) \to 0$ as $\lambda \to 0$, therefore yielding the consistency of $\check{\boldsymbol{\mu}}_\lambda$. ■

Remark 3.4. *The following observations follow from Theorem 3.13 in comparison to R-KMSE.*

*(i) The kernel being c_0-universal is critical for the universal consistency of S-KMSE (*i.e.*, S-KMSE is consistent for any $\mathbb{P}$). This condition ensures that $\mathbf{C}_{XX}$ has a trivial null space, without which the consistency of S-KMSE is guaranteed if $\mu \in \overline{\mathcal{R}(\mathbf{C}_{XX})}$ (*i.e.*, requires the knowledge of $\mathbb{P}$) whereas no such assumption on k or $\mathbb{P}$ is required for the consistency of R-KMSE.*

(ii) A convergence rate for $\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}}$ can be obtained by bounding $\mathcal{A}(\lambda)$ in (3.49) where $\mathcal{A}(\lambda) = \|(\mathbf{C}_{XX} + \lambda I)^{-1}\mathbf{C}_{XX}\boldsymbol{\mu} - \boldsymbol{\mu}\|_{\mathscr{H}}$. The classical approach to bound $\mathcal{A}(\lambda)$ is to assume $\boldsymbol{\mu} \in \mathcal{R}(\mathbf{C}_{XX})$, i.e.*, range space of $\mathbf{C}_{XX}$, which then yields $\mathcal{A}(\lambda) \leq \|\mathbf{C}_{XX}^{-1}\boldsymbol{\mu}\|_{\mathscr{H}}\lambda$ (see Sriperumbudur et al. (2013; Proposition A.2)), thereby obtaining $\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} = O_{\mathbb{P}}(n^{-1/2})$ for $\lambda = cn^{-1/2}$ with $c > 0$ being a constant independent of n. However, it can be shown that $\boldsymbol{\mu}$ never lies in $\mathcal{R}(\mathbf{C}_{XX})$. To this end, suppose $\boldsymbol{\mu} \in \mathcal{R}(\mathbf{C}_{XX})$,* i.e.*, $\exists\, g \in \mathscr{H}$ such that $\boldsymbol{\mu} = \mathbf{C}_{XX}g = \int k(\cdot, \mathbf{x})g(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{x})$, which implies $\int k(\cdot, \mathbf{x})(g(\mathbf{x}) - 1)\,\mathrm{d}\mathbb{P}(\mathbf{x}) = 0$. Define $d\mu_1(\mathbf{x}) \triangleq g(\mathbf{x})\,\mathrm{d}\mathbb{P}(\mathbf{x})$. It is obvious that $\mu_1, \mathbb{P} \in M_b(\mathcal{X})$. Since k is c_0-universal, we therefore have $\mu_1 = \mathbb{P}$ which implies $g = 1$,* i.e.*, $1 \in \mathscr{H}$, yielding a contradiction as the assumption $k(\cdot, \mathbf{x}) \in C_0(\mathcal{X})$ ensures that $1 \notin \mathscr{H}$. Hence $\mu \notin \mathcal{R}(\mathbf{C}_{XX})$ and so the above argument cannot be used to bound $\mathcal{A}(\lambda)$. On the other hand, if $1 \in \mathscr{H}$ is assumed in place of k being c_0-universal, then the above argument can be used (as there exists $g \in \mathscr{H}$ such that $\boldsymbol{\mu} = \mathbf{C}_{XX}g$) to show that $\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}} = O_{\mathbb{P}}(n^{-1/2})$, however, compromising on the universal consistency of $\check{\boldsymbol{\mu}}_\lambda$. This is because if $1 \in \mathscr{H}$, then $\mathbf{C}_{XX}$ may have a non-trivial null space and therefore the consistency of $\check{\boldsymbol{\mu}}_\lambda$ is achieved if $\boldsymbol{\mu} \in \overline{\mathcal{R}(\mathbf{C}_{XX})}$ (see Sriperumbudur et al. (2013; Proposition A.2)). Owing to these issues, may be one should explore a different technique (than the one presented in the proof of Theorem 3.13) to obtain consistency and convergence rate for $\|\check{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_{\mathscr{H}}$.*

Note that the estimator $\check{\boldsymbol{\mu}}_\lambda$ requires the knowledge of the shrinkage or regularization parameter, λ. Similar to R-KMSE, below, we present a data dependent approach to select λ using leave-one-out cross validation. While the shrinkage parameter for R-KMSE can be obtained in a simple closed form (see Proposition 3.9), we will see below that finding the corresponding parameter for S-KMSE is more involved. Evaluating the score function (*i.e.*, (3.38)) naïvely requires one to solve for $\hat{\boldsymbol{\mu}}_\lambda^{(-i)}$ explicitly for every i, which is computationally expensive. The following result provides an alternate expression for the score, which can be evaluated more efficiently.[1]

Proposition 3.14. *The LOOCV score of S-KMSE is given by*

$$\begin{aligned} LOOCV(\lambda) \quad = \quad & \frac{1}{n}\mathrm{tr}\left((\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{K}(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{A}_\lambda\right) - \frac{2}{n}\mathrm{tr}\left((\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{B}_\lambda\right) \\ & + \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i), \end{aligned}$$

where $\lambda_n \triangleq (n-1)\lambda$, $\mathbf{A}_\lambda \triangleq \frac{1}{(n-1)^2}\sum_{i=1}^n \mathbf{c}_{i,\lambda}\mathbf{c}_{i,\lambda}^\top$, $\mathbf{B}_\lambda \triangleq \frac{1}{n-1}\sum_{i=1}^n \mathbf{c}_{i,\lambda}\mathbf{k}_i^\top$, $d_{i,\lambda} \triangleq \mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i$,

$$\begin{aligned} \mathbf{c}_{i,\lambda} \quad \triangleq \quad & \mathbf{K}\mathbf{1} - \mathbf{k}_i - \mathbf{e}_i\mathbf{k}_i^\top\mathbf{1} + \mathbf{e}_i k(\mathbf{x}_i, \mathbf{x}_i) + \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{K}\mathbf{1}}{1 - d_{i,\lambda}} - \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{k}_i}{1 - d_{i,\lambda}} \\ & - \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top\mathbf{1}}{1 - d_{i,\lambda}} + \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i k(\mathbf{x}_i, \mathbf{x}_i)}{1 - d_{i,\lambda}}, \end{aligned}$$

$\mathbf{k}_i$ is the i^{th} column of $\mathbf{K}$, $\mathbf{1} \triangleq (1, \ldots, 1)^\top$ and $\mathbf{e}_i \triangleq (0, 0, \ldots, 1, \ldots, 0)^\top$ with 1 being in the i^{th} position. Here $\mathrm{tr}(\mathbf{A})$ denotes the trace of a square matrix $\mathbf{A}$.

Proof of Proposition 3.14. From Proposition 3.12, we have $\check{\boldsymbol{\mu}}_\lambda^{(-i)} = (\widehat{\mathbf{C}}_{XX}^{(-i)} + \lambda I)^{-1}\widehat{\mathbf{C}}_{XX}^{(-i)}\hat{\boldsymbol{\mu}}^{(-i)}$ where $\widehat{\mathbf{C}}_{XX}^{(-i)} \triangleq \frac{1}{n-1}\sum_{j\neq i} k(\cdot, \mathbf{x}_j) \otimes k(\cdot, \mathbf{x}_j)$ and $\hat{\boldsymbol{\mu}}^{(-i)} \triangleq \frac{1}{n-1}\sum_{j\neq i} k(\cdot, \mathbf{x}_j)$. Define $\mathbf{a} \triangleq$

[1] An alternative–more efficient–formulation of LOOCV for S-KMSE is given in Appendix B.

$k(\cdot, \mathbf{x}_i)$. It is easy to verify that

$$\widehat{\mathbf{C}}_{XX}^{(-i)} = \frac{n}{n-1}\left(\widehat{\mathbf{C}}_{XX} - \frac{\mathbf{a}\otimes\mathbf{a}}{n}\right) \text{ and } \hat{\boldsymbol{\mu}}^{(-i)} = \frac{n}{n-1}\left(\hat{\boldsymbol{\mu}} - \frac{\mathbf{a}}{n}\right).$$

Therefore,

$$\check{\boldsymbol{\mu}}_\lambda^{(-i)} = \frac{n}{n-1}\left((\widehat{\mathbf{C}}_{XX} + \lambda_n' I) - \frac{\mathbf{a}\otimes\mathbf{a}}{n}\right)^{-1}\left(\widehat{\mathbf{C}}_{XX} - \frac{\mathbf{a}\otimes\mathbf{a}}{n}\right)\left(\hat{\boldsymbol{\mu}} - \frac{\mathbf{a}}{n}\right),$$

which after using Sherman-Morrison formula reduces to

$$\begin{aligned}\check{\boldsymbol{\mu}}_\lambda^{(-i)} &= \frac{n}{n-1}\left((\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1} + \frac{(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}}{n - \langle \mathbf{a}, (\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\mathbf{a}\rangle_{\mathscr{H}}}\right) \\ &\quad \left(\widehat{\mathbf{C}}_{XX} - \frac{\mathbf{a}\otimes\mathbf{a}}{n}\right)\left(\hat{\boldsymbol{\mu}} - \frac{\mathbf{a}}{n}\right),\end{aligned}$$

where $\lambda_n' \triangleq \frac{n-1}{n}\lambda$. Using the idea in the proof of Proposition 3.12, the following can be proved:

(i) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}} = n^{-1}\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{K}\mathbf{1}$.

(ii) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\widehat{\mathbf{C}}_{XX}\mathbf{a} = \Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{k}_i$.

(iii) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\mathbf{a} = n\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i$.

Based on the above, it is easy to show that

(iv) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})\hat{\boldsymbol{\mu}} = (\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\mathbf{a}\langle\mathbf{a}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}} = \Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top\mathbf{1}$.

(v) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})\mathbf{a} = (\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\mathbf{a}\langle\mathbf{a}, \mathbf{a}\rangle_{\mathscr{H}} = n\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i k(\mathbf{x}_i, \mathbf{x}_i)$.

(vi) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\widehat{\mathbf{C}}_{XX}\hat{\boldsymbol{\mu}} = \Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{K}\mathbf{1}$.

(vii) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\widehat{\mathbf{C}}_{XX}\mathbf{a} = n\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{k}_i$.

(viii) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})\hat{\boldsymbol{\mu}} = n\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top\mathbf{1}$.

(ix) $(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})(\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}(\mathbf{a}\otimes\mathbf{a})\mathbf{a} = n^2\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i k(\mathbf{x}_i, \mathbf{x}_i)$.

(x) $\langle\mathbf{a}, (\widehat{\mathbf{C}}_{XX} + \lambda_n' I)^{-1}\mathbf{a}\rangle_{\mathscr{H}} = n\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i$.

Using the above in $\check{\boldsymbol{\mu}}_\lambda^{(-i)}$, we obtain

$$\check{\boldsymbol{\mu}}_\lambda^{(-i)} = \frac{1}{n-1}\Phi(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{c}_{i,\lambda}.$$

Substituting the above in (3.38) yields the result. ■

Unlike R-KMSE, a closed form expression for the minimizer of $LOOCV(\lambda)$ in Proposition 3.14 is not possible and so proving the consistency of S-KMSE along with results similar to those in Theorem 3.10 are highly non-trivial. Hence, we are not able to provide any theoretical comparison of $\check{\boldsymbol{\mu}}_\lambda$ (with λ being chosen as a minimizer of $LOOCV(\lambda)$ in Proposition 3.14) with $\hat{\boldsymbol{\mu}}$. However, in the following section, we provide an empirical comparison through simulations where we show that the S-KMSE outperforms the empirical estimator.

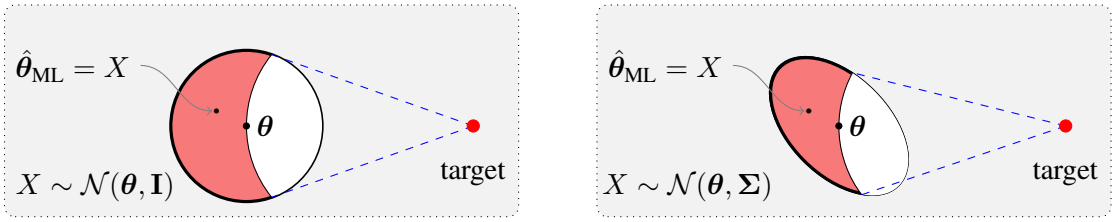


Figure 3.2: Geometric explanation of a shrinkage estimator when estimating a mean of a Gaussian distribution. For isotropic Gaussian, the level sets of the joint density of $\hat{\boldsymbol{\theta}}_{\text{ML}} = X$ are hyperspheres. In this case, shrinkage has the same effect regardless of the direction. Shaded area represents those estimates that get closer to $\boldsymbol{\theta}$ after shrinkage. For anisotropic Gaussian, the level sets are concentric ellipsoids, which makes the effect dependent on the direction of shrinkage.

3.4.2 Other Filtering Functions

As pointed out earlier, the effect of shrinkage achieved by S-KMSE is akin to *spectral filtering*. In this section, I provide extensions of S-KMSE using different filter functions. First, let us return to the shrinkage estimator $\hat{\boldsymbol{\mu}}_\alpha$ considered in Section 3.3, *i.e.*,

$$\hat{\boldsymbol{\mu}}_\alpha = \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}_{\mathbb{P}} = \alpha \sum_i \langle f^*, e_i \rangle e_i + (1-\alpha) \sum_i \langle \hat{\boldsymbol{\mu}}_{\mathbb{P}}, e_i \rangle e_i,$$

where $(e_i)_{i\in\mathbb{N}}$ are the countable orthonormal basis (ONB) of $\mathscr{H}$—countable ONB exist since $\mathscr{H}$ is separable which follows from $\mathcal{X}$ being separable and k being continuous (Steinwart and Christmann 2008; Lemma 4.33). This estimator can be generalized by considering the shrinkage estimator

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} := \sum_i \alpha_i \langle f^*, e_i \rangle e_i + \sum_i (1-\alpha_i) \langle \hat{\boldsymbol{\mu}}_{\mathbb{P}}, e_i \rangle e_i$$

where $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots) \in \mathbb{R}^\infty$ is a sequence of shrinkage parameters. If $\Delta_{\boldsymbol{\alpha}} := \mathbb{E}_{\mathbb{P}} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\mu}_{\mathbb{P}}\|^2$ is the risk of this estimator, the following theorem gives an optimality condition on $\boldsymbol{\alpha}$ for which $\Delta_{\boldsymbol{\alpha}} < \Delta$ (see Appendix C.2 for the proof).

Theorem 3.15. *For some ONB $(e_i)_i$, $\Delta_{\boldsymbol{\alpha}} - \Delta = \sum_i (\Delta_{\boldsymbol{\alpha},i} - \Delta_i)$ where $\Delta_{\boldsymbol{\alpha},i}$ and Δ_i denote the risk of the ith component of $\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$, respectively. Then, $\Delta_{\boldsymbol{\alpha},i} - \Delta_i < 0$ if*

$$0 < \alpha_i < \frac{2\Delta_i}{\Delta_i + (f_i^* - \mu_i)^2}, \tag{3.50}$$

where f_i^* and μ_i denote the Fourier coefficients of f^* and $\boldsymbol{\mu}_{\mathbb{P}}$, respectively.

The condition in (3.50) is a component-wise version of the condition given in Theorem 3.3 (see also Muandet et al. (2014a; Theorem 1)) for a class of estimators $\hat{\boldsymbol{\mu}}_\alpha := \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ which may be expressed here by assuming that we have a constant shrinkage parameter $\alpha_i = \alpha$ for all i. Clearly, as the optimal range of α_i may vary across coordinates, the class of estimators in Muandet et al. (2014a) does not allow us to adjust α_i accordingly. To understand why this property is important, let us consider the problem of estimating the mean of Gaussian distribution illustrated in Figure 3.2. For correlated random variable $X \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, a natural choice of basis is the set of orthonormal eigenvectors which diagonalize the covariance matrix $\boldsymbol{\Sigma}$ of X. Clearly, the optimal range of α_i depends on the corresponding eigenvalues. Allowing for different basis $(e_i)_i$ and shrinkage parameter α_i opens up a wide range of strategies that can be used to construct "better" estimators.

A natural strategy under this representation is as follows: *i*) we specify the ONB $(e_i)_i$ and project $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ onto this basis. *ii*) we shrink each $\hat{\mu}_i$ independently according to a pre-defined

shrinkage rule. *iii*) the shrinkage estimate is reconstructed as a superposition of the resulting components. In other words, an ideal shrinkage estimator can be defined formally as a non-linear mapping:

$$\hat{\boldsymbol{\mu}}_{\mathbb{P}} \longrightarrow \sum_i h(\alpha_i)\langle f^*, e_i\rangle e_i + \sum_i (1 - h(\alpha_i))\langle \hat{\boldsymbol{\mu}}_{\mathbb{P}}, e_i\rangle e_i \tag{3.51}$$

where $h : \mathbb{R} \to \mathbb{R}$ is a shrinkage rule. Since we make no reference to any particular basis $(e_i)_i$, nor to any particular shrinkage rule h, a wide range of strategies can be adopted here. For example, we can view *whitening* as a special case in which f^* is the data average $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and $1 - h(\alpha_i) = 1/\sqrt{\alpha_i}$ where α_i and e_i are the ith eigenvalue and eigenvector of the covariance matrix, respectively.

Inspired by Theorem 3.15, we adopt the spectral filtering approach as one of the strategies to construct the estimators of the form (3.51). To this end, owing to the regularization interpretation, we consider estimators of the form $\sum_{i=1}^n \beta_i k(\mathbf{x}_i, \cdot)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$—looking for such an estimator is equivalent to learning a *signed measure* that is supported on $(\mathbf{x}_i)_{i=1}^n$. Since $\sum_{i=1}^n \beta_i k(\mathbf{x}_i, \cdot)$ is a minimizer of (3.44), β should satisfy $\mathbf{K}\boldsymbol{\beta} = \mathbf{K}\mathbf{1}_n$. Here the solution is trivially $\boldsymbol{\beta} = \mathbf{1}_n$, *i.e.*, the coefficients of the standard estimator $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ if $\mathbf{K}$ is invertible. Since $\mathbf{K}^{-1}$ may not exist and even if it exists, the computation of it can be numerically unstable, the idea of spectral filtering—this is quite popular in the theory of inverse problems (Engl et al. 1996) and has been used in kernel least squares (Vito et al. 2005)—is to replace $\mathbf{K}^{-1}$ by some regularized matrices $g_\lambda(\mathbf{K})$ that approximates $\mathbf{K}^{-1}$ as λ goes to zero. Note that unlike in standard formulation of kernel mean estimation, the regularization is quite important here (*i.e.*, the case of estimators of the form $\sum_{i=1}^n \beta_i k(\mathbf{x}_i, \cdot)$) without which the the linear system is under determined. Therefore, we propose the following class of estimators:

$$\hat{\boldsymbol{\mu}}_\lambda := \sum_{i=1}^n \beta_i k(\mathbf{x}_i, \cdot) \quad \text{with} \quad \boldsymbol{\beta}(\lambda) := g_\lambda(\mathbf{K})\mathbf{K}\mathbf{1}_n, \tag{3.52}$$

where $g_\lambda(\cdot)$ is a filter function and λ is referred to as a shrinkage parameter. The matrix-valued function $g_\lambda(\mathbf{K})$ can be described by a scalar function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ on the spectrum of $\mathbf{K}$. That is, if $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ is the eigen-decomposition of $\mathbf{K}$ where $\mathbf{D} = \operatorname{diag}(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n)$, we have $g_\lambda(\mathbf{D}) = \operatorname{diag}(g_\lambda(\tilde{\gamma}_1), \ldots, g_\lambda(\tilde{\gamma}_n))$ and $g_\lambda(\mathbf{K}) = \mathbf{U}g_\lambda(\mathbf{D})\mathbf{U}^\top$. For example, the scalar filter function of Tikhonov regularization is $g_\lambda(\gamma) = 1/(\gamma + \lambda)$. In the sequel, I will also refer to this class of estimators as *Spectral-KMSE*. As we will see later, the estimator $\breve{\boldsymbol{\mu}}_{\mathbb{P}}$ presented previously belongs to this class.

Similar to Proposition 3.11, the following proposition expresses $\hat{\boldsymbol{\mu}}_\lambda$ in terms of the eigenfunctions of $\widehat{\mathbf{C}}_{XX}$ and the filter function g_λ (the proof, which is similar to that of Proposition 3.11, is given in Appendix C.3).

Proposition 3.16. *Let (γ_i, ϕ_i) be eigenvalue and eigenfunction pairs of the empirical covariance operator $\widehat{\mathbf{C}}_{XX}$. The Spectral-KMSE satisfies $\hat{\boldsymbol{\mu}}_\lambda = \sum_{i=1}^n g_\lambda(\gamma_i)\gamma_i\langle\hat{\boldsymbol{\mu}}, \phi_i\rangle\phi_i$, .*

By virtue of Proposition 3.16, if we choose $1 - h(\gamma) := g_\lambda(\gamma)\gamma$, the Spectral-KMSE is indeed in the form of (3.51) when $f^* = 0$ and $(e_i)_i$ is the kernel PCA (KPCA) basis, with the filter function g_λ determining the shrinkage rule. Since by definition $g_\lambda(\gamma_i)$ approaches the function $1/\gamma_i$ as λ goes to 0, the function $g_\lambda(\gamma_i)\gamma_i$ approaches 1 (no shrinkage). As the value of λ increases, we have more shrinkage because the value of $g_\lambda(\gamma_i)\gamma_i$ deviates from 1, and the behavior of this deviation depends on the filter function g_λ. For example, we can see that Proposition 3.16 generalizes Proposition 3.11 (Theorem 2 in Muandet et al. (2014a)) where the filter function is $g_\lambda(\mathbf{K}) = (\mathbf{K} + n\lambda\mathbf{I})^{-1}$, *i.e.*, $g(\gamma) = 1/(\gamma + \lambda)$. That is, we have $g_\lambda(\gamma_i)\gamma_i = \gamma_i/(\gamma_i + \lambda)$,

Table 3.1: Update equations for $\boldsymbol{\beta}$ and corresponding filter functions.

Algorithm	Update Equation ($\mathbf{a} := \mathbf{K}\mathbf{1}_n - \mathbf{K}\boldsymbol{\beta}^{t-1}$)	Filter Function
L2 Boosting	$\boldsymbol{\beta}^t \leftarrow \boldsymbol{\beta}^{t-1} + \eta\mathbf{a}$	$g(\gamma) = \eta\sum_{i=1}^{t-1}(1-\eta\gamma)^i$
Acc. L2 Boosting	$\boldsymbol{\beta}^t \leftarrow \boldsymbol{\beta}^{t-1} + \omega_t(\boldsymbol{\beta}^{t-1} - \boldsymbol{\beta}^{t-2}) + \frac{\kappa_t}{n}\mathbf{a}$	$g(\gamma) = p_t(\gamma)$
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	$g(\gamma) = \gamma^{-1}\mathbb{1}_{\{\gamma\geq\lambda\}}$

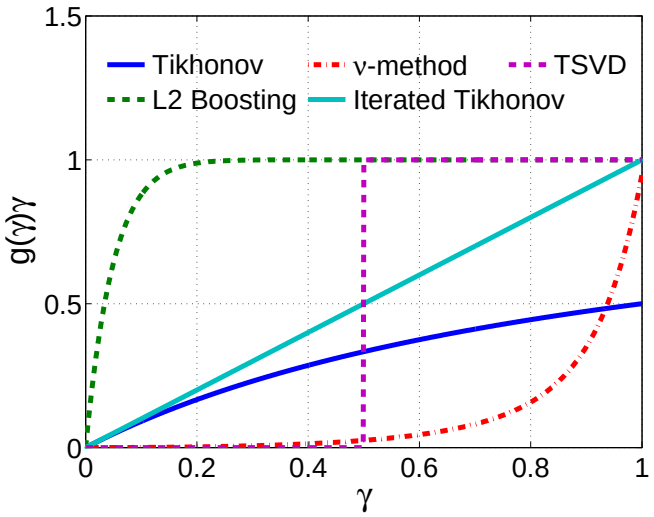


Figure 3.3: Plot of $g(\gamma)\gamma$.

implying that the effect of shrinkage is relatively larger in the low-variance direction. In the following, we discuss well-known examples of spectral filtering algorithms obtained by various choices of g_λ. Update equations for $\boldsymbol{\beta}(\lambda)$ and corresponding filter functions are summarized in Table 3.1. Figure 3.3 illustrates the behavior of these filter functions.

L2 Boosting. This algorithm, also known as gradient descent or Landweber iteration, finds a weight $\boldsymbol{\beta}$ by performing a gradient descent iteratively. Thus, we can interpret *early stopping* as shrinkage and the reciprocal of iteration number as shrinkage parameter, *i.e.*, $\lambda \approx 1/t$. The step-size η does not play any role for shrinkage (Vito et al. 2006), so we use the fixed step-size $\eta = 1/\kappa^2$ throughout.

Accelerated L2 Boosting. This algorithm, also known as ν-method, uses an accelerated gradient descent step, which is faster than L2 Boosting because we only need $\sqrt{t}$ iterations to get the same solution as the L2 Boosting would get after t iterations. Consequently, we have $\lambda \approx 1/t^2$.

Iterated Tikhonov. This algorithm can be viewed as a combination of Tikhonov regularization and gradient descent. Both parameters λ and t play the role of shrinkage parameter.

Truncated Singular Value Decomposition. This algorithm can be interpreted as a projection onto the first principal components of the KPCA basis. Hence, we may interpret *dimensionality reduction* as shrinkage and the size of reduced dimension as shrinkage parameter. This approach

has been used in Song and Dai (2013) to improve the kernel mean estimation under the low-rank assumption.

Most of the above spectral filtering algorithms allow one to compute the coefficients $\boldsymbol{\beta}$ without explicitly computing the eigen-decomposition of $\mathbf{K}$, as we can see in Table 3.1, and some of which may have no natural interpretation in terms of regularized risk minimization considered previously. Lastly, an initialization of $\boldsymbol{\beta}$ corresponds to the target of shrinkage. In what follows, I assume that $\boldsymbol{\beta}^0 = 0$ throughout.

3.4.3 Theoretical Properties of Spectral-KMSE

It is of interest to study the consistency and convergence rate of $\hat{\boldsymbol{\mu}}_\lambda$. Our main goal here is to derive convergence rates for a broad class of algorithms given a set of sufficient conditions on the filter function g_λ. We believe that for some algorithms it is possible to derive the best achievable bounds, which requires ad-hoc proofs for each algorithm. To this end, we provide a set of conditions any *admissible* filter function, g_λ must satisfy.

Definition 3.1. *A family of filter functions $g_\lambda : [0, \kappa^2] \to \mathbb{R}, 0 < \lambda \leq \kappa^2$ is said to be admissible if there exists finite positive constants B, C, D, and η_0 (all independent of λ) such that*

(C1) $\sup_{\gamma \in [0,\kappa^2]} |\gamma g_\lambda(\gamma)| \leq B$,

(C2) $\sup_{\gamma \in [0,\kappa^2]} |r_\lambda(\gamma)| \leq C$,

(C3) $\sup_{\gamma \in [0,\kappa^2]} |r_\lambda(\gamma)| \gamma^\eta \leq D\lambda^\eta$, $\forall\, \eta \in (0, \eta_0]$ *hold, where* $r_\lambda(\gamma) := 1 - \gamma g_\lambda(\gamma)$.

These conditions are quite standard in the theory of inverse problems (Engl et al. 1996, Gerfo et al. 2008). The constant η_0 is called the *qualification* of g_λ and is a crucial factor that determines the rate of convergence in inverse problems. As we will see below, that the rate of convergence of $\hat{\boldsymbol{\mu}}_\lambda$ depends on two factors: (a) smoothness of $\boldsymbol{\mu}_\mathbb{P}$ which is usually unknown as it depends on the unknown $\mathbb{P}$ and (b) qualification of g_λ which determines how well the smoothness of $\boldsymbol{\mu}_\mathbb{P}$ is captured by the spectral filter, g_λ.

Theorem 3.17. *Suppose g_λ is admissible in the sense of Definition 3.1. Let $\kappa = \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{k(\mathbf{x}, \mathbf{x})}$. If $\boldsymbol{\mu}_\mathbb{P} \in \mathcal{R}(\mathbf{C}_{XX}^\beta)$ for some $\beta > 0$, then for any $\delta > 0$, with probability at least $1 - 3e^{-\delta}$,*

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\mathbb{P}\| \leq \frac{2\kappa B + \kappa B\sqrt{2\delta}}{\sqrt{n}} + D\lambda^{\min\{\beta,\eta_0\}} \|\mathbf{C}_{XX}^{-\beta} \boldsymbol{\mu}_\mathbb{P}\| + C\tau \frac{(2\sqrt{2}\kappa^2\sqrt{\delta})^{\min\{1,\beta\}}}{n^{\min\{1/2,\beta/2\}}} \|\mathbf{C}_{XX}^{-\beta} \boldsymbol{\mu}_\mathbb{P}\|,$$

where $\mathcal{R}(A)$ denotes the range space of A and τ is some universal constant that does not depend on λ and n. Therefore,

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\mathbb{P}\| = O_\mathbb{P}\left(n^{-\min\{1/2,\beta/2\}}\right) \qquad \textit{with} \qquad \lambda = o\left(n^{-\frac{\min\{1/2,\beta/2\}}{\min\{\beta,\eta_0\}}}\right).$$

Theorem 3.17 shows that the convergence rate depends on the smoothness of $\boldsymbol{\mu}_\mathbb{P}$ which is imposed through the range space condition that $\boldsymbol{\mu}_\mathbb{P} \in \mathcal{R}(\mathbf{C}_{XX}^\beta)$ for some $\beta > 0$. Note that this is in contrast to the estimator in Section 3.3 which does not require any smoothness assumptions on $\boldsymbol{\mu}_\mathbb{P}$. It can be shown that the smoothness of $\boldsymbol{\mu}_\mathbb{P}$ increases with increase in β. This means, irrespective of the smoothness of $\boldsymbol{\mu}_\mathbb{P}$ for $\beta > 1$, the best possible convergence rate is $n^{-1/2}$ which matches with that of KMSE in Section 3.3. While the qualification η_0 does not seem to directly affect the rates, it controls the rate at which λ converges to zero. For example, if $g_\lambda(\gamma) = 1/(\gamma + \lambda)$ which corresponds to Tikhonov regularization, it can be shown that $\eta_0 = 1$ which means for $\beta > 1$, $\lambda = o(n^{-1/2})$ implying that λ cannot decay to zero slower than $n^{-1/2}$.

Ideally, one would require a larger η_0 (preferably infinity which is the case with truncated SVD) so that the convergence of λ to zero can be made arbitrarily slow if β is large. This way, both β and η_0 control the behavior of the estimator.

In fact, Theorem 3.17 provides a choice for λ to construct the Spectral-KMSE. However, this choice of λ depends on β which is not known in practice (although η_0 is known as it is determined by the choice of g_λ). Therefore, λ is usually learnt from data through cross-validation or through Lepski's method (Lepski et al. 1997) for which guarantees similar to the one presented in Theorem 3.17 can be provided. However, irrespective of the data-dependent/independent choice for λ, checking for the admissibility of Spectral-KMSE is very difficult and we intend to consider it in future work.

3.5 Sparse Approximation

The regression perspective of kernel mean estimation allows us to construct diverse estimators of kernel mean by imposing different regularizers. Assume that $\hat{\boldsymbol{\mu}} = \sum_{i=1}^n \beta_j \phi(\mathbf{x}_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. An interesting choice of regularizer is the ell_1-norm of $\boldsymbol{\beta}$, which in addition to shrinkage also induces the sparsity on $\boldsymbol{\beta}$. Hence, the sparse KME can be formulated as follow:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^n \left\| k(\mathbf{x}_i, \cdot) - \sum_{j=1}^n \beta_j k(\mathbf{x}_j, \cdot) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{3.53}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_i |\beta_i|$. It is important to note that this formulation is fundamentally different from the standard lasso formulation (Tibshirani 1996). That is, we essentially performs a *sample selection*, whereas the lasso (and most of its extensions) often performs a *variable selection*.

Despite the difference, it is possible to adopt a lasso software package in our problem. First, recall that the solution of the standard KME can be obtained by solving the systems of linear equations $\mathbf{K}\boldsymbol{\beta} = \mathbf{K}\mathbf{1}_n$. Letting $\mathbf{y} = \mathbf{K}\mathbf{1}_n$, the sparse KME formulation can then be rewritten in the standard form as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2n} \|\mathbf{K}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{3.54}$$

That is, the matrix $\mathbf{K}$ acts as a design matrix and $\mathbf{y}$ is a regression target. Consequently, several solvers for lasso can be used to learn the sparse KME.

The sparse representation of the kernel mean is useful in some applications of kernel mean embedding such as reinforcement learning and the state-space model because the kernel mean has to be applied repeatedly (Kanagawa et al. 2013, McCalman et al. 2013). In Grünewälder et al. (2012), the sparse conditional mean embedding has been proposed using the regression formulation of the conditional kernel mean. Additionally, the proposed sparse kernel mean can be applied in applications where one need to find a summary of the dataset, *e.g.*, choosing pivot points in Nyström method, data summarization and squashing, etc. I will consider this in greater detail in future works.

As a demonstration I perform a simple simulation using the sparse approximation of kernel mean. Figure 3.4 depicts the average losses of the KME and its approximations using subsampling, LASSO, and elastic net. Clearly, the estimators obtained from minimizing (3.54) using LASSO and elastic net provides a much better approximation than the subsampling method at the same level of sparsity.

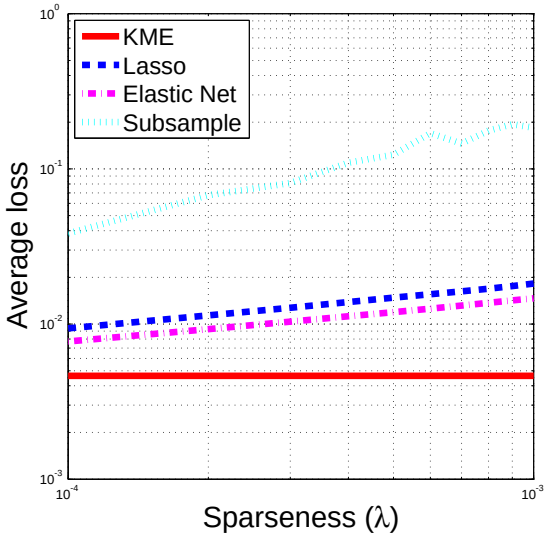


Figure 3.4: The comparison between the KME and its sparse approximations obtained from (3.54).

3.6 Probabilistic View

This section presents the probabilistic account of the kernel mean estimator and shrinkage. We will see that while the S-KMSE can be viewed as a posterior mean obtained from the product of a prior and a data-dependent likelihood, the R-KMSE cannot. First, recall the primal form of loss functional:

$$\mathcal{E}(g) \triangleq \frac{1}{n}\sum_{i=1}^{n} \|\phi(\mathbf{x}_i) - g\|_{\mathscr{H}}^2 , \quad g \in \mathscr{H}. \tag{3.55}$$

Estimating g directly using (3.55) can be difficult as the RKHS $\mathscr{H}$ is usually high-dimensional, if not infinite. By representer theorem, we have $g = \sum_{i=1}^{n} \beta_i \phi(\mathbf{x}_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. As a result, we can transform Equation (3.55) into its dual form

$$\mathcal{E}^*(\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \left\| \phi(\mathbf{x}_i) - \sum_{j=1}^{n} \beta_j \phi(\mathbf{x}_j) \right\|_{\mathscr{H}}^2 , \quad \boldsymbol{\beta} \in \mathbb{R}^n. \tag{3.56}$$

Consequently, the estimation of g is amount to estimating the weight vector $\boldsymbol{\beta}$. Simple calculation gives the dual form (3.56) in term of the kernel matrix $\mathbf{K}$ as

$$\mathcal{E}^*(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{K} \mathbf{1}_n + \frac{1}{n}\mathrm{trace}(\mathbf{K}) . \tag{3.57}$$

The standard kernel mean estimator

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

can be obtained as a minimizer of the primal form (3.55) and the corresponding value of $\boldsymbol{\beta}$, *i.e.*, $\boldsymbol{\beta} = \mathbf{1}_n$, is a minimizer of the dual form (3.56). We assume that the kernel matrix $\mathbf{K}$ is invertible.

It is straightforward to see that the dual form (3.57) is quadratic in $\boldsymbol{\beta}$, which implies that the same solution can be obtained by minimizing a negative log-likelihood of some Gaussian

distribution over $\boldsymbol{\beta}$. That is, let $\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\nu}, \boldsymbol{\Sigma})$ be the Gaussian distribution over $\boldsymbol{\beta}$ with mean $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma}$. Consequently, we have

$$\begin{aligned}
\mathcal{E}'(\boldsymbol{\beta}) &\triangleq -\ln \mathcal{N}(\boldsymbol{\beta}; \mathbf{1}_n, \mathbf{K}^{-1}) \\
&= -\ln \left[\frac{1}{\sqrt{(2\pi)^n |\mathbf{K}^{-1}|}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \mathbf{1}_n)^\top \mathbf{K} (\boldsymbol{\beta} - \mathbf{1}_n) \right) \right] \\
&= \ln \sqrt{(2\pi)^n |\mathbf{K}^{-1}|} + \frac{1}{2} (\boldsymbol{\beta} - \mathbf{1}_n)^\top \mathbf{K} (\boldsymbol{\beta} - \mathbf{1}_n) \\
&= \ln \sqrt{(2\pi)^n |\mathbf{K}^{-1}|} + \frac{1}{2} \mathbf{1}_n \mathbf{K} \mathbf{1}_n + \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{K} \mathbf{1}_n \\
&= \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{K} \mathbf{1}_n + const,
\end{aligned}$$

where $const$ denotes constant terms that do not depend on $\boldsymbol{\beta}$. It is easy to see that $\mathcal{E}^*(\boldsymbol{\beta})$ and $\mathcal{E}'(\boldsymbol{\beta})$ have the same minimizer, *i.e.*, $\boldsymbol{\beta} = \mathbf{1}_n$. If the Gram matrix $\mathbf{K}$ is strictly positive-definite, the minimizer is unique.

As a result, the weight vector $\boldsymbol{\beta}$ of the standard kernel mean estimator can be considered as a maximum-likelihood estimate of the probability distribution $\mathcal{N}(\boldsymbol{\beta}; \mathbf{1}_n, \mathbf{K}^{-1})$, which differs from the likelihood in the usual sense, *i.e.*, the probability density of the observations given the parameters. Instead, it specifies the probability density of the weight vector $\boldsymbol{\beta}$. In the following we will denote $\mathcal{N}(\boldsymbol{\beta}; \mathbf{1}_n, \mathbf{K}^{-1})$ by $\mathbb{P}_X$ to emphasize its dependence on the data.

Despite being different from the standard likelihood, the distribution $\mathbb{P}_X$ may still be interpreted as a *data-dependent* belief over possible values of $\boldsymbol{\beta}$. Following standard Bayesian formalism, one may want to specify alternative belief over the values of $\boldsymbol{\beta}$. For example,

$$\mathbb{P}_M \triangleq \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \boldsymbol{\Sigma}).$$

Combining $\mathbb{P}_X$ and $\mathbb{P}_M$ yields

$$\begin{aligned}
Q \triangleq \mathbb{P}_X \cdot \mathbb{P}_M &= \mathcal{N}(\boldsymbol{\beta}; \mathbf{1}_n, \mathbf{K}^{-1}) \cdot \mathcal{N}(\boldsymbol{\beta}; \mathbf{0}, \boldsymbol{\Sigma}) \\
&\propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \mathbf{1}_n)^\top \mathbf{K} (\boldsymbol{\beta} - \mathbf{1}_n) \right) \exp\left(-\frac{1}{2} \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} \right) \\
&\propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) (\mathbf{K} + \boldsymbol{\Sigma}^{-1}) (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \right)
\end{aligned}$$

where $\bar{\boldsymbol{\beta}} = (\mathbf{K} + \boldsymbol{\Sigma}^{-1})^{-1} \mathbf{K} \mathbf{1}_n$ and this is recognized as the form of Gaussian with mean $\bar{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{A}^{-1}$

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}; \bar{\boldsymbol{\beta}}, \mathbf{A}^{-1})$$

where $\mathbf{A} = \mathbf{K} + \boldsymbol{\Sigma}^{-1}$. By imposing different prior $\mathbb{P}_M$ on $\boldsymbol{\beta}$, we would obtain different mean $\bar{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{A}^{-1}$. For example, if $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ where σ^2 specifies the uncertainty of our belief, we have

$$\bar{\boldsymbol{\beta}} = (\mathbf{K} + \sigma^{-2} \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$$

which corresponds to the S-KMSE if we set $n\lambda = \sigma^{-2}$. Alternatively, one may consider the covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{K}^{-1}$ which reflects covariance structure obtained from the observations. In which case, we have

$$\bar{\boldsymbol{\beta}} = (\mathbf{K} + \sigma^{-2} \mathbf{K})^{-1} \mathbf{K} \mathbf{1}_n = \frac{1}{1 + \sigma^{-2}} \mathbf{K}^{-1} \mathbf{K} \mathbf{1}_n = \frac{1}{1 + \sigma^{-2}} \mathbf{1}_n$$

which corresponds to the R-KMSE if we set $\lambda = \sigma^{-2}$.

In other words, if we think of $\mathbb{P}_X$ as a likelihood, then it encodes the dependence of $\boldsymbol{\beta}$ on the observations $\mathbf{x}$ through the Gram matrix $\mathbf{K}$. For S-KMSE, the prior $\mathbb{P}_M$ is independent of the observations, whereas the "prior" of R-KMSE is data-dependent - it is a function of the $\mathbf{x}$. Hence, S-KMSE can be written as the product of a prior and a data-dependent likelihood as in the standard Bayesian formalism, but R-KMSE cannot. Thus, it is different from standard Bayesian formalism (Rasmussen and Williams 2005; Chapter 2.1). Moreover, the variance term σ^2 plays similar role to the regularization parameter λ. That is, the more we are uncertain about the alternative value of $\boldsymbol{\beta}$, the less we should shrink toward it.

3.7 Experimental Results

In this section, we empirically compare the proposed shrinkage estimators to the standard estimator of the kernel mean on both synthetic and real-world datasets. Specifically, we consider the following estimators: *i*) empirical/standard kernel mean estimator (KME), *ii*) KMSE whose parameter is obtained via empirical bound (B-KMSE), *iii*) regularized KMSE whose parameter is obtained via Proposition 3.9 (R-KMSE), and *iv*) spectral KMSE whose parameter is obtained via Proposition 3.14 (S-KMSE).

3.7.1 Synthetic Data

Given the true data-generating distribution $\mathbb{P}$ and the i.i.d. sample $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ from $\mathbb{P}$, we evaluate different estimators using the loss function

$$L(\boldsymbol{\beta}, \mathbf{X}, \mathbb{P}) \triangleq \left\| \sum_{i=1}^{n} \beta_i k(\mathbf{x}_i, \cdot) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \cdot)] \right\|_{\mathscr{H}}^2, \tag{3.58}$$

where $\boldsymbol{\beta}$ is the weight vector associated with different estimators. Then, we can estimate the risk of the estimator by averaging over m independent copies of X, *i.e.*, $\widehat{R} = \frac{1}{m}\sum_{j=1}^{m} L(\boldsymbol{\beta}_j, \mathbf{X}_j, \mathbb{P})$.

To allow for an exact calculation of $L(\boldsymbol{\beta}, \mathbf{X}, \mathbb{P})$, we consider $\mathbb{P}$ to be a mixture-of-Gaussians distribution and k being one of the following kernel functions: *i*) linear kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$, *ii*) polynomial degree-2 kernel $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + 1)^2$, *iii*) polynomial degree-3 kernel $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + 1)^3$ and *iv*) Gaussian RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2\right)$. We refer to them as LIN, POLY2, POLY3, and RBF, respectively. The analytic forms of $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \cdot)]$ for Gaussian distribution are given in Song et al. (2008) and Muandet et al. (2012). Unless otherwise stated, we set the bandwidth parameter of the Gaussian kernel as $\sigma^2 = \text{median}\left\{\|\mathbf{x}_i - \mathbf{x}_j\|^2 : i, j = 1, \ldots, n\right\}$, *i.e.*, the median heuristic.

Gaussian Distribution

We begin our empirical studies by considering the simplest case in which the distribution $\mathbb{P}$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$ on $\mathbb{R}^d$ where $d = 1, 2, 3$ and k is a linear kernel. In this case, the problem of kernel mean estimation reduces to just estimating the mean $\boldsymbol{\mu}$ of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$. We consider only shrinkage estimators of form $\hat{\boldsymbol{\mu}}_\alpha = \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}$. The true mean $\boldsymbol{\mu}$ of the distribution is chosen to be 1, $(1, 0)^\top$, and $(1, 0, 0)^\top$, respectively. Figure 3.5 depicts the comparison between the standard estimator and the shrinkage estimator, $\hat{\boldsymbol{\mu}}_\alpha$ when the target f^* is the origin. We can clearly see that even in this simple case, an improvement can be gained by applying a small shrinkage. Furthermore, the improvement becomes more substantial as we increase the dimensionality of the underlying space. Figure 3.6 illustrates similar results when $f^* \neq 0$ but $f^* \in \{2, (2, 0)^\top, (2, 0, 0)^\top\}$. Interestingly, we can still observe

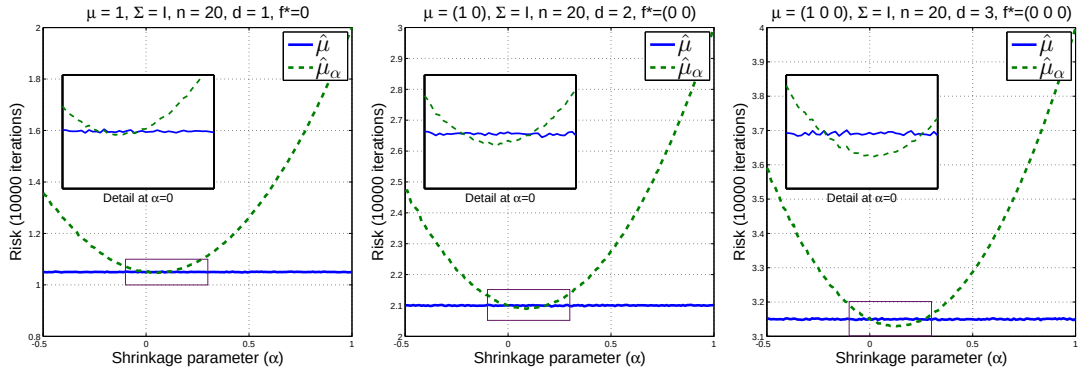


Figure 3.5: The comparison between standard estimator, $\hat{\boldsymbol{\mu}}$ and shrinkage estimator, $\hat{\boldsymbol{\mu}}_\alpha$ (with $f^* = 0$) of the mean of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^d$ where $d = 1, 2, 3$.

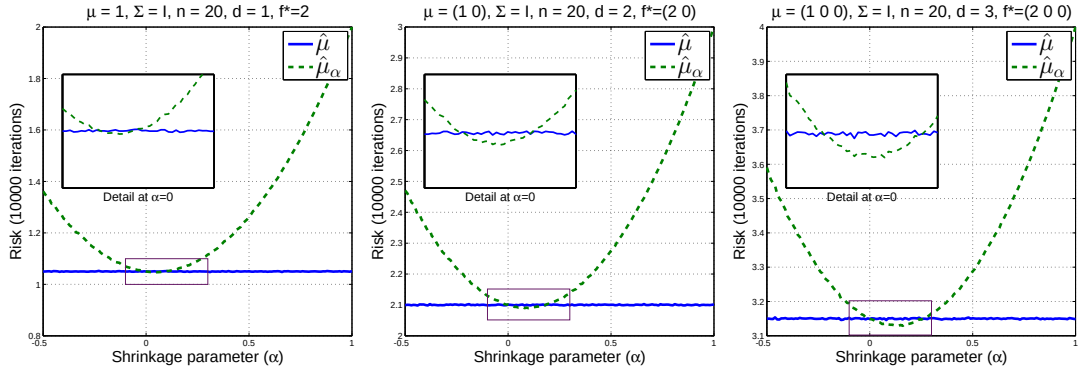


Figure 3.6: The risk comparison between standard estimator, $\hat{\boldsymbol{\mu}}$ and shrinkage estimator, $\hat{\boldsymbol{\mu}}_\alpha$ (with $f^* \in \{2, (2,0)^\top, (2,0,0)^\top\}$) of the mean of the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on $\mathbb{R}^d$ where $d = 1, 2, 3$.

similar improvement, which demonstrates that the choice of target f^* can be arbitrary when no prior knowledge about $\boldsymbol{\mu}_\mathbb{P}$ is available.

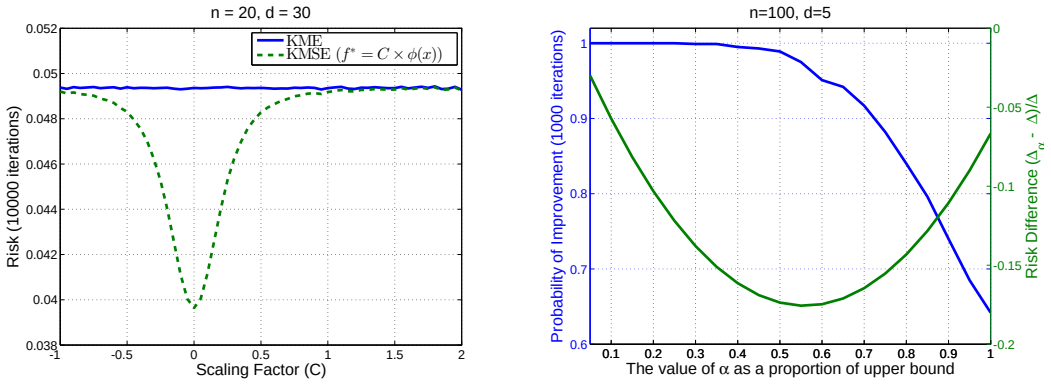
Mixture of Gaussians Distributions

To simulate a more realistic case, let $\mathbf{y}$ be a sample from $\mathbb{P} \triangleq \sum_{i=1}^4 \pi_i \mathcal{N}(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$. In the following experiments, the sample $\mathbf{x}$ is generated from the following generative process:

$$\mathbf{x} = \mathbf{y} + \varepsilon, \quad \theta_{ij} \sim \mathcal{U}(-10, 10), \quad \Sigma_i \sim \mathcal{W}(2 \times \mathbf{I}_d, 7), \quad \varepsilon \sim \mathcal{N}(0, 0.2 \times \mathbf{I}_d),$$

where $\mathcal{U}(a, b)$ and $\mathcal{W}(\Sigma_0, df)$ represent the uniform distribution and Wishart distribution, respectively. We set $\boldsymbol{\pi} = (0.05, 0.3, 0.4, 0.25)^\top$. The choice of parameters here is quite arbitrary; we have experimented using various parameter settings and the results are similar to those presented here.

Figure 3.7a depicts the comparison between the standard kernel mean estimator and the shrinkage estimator, $\hat{\boldsymbol{\mu}}_\alpha$ when the kernel k is the Gaussian RBF kernel. For shrinkage estimator $\hat{\boldsymbol{\mu}}_\alpha$, we consider $f^* = C \times k(\mathbf{x}, \cdot)$ where C is a scaling factor and each element of $\mathbf{x}$ is a realization of uniform random variable on $(0, 1)$. That is, we allow the target f^* to change depending on the value of C. As the absolute value of C increases, the target function f^* will move further away from the origin. The shrinkage parameter α is determined using the empirical bound, *i.e.*, $\tilde{\alpha} = \hat{\Delta}/(\hat{\Delta} + \|f^* - \hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}})$. As we can see in Figure 3.7a, the results reveal how important the choice of f^* is. That is, we may get substantial improvement over the empirical



(a) **(b)**

Figure 3.7: (a) The risk comparison between $\hat{\boldsymbol{\mu}}$ (KME) and $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}}$ (KMSE) where $\tilde{\alpha} = \hat{\Delta}/(\hat{\Delta} + \|f^* - \hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}})$. We consider when $f^* = C \times k(\mathbf{x}, \cdot)$ where $\mathbf{x}$ is drawn uniformly from a pre-specified range and C is a scaling factor. (b) The probability of improvement and the risk difference as a function of shrinkage parameter α averaged over 1,000 iterations. As the value of α increases, we get more improvement in term of the risk, whereas the probability of improvement decreases as a function of α.

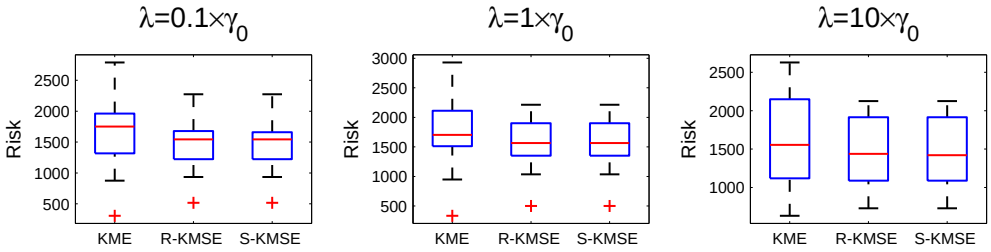
estimator if appropriate prior knowledge is incorporated through f^*, which in this case suggests that f^* should lie close to the origin. We intend to investigate the topic of prior knowledge in more detail in our future work.

Previous comparisons between standard estimator and shrinkage estimator is based entirely on the notion of a risk, which is in fact not useful in practice as we only observe a single copy of sample from the probability distribution. Instead, one should also look at the probability that, given a single copy of sample, the shrinkage estimator outperforms the standard one in term of a loss. To this end, we conduct an experiment comparing the standard estimator and shrinkage estimator using the Gaussian RBF kernel. In addition to the risk comparison, we also compare the probability that the shrinkage estimator gives smaller loss than that of the standard estimator. To be more precise, the probability is defined as a proportion of the samples drawn from the same distribution whose shrinkage loss is smaller than the loss of the standard estimator. Figure 3.7b illustrates the risk difference ($\Delta_\alpha - \Delta$) and the probability of improvement (*i.e.*, the fraction of times $\Delta_\alpha < \Delta$) as a function of shrinkage parameter α. In this case, the value of α is specified as a proportion of empirical upper bound $2\hat{\Delta}/(\hat{\Delta} + \|\hat{\boldsymbol{\mu}}\|^2_{\mathscr{H}})$. The results suggest that the shrinkage parameter α controls the trade-off between the amount of improvement in terms of risk and the probability that the shrinkage estimator will improve upon the standard one. However, this trade-off only holds up to a certain value of α. As α becomes too large, both the probability of improvement and the amount of improvement itself decrease, which coincides with the intuition given for the positive-part shrinkage estimators (cf. Section 3.3.1).

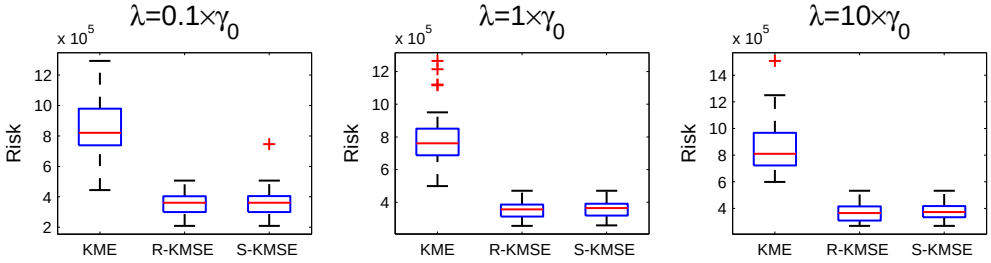
Shrinkage Estimators via Leave-One-Out Cross-Validation

In addition to the empirical upper bound, one can alternatively compute the shrinkage parameter using leave-one-out cross-validation proposed in Section 3.4. Our goal here is to compare the B-KMSE, R-KMSE and S-KMSE on synthetic data when the shrinkage parameter λ is chosen via leave-one-out cross-validation procedure. Note that the only difference between B-KMSE and R-KMSE is the way we compute the shrinkage parameter.

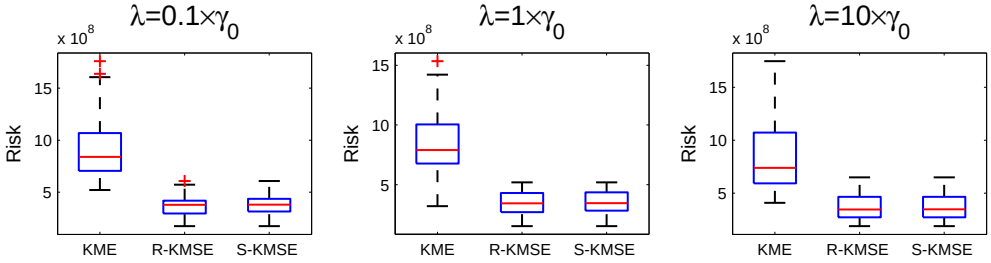
Figure 3.8 shows the empirical risk of different estimators using different kernels as we increase the value of shrinkage parameter λ (note that R-KMSE and S-KMSE in Figure 3.8 refer to those in (3.36) and (3.45) respectively). Here we scale the shrinkage parameter by the



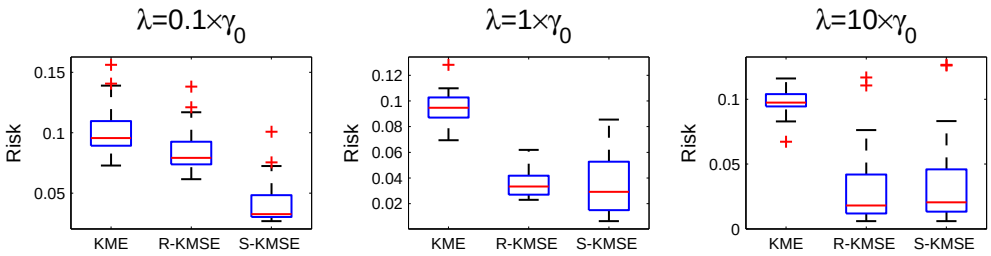
(a) LIN



(b) POLY2



(c) POLY3



(d) RBF

Figure 3.8: The average loss of KME (left), R-KMSE (middle) and S-KMSE (right) estimators with different values of shrinkage parameter. We repeat the experiments over 30 different distributions with $n = 10$ and $d = 30$.

smallest non-zero eigenvalue γ_0 of the kernel matrix $\mathbf{K}$. In general, we find that R-KMSE and S-KMSE outperforms KME. Nevertheless, as the shrinkage parameter λ becomes large, there is a tendency that the specific shrinkage estimate might actually perform worse than the KME, *e.g.*, see LIN kernel and outliers in Figure 3.8. The result also supports our previous observation regarding Figure 3.7b, which suggests that it is very important to choose the parameter λ appropriately.

To demonstrate the leave-one-out cross-validation procedure, we conduct similar experiments in which the parameter λ is chosen by the proposed LOOCV procedure. Figure 3.9

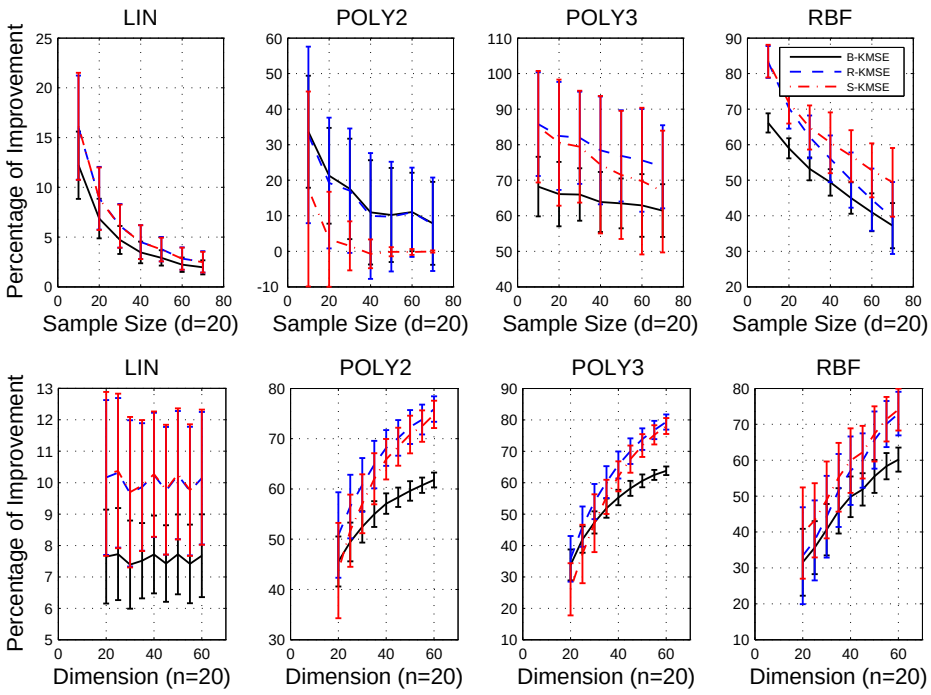


Figure 3.9: The percentage of improvement compared to KME over 30 different distributions of B-KMSE, R-KMSE and S-KMSE with varying sample size (n) and dimension (d). For B-KMSE, we calculate α using (3.20), whereas R-KMSE and S-KMSE use LOOCV to choose λ.

depicts the percentage of improvement (w.r.t. the empirical risk of the KME[2]) as we vary the sample size and dimension of the data. Clearly, B-KMSE, R-KMSE and S-KMSE outperform the standard estimator. Moreover, both R-KMSE and S-KMSE tend to outperform the B-KMSE. We can also see that the performance of S-KMSE depends on the choice of kernel. This makes sense intuitively because S-KMSE also incorporates the eigen-spectrum of $\mathbf{K}$, whereas R-KMSE does not. The effects of both sample size and data dimensionality are also transparent from Figure 3.9. While it is intuitive to see that the improvement gets smaller with increase in sample size, it is a bit surprising to see that we can gain much more in high-dimensional input space, especially when the kernel function is non-linear, because the estimation happens in the feature space associated with the kernel function rather than in the input space. Lastly, we note that the improvement is more substantial in the "large d, small n" paradigm.

3.7.2 Real Data

To evaluate the proposed estimators on real-world data, we consider several benchmark applications, namely, classification via Parzen window classifier, density estimation via kernel mean matching (Song et al. 2008), and discriminative learning on distributions (Muandet et al. 2012, Muandet and Schölkopf 2013). For some of these tasks we employ datasets from the UCI repositories. We use only real-valued features, each of which is normalized to have zero mean and unit variance.

[2]If we denote the loss of KME and KMSE as ℓ_{KME} and ℓ_{KMSE}, respectively, the percentage of improvement is calculated as $100 \times (\ell_{\text{KME}} - \ell_{\text{KMSE}})/\ell_{\text{KME}}$.

Table 3.2: The classification error rate of Parzen window classifier via different kernel mean estimators. The boldface represents the result whose difference from the baseline, *i.e.*, KME, is statistically significant.

Dataset	Classification Error Rate			
	KME	B-KMSE	R-KMSE	S-KMSE
Climate Model	0.0348±0.0118	0.0348±0.0118	0.0348±0.0118	0.0348±0.0118
Ionosphere	0.2873±0.0343	**0.2768±0.0359**	**0.2749±0.0341**	0.2800±0.0367
Parkinsons	0.1318±0.0441	0.1250±0.0366	**0.1157±0.0395**	0.1309±0.0396
Pima	0.2951±0.0462	0.2921±0.0442	0.2937±0.0458	0.2943±0.0471
SPECTF	0.2583±0.0829	0.2597±0.0817	**0.2263±0.0626**	0.2417±0.0651
Iris	0.1079±0.0379	0.1071±0.0389	0.1055±0.0389	0.1040±0.0383
Wine	0.1301±0.0381	**0.1183±0.0445**	**0.1161±0.0414**	**0.1183±0.0431**

Parzen Window Classifiers

One of the oldest and best-known classification algorithms is the *Parzen window classifier* (**?**). It is easy to implement and is one of the powerful non-linear supervised learning techniques. Suppose we have data points from two classes, namely, positive class and negative class. For positive class, we observe $\mathfrak{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathcal{X}$, while for negative class we have $\mathfrak{Y} \triangleq \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m\} \subset \mathcal{X}$. The Parzen window classifier is given by

$$f(\mathbf{z}) = \operatorname{sgn}\left(\frac{1}{n}\sum_{i=1}^{n} k(\mathbf{z}, \mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} k(\mathbf{z}, \mathbf{y}_j) + b\right) = \operatorname{sgn}\left(\hat{\boldsymbol{\mu}}_{\mathfrak{X}}(\mathbf{z}) - \hat{\boldsymbol{\mu}}_{\mathfrak{Y}}(\mathbf{z}) + b\right), \quad (3.59)$$

where b is a bias term given by $b = \frac{1}{2}(\|\hat{\boldsymbol{\mu}}_{\mathfrak{Y}}\|^2_{\mathscr{H}} - \|\hat{\boldsymbol{\mu}}_{\mathfrak{X}}\|^2_{\mathscr{H}})$. This algorithm is often referred to as the lazy algorithm as it does not require training.

In brief, the classifier (3.59) assigns the data point $\mathbf{z}$ to the class whose empirical kernel mean $\hat{\boldsymbol{\mu}}$ is closer to the feature map $k(\mathbf{z}, \cdot)$ of the data point in the RKHS. On the other hand, we may view the empirical kernel mean $\hat{\boldsymbol{\mu}}_{\mathfrak{X}} \triangleq \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \cdot)$ (resp. $\hat{\boldsymbol{\mu}}_{\mathfrak{Y}} \triangleq \frac{1}{m}\sum_{j=1}^{m} k(\mathbf{y}_j, \cdot)$) as a standard empirical estimate, *i.e.*, KME, of the true kernel mean representation of the class-conditional distribution $\mathbb{P}(X|Y = +1)$ (resp. $\mathbb{P}(X|Y = -1)$). Given the improvement of shrinkage estimators over the empirical estimator of kernel mean, it is natural to expect that the performance of Parzen window classifier can be improved by employing shrinkage estimators of the true mean representation.

Our goal in this experiment is to compare the performance of Parzen window classifier using different kernel mean estimators. That is, we replace $\hat{\boldsymbol{\mu}}_{\mathfrak{X}}$ and $\hat{\boldsymbol{\mu}}_{\mathfrak{Y}}$ by their shrinkage counterparts and evaluate the resulting classifiers across several datasets taken from the UCI machine learning repository. In this experiment, we only consider the Gaussian RBF kernel whose bandwidth parameter is chosen by cross-validation procedure over a uniform grid $\sigma \in [0.1, 2]$. We use 30% of each dataset as a test set and the rest as a training set. We employ a simple pairwise coupling and majority vote for multi-class classification. We repeat the experiments 100 times and perform the paired-sample t-test on the results at 5% significance level. Table 3.2 reports the classification error rates of the Parzen window classifiers with different kernel mean estimators. Although the improvement is not substantial, we can see that the shrinkage estimators consistently give better performance than the standard estimator.

Density Estimation

We perform density estimation via kernel mean matching (Song et al. 2008), wherein we fit the density $Q = \sum_{j=1}^{m} \pi_j \mathcal{N}(\boldsymbol{\theta}_j, \boldsymbol{\sigma}_j^2 \mathbf{I})$ to each dataset by the following minimization problem:

$$\min_{\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\sigma}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_Q\|_{\mathscr{H}}^2 \quad \text{subject to} \quad \sum_{j=1}^{m} \pi_j = 1,\ \pi_j \geq 0\,. \tag{3.60}$$

The empirical mean map $\hat{\boldsymbol{\mu}}$ is obtained from samples using different estimators, whereas $\boldsymbol{\mu}_Q$ is the kernel mean embedding of the density Q. Unlike experiments in Song et al. (2008), our goal is to compare different estimators of $\boldsymbol{\mu}_{\mathbb{P}}$ (where $\mathbb{P}$ is the true data distribution), by replacing $\hat{\boldsymbol{\mu}}$ in (3.60) with different shrinkage estimators. A better estimate of $\boldsymbol{\mu}_{\mathbb{P}}$ should lead to better density estimation, as measured by the negative log-likelihood of Q on the test set, which we choose to be 30% of the dataset. For each dataset, we set the number of mixture components m to be 10. The model is initialized by running 50 random initializations using the k-means algorithm and returning the best. We repeat the experiments 30 times and perform the paired sign test on the results at 5% significance level.[3]

The average negative log-likelihood of the model Q, optimized via different estimators, is reported in Table 3.3. In most cases, both R-KMSE and S-KMSE consistently achieve smaller negative log-likelihood when compared to KME. B-KMSE also tends to outperform the KME. However, in few cases the KMSEs achieve larger negative log-likelihood, especially when we use linear and degree-2 polynomial kernels. This highlight the potential of our estimators in a non-linear setting.

Discriminative Learning on Probability Distributions

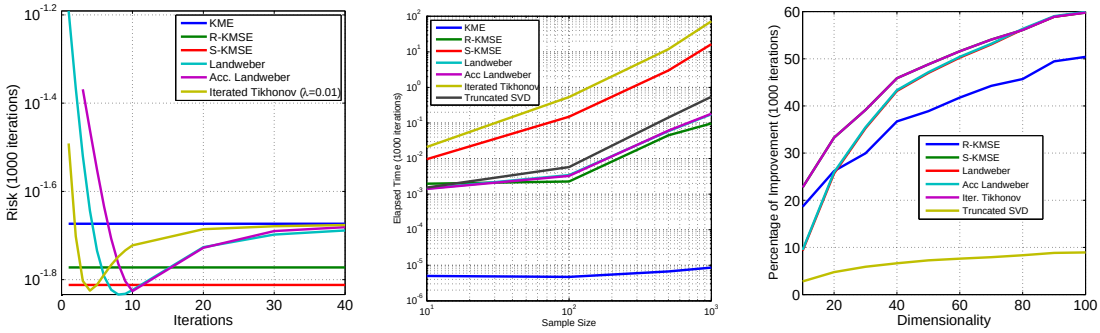
The last experiment involves the discriminative learning on a collection of probability distributions via the kernel mean representation. A positive semi-definite kernel between distributions can be defined via their kernel mean embeddings. That is, given a training sample $(\widehat{\mathbb{P}}_1, y_1), \ldots, (\widehat{\mathbb{P}}_m, y_m) \in \mathscr{P} \times \{-1, +1\}$ where $\widehat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{p=1}^{n_i} \delta_{\mathbf{x}_p^i}$ and $\mathbf{x}_p^i \sim \mathbb{P}_i$, the linear kernel between two distributions is approximated by

$$\langle \hat{\boldsymbol{\mu}}_{\mathbb{P}_i}, \hat{\boldsymbol{\mu}}_{\mathbb{P}_j} \rangle_{\mathscr{H}} = \left\langle \sum_{p=1}^{n_i} \beta_p^i \phi(\mathbf{x}_p^i), \sum_{q=1}^{n_j} \beta_q^j \phi(\mathbf{x}_q^j) \right\rangle_{\mathscr{H}} = \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \beta_p^i \beta_q^j k(\mathbf{x}_p^i, \mathbf{x}_q^j),$$

where the weight vectors $\boldsymbol{\beta}^i$ and $\boldsymbol{\beta}^j$ come from the kernel mean estimates of $\boldsymbol{\mu}_{\mathbb{P}_i}$ and $\boldsymbol{\mu}_{\mathbb{P}_j}$, respectively. The non-linear kernel can then be defined accordingly, *e.g.*, $\kappa(\mathbb{P}_i, \mathbb{P}_j) = \exp(\|\hat{\boldsymbol{\mu}}_{\mathbb{P}_i} - \hat{\boldsymbol{\mu}}_{\mathbb{P}_j}\|_{\mathscr{H}}^2 / 2\sigma^2)$. Our goal in this experiment is to investigate if the shrinkage estimators of the kernel mean improve the performance of discriminative learning on distributions. To this end, we conduct experiments on natural scene categorization using support measure machine (SMM) (Muandet et al. 2012) and group anomaly detection on a high-energy physics dataset using one-class SMM (OCSMM) (Muandet and Schölkopf 2013). We use both linear and non-linear kernels where the Gaussian RBF kernel is employed as an embedding kernel (Muandet et al. 2012). All hyper-parameters are chosen by 10-fold cross-validation.[4] For our unsupervised problem, we repeat the experiments using several parameter settings and report the best results. Table 3.4 reports the classification accuracy of SMM and the area under ROC curve (AUC) of

[3]The paired sign test is a nonparametric test that can be used to examine whether two paired samples have the same distribution. In our case, we compare B-KMSE, R-KMSE and S-KMSE against KME.

[4]In principle one can incorporate the shrinkage parameter into the cross-validation procedure. In this work we are only interested in the value of λ returned by the proposed LOOCV procedure.



(a) risk vs. iteration **(b)** runtime vs. sample size **(c)** risk vs. dimension

Figure 3.10: (a) For iterative algorithms, the number of iterations acts as shrinkage parameter. (b) The iterative algorithms such as Landweber and accelerated Landweber are more efficient than the S-KMSE. (c) A percentage of improvement w.r.t. the KME, i.e., $100 \times (R - R_\lambda)/R$ where R and R_λ denote the approximated risk of KME and KMSE, respectively. Most Spectral-KMSE algorithms outperform R-KMSE which does not take into account the geometric information of the RKHS.

OCSMM using different kernel mean estimators. All shrinkage estimators consistently lead to better performance on both SMM and OCSMM when compared to KME.

In summary, the proposed shrinkage estimators outperform the standard KME. While B-KMSE and R-KMSE are very competitive compared to KME, S-KMSE tends to outperform both B-KMSE and R-KMSE, however, sometimes leading to poor estimates depending on the dataset and the kernel function.

3.7.3 Comparison of Filter Functions

The main objective of our empirical studies in this section is to compare different filter functions for the Spectral-KMSE.

Synthetic data. Given the i.i.d. sample $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ from $\mathbb{P}$ where $\mathbf{x}_i \in \mathbb{R}^d$, we evaluate different estimators using the loss function (3.58). The risk of the estimator is subsequently approximated by averaging over m independent copies of $\mathbf{X}$. In this experiment, we set $n = 50$, $d = 20$, and $m = 1000$. Throughout, we use the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$ whose bandwidth parameter is calculated using the median heuristic. To allow for an analytic calculation of the loss $L(\boldsymbol{\beta}, \mathbf{X}, \mathbb{P})$, we assume that the distribution $\mathbb{P}$ is a d-dimensional mixture of Gaussians. Specifically, the data are generated as follows: $\mathbf{x} \sim \sum_{i=1}^4 \pi_i \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i) + \varepsilon, \theta_{ij} \sim \mathcal{U}(-10, 10), \Sigma_i \sim \mathcal{W}(3 \times \mathbf{I}_d, 7), \varepsilon \sim \mathcal{N}(0, 0.2 \times \mathbf{I}_d)$ where $\mathcal{U}(a, b)$ and $\mathcal{W}(\Sigma_0, df)$ are the uniform distribution and Wishart distribution, respectively. We set $\boldsymbol{\pi} = [0.05, 0.3, 0.4, 0.25]$.

A natural approach for choosing λ is cross-validation procedure, which can be performed efficiently for the iterative methods such as Landweber and accelerated Landweber. For these two algorithms, we evaluate the leave-one-out score and select $\boldsymbol{\beta}^t$ at the iteration t that minimizes this score (see, e.g., Figure 3.10a). Note that these methods have the built-in property of computing the whole *regularization path* efficiently. Since each iteration of the iterated Tikhonov is in fact equivalent to the S-KMSE, we assume $t = 3$ for simplicity and use the efficient LOOCV procedure proposed earlier to find λ at each iteration. Lastly, the truncation limit of TSVD can be identified efficiently by mean of generalized cross-validation (GCV) procedure (Golub et al. 1979). To allow for an efficient calculation of GCV score, we resort to the alternative loss function $\mathcal{L}(\boldsymbol{\beta}) := \|\mathbf{K}\boldsymbol{\beta} - \mathbf{K}\mathbf{1}_n\|_2^2$.

Table 3.3: Average negative log-likelihood of the model Q on test points over 30 randomizations. The boldface represents the result whose difference from the baseline, *i.e.*, KME, is statistically significant.

Dataset	LIN				POLY2				POLY3				RBF			
	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE
1. ionosphere	39.878	40.038	39.859	39.823	34.651	**34.352**	34.390	**34.009**	35.943	**35.575**	35.543	**34.617**	41.601	40.976	**40.817**	41.229
2. sonar	72.240	**72.044**	72.198	**72.157**	100.420	**99.573**	**97.844**	**97.783**	72.294	**71.933**	72.003	71.835	98.540	**95.815**	**93.458**	**93.010**
3. Australian	18.277	18.280	18.294	18.293	18.357	18.381	18.391	18.429	18.611	18.463	18.466	18.495	19.428	19.325	19.418	19.393
4. specft	57.444	**57.2808**	57.218	**57.224**	67.018	66.979	**66.431**	**66.391**	59.585	**58.969**	60.006	60.616	65.674	65.138	65.039	**64.699**
5. wdbc	31.801	31.759	31.776	31.781	32.421	32.310	32.373	32.316	31.183	**31.167**	**31.127**	31.110	36.471	36.453	36.335	35.898
6. wine	16.019	16.000	16.039	16.009	17.070	16.920	**16.886**	16.960	16.393	16.300	16.309	16.202	17.569	17.546	17.498	17.498
7. satimage	25.258	25.317	25.219	25.186	24.214	24.111	24.132	24.259	25.284	25.276	25.239	25.263	23.741	23.753	23.728	**24.384**
8. segment	18.326	**17.868**	18.055	**18.124**	18.571	18.292	18.277	18.631	19.642	19.549	19.404	19.628	21.946	**21.598**	**21.580**	**21.822**
9. vehicle	16.633	16.519	16.521	16.499	16.096	**15.998**	16.031	16.041	16.288	16.278	**16.281**	**16.263**	18.260	18.056	18.119	**17.911**
10. svmguide2	27.298	27.273	27.281	27.276	27.812	**28.030**	27.985	**27.975**	28.014	**28.177**	**28.321**	28.250	28.132	28.122	28.119	28.020
11. vowel	12.632	12.626	12.629	12.656	12.532	12.471	12.479	12.472	13.069	13.061	13.056	**13.054**	13.526	13.486	13.462	**13.453**
12. housing	14.637	14.441	14.469	**14.296**	15.543	15.467	15.414	15.390	15.592	**15.543**	**15.509**	**15.408**	16.487	16.239	16.424	**16.019**
13. bodyfat	17.527	17.362	**17.348**	17.396	17.386	17.358	17.356	17.329	16.418	16.393	**16.305**	**16.194**	17.875	17.652	17.607	**17.651**
14. abalone	5.706	5.665	5.708	5.722	7.281	7.116	7.185	7.025	5.864	5.847	5.853	5.832	6.068	6.039	6.049	5.910
15. glass	9.245	9.211	9.198	9.217	8.571	8.473	8.457	8.414	9.050	8.991	9.012	**8.737**	9.606	**9.605**	**9.575**	**9.573**

Table 3.4: The classification accuracy of SMM and the area under ROC curve (AUC) of OCSMM using different estimators to construct the kernel on distributions.

Estimator	Linear Kernel		Non-linear Kernel	
	SMM	OCSMM	SMM	OCSMM
KME	0.5432	0.6955	0.6017	0.9085
B-KMSE	0.5455	0.6964	0.6106	0.9088
R-KMSE	0.5521	0.6970	0.6303	0.9105
S-KMSE	0.5606	0.6970	0.6412	0.9063

Figure 3.10 reveals interesting aspects of the Spectral-KMSE. Firstly, as we can see in Figure 3.10a, the number of iterations acts as shrinkage parameter whose optimal value can be attained within just a few iterations. Moreover, these methods do not suffer from "over-shrinking" because $\lambda \to 0$ as $t \to \infty$. In other words, if the chosen t happens to be too large, the worst we can get is the standard empirical estimator. Secondly, Figure 3.10b demonstrates that both Landweber and accelerated Landweber are more computationally efficient than the S-KMSE. Lastly, Figure 3.10c suggests that the improvement of shrinkage estimators becomes increasingly remarkable in a high-dimensional setting. Interestingly, we can observe that most Spectral-KMSE algorithms outperform the R-KMSE, which supports our hypothesis on the importance of the geometric information of RKHS mentioned in Section 3.4.2. In addition, although the TSVD still gain from shrinkage, the improvement is smaller than other algorithms. This highlights the importance of filter functions and associated parameters.

Real data. We apply Spectral-KMSE to the density estimation problem via kernel mean matching (Song et al. 2008). The datasets were taken from the UCI repository[5] and pre-processed by standardizing each feature. Then, we fit a mixture model $Q = \sum_{j=1}^{r} \pi_j \mathcal{N}(\boldsymbol{\theta}_j, \sigma_j^2 \mathbf{I})$ to the pre-processed dataset $\mathbf{X} := \{x_i\}_{i=1}^{n}$ by minimizing $\|\boldsymbol{\mu}_Q - \hat{\boldsymbol{\mu}}_X\|^2$ subject to the constraint $\sum_{j=1}^{r} \pi_j = 1$. Here $\boldsymbol{\mu}_Q$ is the mean embedding of the mixture model Q and $\hat{\boldsymbol{\mu}}_X$ is the empirical mean embedding obtained from $\mathbf{X}$. Based on different estimators of $\boldsymbol{\mu}_X$, we evaluate the resultant model Q by the negative log-likelihood score on the test data. The parameters $(\pi_j, \boldsymbol{\theta}_j, \sigma_j^2)$ are initialized by the best one obtained from the K-means algorithm with 50 initializations. Throughout, we set $r = 5$ and use 25% of each dataset as a test set.

Table 3.5: The average negative log-likelihood evaluated on the test set. The results are obtained from 30 repetitions of the experiment. The boldface represents the statistically significant results.

Dataset	KME	R-KMSE	S-KMSE	Landweber	Acc Land	Iter Tik	TSVD
ionosphere	36.1769	36.1402	36.1622	**36.1204**	**36.1554**	36.1334	36.1442
glass	10.7855	10.7403	10.7448	**10.7099**	10.7541	10.9078	10.7791
bodyfat	18.1964	18.1158	18.1810	18.1607	18.1941	18.1267	18.1061
housing	14.3016	**14.2195**	**14.0409**	14.2499	14.1983	**14.2868**	**14.3129**
vowel	13.9253	13.8426	13.8817	13.8337	14.1368	13.8633	13.8375
svmguide2	28.1091	**28.0546**	**27.9640**	28.1052	**27.9693**	**28.0417**	28.1128
vehicle	18.5295	18.3693	**18.2547**	**18.4873**	**18.3124**	**18.4128**	18.3910
wine	16.7668	**16.7548**	**16.7457**	**16.7596**	**16.6790**	16.6954	**16.5719**
wdbc	35.1916	**35.1814**	**35.0023**	**35.1402**	35.1366	35.1881	**35.1850**

Table 3.5 reports the results on real data. In general, the mixture model Q obtained from the proposed shrinkage estimators tend to achieve lower negative log-likelihood score than that obtained from the standard empirical estimator. Moreover, we can observe that the relative performance of different filter functions vary across datasets, suggesting that, in addition to potential gain from shrinkage, incorporating prior knowledge through the choice of filter function could lead to further improvement.

3.8 Discussions

Motivated by the classical James-Stein phenomenon, we proposed a shrinkage estimator for the kernel mean $\boldsymbol{\mu}$ in an RKHS $\mathscr{H}$ and showed they improve upon the empirical estimator $\hat{\boldsymbol{\mu}}$ in the mean squared sense. We showed the proposed shrinkage estimator $\tilde{\boldsymbol{\mu}}$ (with the

[5] http://archive.ics.uci.edu/ml/

shrinkage parameter being learned from data) to be $\sqrt{n}$-consistent and satisfies $\mathbb{E}\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2_{\mathscr{H}} < \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2_{\mathscr{H}} + O(n^{-3/2})$ as $n \to \infty$. We also provided a regularization interpretation to shrinkage estimation, using which we also presented two shrinkage estimators, namely regularized shrinkage estimator and spectral shrinkage estimator, wherein the first one is closely related to $\tilde{\boldsymbol{\mu}}$ while the latter exploits the spectral decay of the covariance operator in $\mathscr{H}$. We showed through numerical experiments that the proposed estimators outperform the empirical estimator in various scenarios. Most importantly, the shrinkage estimators not only provide more accurate estimation, but also lead to superior performance on many real-world applications.

In this chapter, while we focused mainly on an estimation of the mean function in RKHS, it is quite straightforward to extend the shrinkage idea to estimate covariance (and cross-covariance) operators and tensors in RKHS. The key observation is that the covariance operator can be viewed as a mean function in a tensor RKHS. Covariance operators in RKHS are ubiquitous in many machine learning algorithms such as kernel PCA, kernel FDA, and kernel CCA. To this end, we carried out a preliminary investigation on extending the shrinkage idea to estimate covariance (and cross-covariance) operators and present below some numerical results that demonstrate the performance of the corresponding shrinkage estimator.

Let $(\mathscr{H}, k)$ and $(\mathscr{F}, l)$ be RKHS of functions on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, with reproducing kernels k and l, respectively. We consider a random vector $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with distribution $\mathbb{P}_{XY}$. The marginal distributions of X and Y are denoted by $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. If $\mathbb{E}_X k(X, X) < \infty$ and $\mathbb{E}_Y l(Y, Y) < \infty$, then there exists a unique *cross-covariance operator* $\mathbf{C}_{YX} : \mathscr{H} \to \mathscr{F}$ such that

$$\langle g, \mathbf{C}_{YX} f \rangle_{\mathscr{F}} = \mathbb{E}_{XY}[(f(X) - \mathbb{E}_X[f(X)])(g(Y) - \mathbb{E}_Y[g(Y)])] = Cov(f(X), g(Y))$$

holds for all $f \in \mathscr{H}$ and $g \in \mathscr{F}$ (see Fukumizu et al. (2004)). If X is equal to Y, we obtain the self-adjoint operator $\mathbf{C}_{XX}$ called the *covariance operator*. Given an i.i.d sample $((\mathbf{x}_i, \mathbf{y}_i))_{i=1}^n$ from $\mathbb{P}_{XY}$, we can write the empirical cross-covariance operator $\widehat{\mathbf{C}}_{YX}$ as

$$\widehat{\mathbf{C}}_{YX} \triangleq \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \varphi(\mathbf{y}_i) - \hat{\boldsymbol{\mu}}_X \otimes \hat{\boldsymbol{\mu}}_Y \tag{3.61}$$

where $\hat{\boldsymbol{\mu}}_X = \frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i)$, $\hat{\boldsymbol{\mu}}_Y = \frac{1}{n}\sum_{i=1}^n \varphi(\mathbf{y}_i)$, $\phi(\mathbf{x}_i) := k(\cdot, \mathbf{x}_i)$ and $\varphi(\mathbf{y}_i) := l(\cdot, \mathbf{y}_i)$. Let $\tilde{\phi}$ and $\tilde{\varphi}$ be the centered version of the feature map ϕ and φ defined as $\tilde{\phi}(\mathbf{x}) = \phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_X$ and $\tilde{\varphi}(\mathbf{y}) = \varphi(\mathbf{y}) - \hat{\boldsymbol{\mu}}_Y$, respectively. Then, the empirical cross-covariance operator in (3.61) can be rewritten as

$$\widehat{\mathbf{C}}_{YX} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(\mathbf{x}_i) \otimes \tilde{\varphi}(\mathbf{y}_i),$$

and therefore a shrinkage estimator of $\mathbf{C}_{YX}$ (*e.g.*, an equivalent of B-KMSE) can be constructed based on the ideas presented in this paper. We will call this estimator a *covariance-operator shrinkage estimator* (COSE). The same trick can be easily generalized to tensors of higher order, which have been previously used, for example, in Song et al. (2011b).

In our empirical studies, we perform KPCA using different estimates of the mean and covariance operators. We compare the reconstruction error $\mathcal{E}_{proj}(\mathbf{z}) = \|\phi(\mathbf{z}) - \mathbf{P}\phi(\mathbf{z})\|^2$ on test samples where $\mathbf{P}$ is the projection constructed from the first 20 principal components. We use a Gaussian RBF kernel for all datasets and compare 5 different scenarios: *i*) standard KPCA *ii*) shrinkage centering with R-KMSE *iii*) shrinkage centering with S-KMSE *iv*) KPCA with R-COSE *v*) KPCA with S-COSE. Given the similarity between B-KMSE and R-KMSE, we omit the B-KMSE in this experiment for ease of analysis. To perform KPCA on shrinkage covariance operator, we solve the generalized eigenvalue problem $\mathbf{K}_c\mathbf{B}\mathbf{K}_c\mathbf{V} = \mathbf{K}_c\mathbf{V}\mathbf{D}$ where

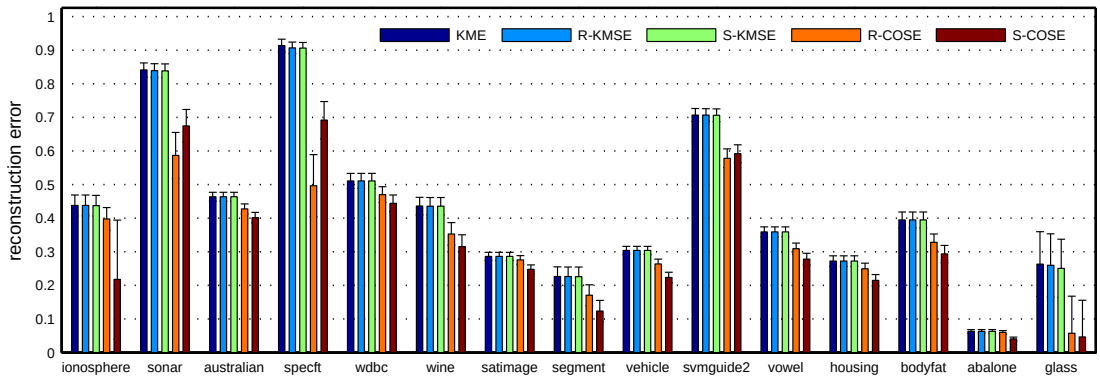


Figure 3.11: The average reconstruction error of KPCA on hold-out test samples over 100 repetitions. KME represents the standard approach, whereas R-KMSE and S-KMSE use shrinkage means to perform centering. R-COSE and S-COSE directly use the shrinkage estimate of the covariance operator.

$\mathbf{B} = \mathrm{diag}(\boldsymbol{\beta})$ and $\mathbf{K}_c$ is the centered Gram matrix. The weight vector $\boldsymbol{\beta}$ is obtained from shrinkage estimators using the kernel matrix $\mathbf{K}_c \circ \mathbf{K}_c$ where $\circ$ denotes the Hadamard product. We use 30% of the dataset as a test set.

Figure 3.11 illustrates the results of KPCA. Clearly, R-COSE and S-COSE consistently outperform all the other estimators. Although we observe an improvement of R-KMSE and S-KMSE over KME, it is very small compared to that of R-COSE and S-COSE. Intuitively, this makes sense as changing the mean point or shifting data does not change the covariance structure considerably, so it will not significantly affect the reconstruction error. In summary, the results are very encouraging and we intend to pursue this aspect further in our future works.

END OF CHAPTER 3

Chapter 4

Supervised Learning on Distributions

Owing to kernel mean embedding of distributions and its estimators presented in previous chapter, this chapter generalizes the standard supervised learning framework on points to probability measures and provides theoretical insights. I will also discuss connections to existing frameworks. Lastly, I will give some suggestions for future research.

4.1 Introduction

Discriminative learning algorithms are typically trained from large collections of vectorial training examples. In many classical learning problems, however, it is arguably more appropriate to represent training data not as individual data points, but as probability distributions. There are, in fact, multiple reasons why probability distributions may be preferable.

Firstly, uncertain or missing data naturally arises in many applications. For example, gene expression data obtained from the microarray experiments are known to be very noisy due to various sources of variabilities (Yang and Speed 2002). In order to reduce uncertainty, and to allow for estimates of confidence levels, experiments are often replicated. Unfortunately, the feasibility of replicating the microarray experiments is often inhibited by cost constraints, as well as the amount of available mRNA. To cope with experimental uncertainty given a limited amount of data, it is natural to represent each array as a probability distribution that has been designed to approximate the variability of gene expressions across slides.

Likewise, it may impossible or very difficult to make an exact measurement of certain properties of the objects such as spectra of celestrial objects. Hence, one has to rely on the measurement possessing substantial and heterogenous uncertainty. For instance, SDSS-III's Baryon Oscillation Spectroscopic Survey (BOSS) aims to obtain spectra for about 160,000 quasars in the $2.2 \leq z \leq 3.5$ redshift range, which are important tools for studying the intervening intergalactic medium and the angular diameter distance of the universe. Unfortunately, quasar target selection in this redshift range possesses several challenges mainly because of stella loci contamination and substantial photometric uncertainties. Incorporating the photometric uncertainty has been shown to improve the efficiency of the quasar target selection (Kirkpatrick et al. 2011, Bovy et al. 2011, Ross et al. 2012). Probability distributions are natural representation of the uncertain data.

Secondly, many application domains call for methods that can deal with complex structured data such as DNA sequences, text documents, and graphs. While many available methods rely on kernels defined over such structured data, it is often easier to capture the structure of complex objects with generative models than directly with kernels. For example, in natural language processing, a text document is often modeled as a distribution over topics comprising of words in a dictionary (Blei et al. 2001). In bioinformatics, hidden Markov models (HMMs) are often employed as the basis for methods used in biological sequence analysis. Thus, learning directly from distributions defined by such generative models alleviates the complexity in designing the kernels and can also give a new insight into the structure of data.

Probability distributions may be equally appropriate given an abundance of training data. In data-rich disciplines such as neuroinformatics, climate informatics, and astronomy, a high throughput experiment can easily generate a huge amount of data, leading to significant computational challenges in both time and space. Instead of scaling up one's learning algorithms, one can scale down one's dataset by constructing a smaller collection of distributions which represents groups of similar samples. Besides computational efficiency, aggregate statistics can potentially incorporate higher-level information that represents the collective behavior of multiple data points.

4.2 Related Works

Several attempts have previously been made to learn from distributions by creating positive definite kernels on probability measures. In Jebara et al. (2004b), the probability product kernel (PPK) was proposed as a generalized inner product between two input objects, which 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 nonextensive 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 $K(\mathbb{P}, \mathbb{Q}) = \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ considered in this thesis is in fact a special case of the Hilbertian metric (Hein and Bousquet 2005), with the associated kernel $K(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}, \tilde{\mathbf{x}}\sim\mathbb{Q}}[k(\mathbf{x}, \tilde{\mathbf{x}})]$, 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. That is, a probabilistic model $\hat{p}_{\mathbf{x}}$ is learned for each example and used as a surrogate to construct the kernel between those examples. The idea of surrogate kernels has also been adopted by the PPK (Jebara et al. 2004b). In this case, we have $K_\rho(p, p') = \int_{\mathcal{X}} p(\mathbf{x})^\rho p'(\mathbf{x})^\rho \, \mathrm{d}\mathbf{x}$, which has been shown to be a special case of GMMK when $\rho = 1$ (Mehta and Gray 2010). Consequently, GMMK, PPK with $\rho = 1$, and linear kernel $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ are equivalent when the embedding kernel is $k(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$. More 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. 2011a). In contrast, we treat the probability distributions in a supervised way (cf. the regularized functional (4.10)) and the kernel is not restricted to only the empirical kernel.

The use of expected kernels in dealing with the uncertainty in the input data has a connection to robust SVMs. For instance, 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 chapter, 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).

A major drawback of the previous works is that they usually impose a strong *parametric assumption* on the form of probability distribution. As we will see, the kernel mean representation allows one to learn directly from distributions without making such an assumption. For example, Szabó et al. (2015) has recently studied the *nonparametric* distributions regression problem based on kernel mean embedding (Smola et al. 2007) and the kernel ridge regression algorithm (cf. Section 2.2.3). 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. 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 regression 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.

4.3 Learning with Empirical Risk Minimization

I will first give a basic idea of empirical risk minimization (ERM) for supervised learning (Vapnik 1992) and then generalize it to a space of probability distributions. For simplicity, I will focus on binary classification problem, although many of the main features can be generalized to multiclass classification and regression problem.

The problem of learning is to choose from a given set of functions $\mathcal{F}$ the one that minimizes a pre-specified risk functional. We assume that there is a joint probability distribution $\mathbb{P}(\mathbf{x}, y)$ over X and Y. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be an arbitrary loss function which measures the discrepancy between the true response y and the response $f(\mathbf{x})$ provided by the learning machine. Our goal is to find $f \in \mathcal{F}$ that minimizes the risk functional

$$R(f) = \int \ell(y, f(\mathbf{x}))\, \mathrm{d}\mathbb{P}(\mathbf{x}, y) \tag{4.1}$$

We denote by f^* a function for which (4.1) is minimal. Since $\mathbb{P}(\mathbf{x}, y)$ is unknown in practice, we cannot evaluate (4.1). Based on a training set of n i.i.d. observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ distributed according to some unknown distribution $\mathbb{P}(X, Y)$, the ERM aims to minimize

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)). \tag{4.2}$$

The principle of ERM states that the learning algorithm should choose $\hat{f}$ which minimizes the empirical risk: $\hat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}(f)$.

A loss function commonly used in the theory of classification is the 0-1 loss function: $\ell(y, \hat{y}) = \mathbb{1}_{\{y \neq \hat{y}\}}$ where $\mathbb{1}$ is an indicator function. In practice, it is difficult to optimize with 0-1 loss as it is an NP-hard problem, so several surrogate losses have been proposed. For instance, common loss functions include *hinge loss* $\ell(y, \hat{y}) = \max(0, 1 + y\hat{y})$ used in SVM, the *exponential loss* $\ell(y, \hat{y}) = \exp(-y\hat{y})$ used in AdaBoost, and the *logistic loss* $\ell(y, \hat{y}) = \log_2(1 + e^{-y\hat{y}})$ used in logistic regression. A typical loss function for regression is a square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$.

Note that one can rewrite (4.2) as follows:

$$\widehat{R}(f) = \int \ell(y, f(\mathbf{x}))\, \mathrm{d}\hat{\mathbb{P}}_n(\mathbf{x}, y), \tag{4.3}$$

where $\hat{\mathbb{P}}_n(\mathbf{x}, y) := \frac{1}{n} \sum_{i=1}^{n} \delta_{(\mathbf{x}_i, y_i)}$ represents an empirical distribution contructed from the sample. Owing to this observation, Chapelle et al. (2000) proposes a *vicinal risk minimization* (VRM) framework in which the Dirac measure $\delta_{(\mathbf{x}_i, y_i)}$ is replaced by a more general probability distribution. On the other hand, the approach proposed in this chapter extends the ERM in a different way.

4.4 Distributional Risk Minimization

Given a non-empty set $\mathcal{X}$, let $\mathscr{P}$ denote the set of all probability measures $\mathbb{P}$ on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$. The goal of *distributional risk minimization* (DRM) is to learn a function $h : \mathscr{P} \to \mathcal{Y}$ given a set of example pairs $\{(\mathbb{P}_i, y_i)\}_{i=1}^n$, where $\mathbb{P}_i \in \mathscr{P}$ and $y_i \in \mathcal{Y}$. In other words, we consider a supervised setting in which input training examples are probability distributions. Throughout this chaper, I focus on the binary classification problem, *i.e.*, $\mathcal{Y} = \{+1, -1\}$.

For a function class $\mathcal{F}$, the DRM minimizes the following loss functional

$$\widehat{R}(f) = \ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_n, y_n, \mathbb{E}_{\mathbb{P}_n}[f]\right) + \Omega\left(\|f\|_{\mathcal{F}}\right) \tag{4.4}$$

where $\ell(\cdot)$ is the loss functional and $\Omega(\cdot)$ is a monotonically increasing regularization functional. One example of (4.4) is when $\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_n, y_n, \mathbb{E}_{\mathbb{P}_n}[f]\right) = \sum_{i=1}^{n} \ell(y_i, \mathbb{E}_{\mathbb{P}_i}[f])$. Note that if we substitute the distributions $\mathbb{P}_i$ in (4.4) by Dirac measures $\delta_{\mathbf{x}_i}$, it follows that $\mathbb{E}_{\delta_{\mathbf{x}_i}}[f] = f(\mathbf{x}_i)$, and the DRM consequently reduces to ERM on the i.i.d. sample $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.

On the one hand, the minimization problem (4.4) is different from minimizing the functional

$$\mathbb{E}_{\mathbb{P}_1} \ldots \mathbb{E}_{\mathbb{P}_n} \ell(\mathbf{x}_1, y_1, f(\mathbf{x}_1), \ldots, \mathbf{x}_n, y_n, f(\mathbf{x}_n)) + \Omega(\|f\|_{\mathcal{F}}) \tag{4.5}$$

for the special case of the additive loss ℓ, which is similar to the VRM proposed in Chapelle et al. (2000). Therefore, the solution of our regularization problem is different from what one would get in the limit by training on an infinitely many points sampled from $\mathbb{P}_1, \ldots, \mathbb{P}_m$. On the other hand, it is also different from minimizing the functional

$$\ell(\mathbf{m}_1, y_1, f(\mathbf{m}_1), \ldots, \mathbf{m}_n, y_n, f(\mathbf{m}_n)) + \Omega(\|f\|_{\mathcal{F}}) \tag{4.6}$$

where we substitute each distribution $\mathbb{P}_i$ by its mean $\mathbf{m}_i = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_i}[\mathbf{x}]$. In a sense, our framework is something in between.

Next I argue that when a function class $\mathcal{F}$ in (4.4) is chosen to be an RKHS, the corresponding problem amounts to learning on distributions when each of them is represented by the kernel mean embedding.

4.4.1 Hilbert Space Representation of Distributions

In order to learn from distributions, we use kernel mean embedding as a feature representation of distribution. It not only preserves necessary information of individual distributions, but also permits efficient computations. Briefly, let $\mathscr{H}$ denote an RKHS of functions $f : \mathcal{X} \to \mathbb{R}$, endowed with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The mean map from $\mathscr{P}$ into $\mathscr{H}$ is defined as

$$\boldsymbol{\mu} : \mathscr{P} \to \mathscr{H}, \;\; \mathbb{P} \longmapsto \int_{\mathcal{X}} k(\mathbf{x}, \cdot) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \;\; . \tag{4.7}$$

We assume that $k(\mathbf{x}, \cdot)$ is bounded for any $\mathbf{x} \in \mathcal{X}$. It can be shown that, if k is characteristic, the map (4.7) is injective, *i.e.*, all the information about the distribution is preserved (Sriperumbudur et al. 2010). For any $\mathbb{P}$, letting $\boldsymbol{\mu}_{\mathbb{P}} = \boldsymbol{\mu}(\mathbb{P})$, we have the reproducing property

$$\mathbb{E}_{\mathbb{P}}[f] = \langle \boldsymbol{\mu}_{\mathbb{P}}, f \rangle_{\mathscr{H}}, \;\; \forall f \in \mathscr{H} \;\; . \tag{4.8}$$

That is, we can see the mean embedding $\boldsymbol{\mu}_{\mathbb{P}}$ as a feature map associated with the kernel $K : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$, defined as $K(\mathbb{P}, \mathbb{Q}) = \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$. Since $\sup_{\mathbf{x}} \|k(\mathbf{x}, \cdot)\|_{\mathscr{H}} < \infty$, it also follows that

$$\begin{aligned} K(\mathbb{P}, \mathbb{Q}) &= \iint \langle k(\mathbf{x}, \cdot), k(\mathbf{z}, \cdot) \rangle_{\mathscr{H}} \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{z}) \\ &= \iint k(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{z}), \end{aligned} \tag{4.9}$$

where the second equality follows from the reproducing property of $\mathscr{H}$. It is immediate that K is a p.d. kernel on $\mathscr{P}$.

4.4.2 Representer Theorem for Distributions

The following theorem shows that optimal solutions of a suitable class of regularization problems involving distributions can be expressed as a finite linear combination of mean embeddings.

Theorem 4.1. *Given training examples $(\mathbb{P}_i, y_i) \in \mathscr{P} \times \mathbb{R}$, $i = 1, \ldots, m$, a strictly monotonically increasing function $\Omega : [0, +\infty) \to \mathbb{R}$, and a loss function $\ell : (\mathscr{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, any $f \in \mathcal{H}$ minimizing the regularized risk functional*

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathscr{H}}\right) \tag{4.10}$$

admits a representation of the form

$$f = \sum_{i=1}^{m} \alpha_i \boldsymbol{\mu}_{\mathbb{P}_i}$$

for some $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, m$.

Proof. By virtue of Proposition 2 in Sriperumbudur et al. (2010), the linear functional $\mathbb{E}_{\mathbb{P}}[\cdot]$ are bounded for all $\mathbb{P} \in \mathscr{P}$. Then, given $\mathbb{P}_1, \mathbb{P}_2, ..., \mathbb{P}_m$, any $f \in \mathcal{H}$ can be decomposed as

$$f = f_{\boldsymbol{\mu}} + f^{\perp}$$

where $f_{\boldsymbol{\mu}} \in \mathcal{H}$ lives in the span of $\boldsymbol{\mu}_{\mathbb{P}_i}$, *i.e.*, $f_{\boldsymbol{\mu}} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\mu}_{\mathbb{P}_i}$ and $f^{\perp} \in \mathscr{H}$ satisfying, for all j, $\langle f^{\perp}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle = 0$. Hence, for all j, we have

$$\mathbb{E}_{\mathbb{P}_j}[f] = \mathbb{E}_{\mathbb{P}_j}[f_{\boldsymbol{\mu}} + f^{\perp}] = \langle f_{\boldsymbol{\mu}} + f^{\perp}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle = \langle f_{\boldsymbol{\mu}}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle + \langle f^{\perp}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle = \langle f_{\boldsymbol{\mu}}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle$$

which is independent of $f^\perp$. As a result, the loss functional ℓ in (4.10) does not depend on $f^\perp$. For the regularization functional Ω, since $f^\perp$ is orthogonal to $\sum_{i=1}^m \alpha_i \boldsymbol{\mu}_{\mathbb{P}_i}$ and Ω is strictly monotonically increasing, we have

$$\Omega(\|f\|) = \Omega(\|f_{\boldsymbol{\mu}} + f^\perp\|) = \Omega(\sqrt{\|f_{\boldsymbol{\mu}}\|^2 + \|f^\perp\|^2}) \geq \Omega(\|f_{\boldsymbol{\mu}}\|)$$

with equality if and only if $f^\perp = 0$ and thus $f = f_{\boldsymbol{\mu}}$. Consequently, any minimizer must take the form $f = \sum_{i=1}^m \alpha_i \boldsymbol{\mu}_{\mathbb{P}_i} = \sum_{i=1}^m \alpha_i \mathbb{E}_{\mathbb{P}_i}[k(\mathbf{x}, \cdot)]$. ■

Theorem 4.1 clearly indicates how each distribution contributes to the minimizer of (4.10). Roughly speaking, the coefficients α_i controls the contribution of the distributions through the mean embeddings $\boldsymbol{\mu}_{\mathbb{P}_i}$. Furthermore, if we restrict $\mathscr{P}$ to a class of Dirac measures $\delta_{\mathbf{x}}$ on $\mathcal{X}$ and consider the training set $\{(\delta_{\mathbf{x}_i}, y_i)\}_{i=1}^m$, the functional (4.10) reduces to the usual regularization functional (Schölkopf et al. 2001a) and the solution reduces to $f = \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \cdot)$. Therefore, the standard representer theorem is recovered as a particular case (see also Dinuzzo and Schölkopf (2012) for more general results on representer theorem).

4.5 Support Measure Machines

This subsection extends SVMs to deal with probability distributions, leading to *support measure machines* (SMMs). In its general form, an SMM amounts to solving an SVM problem with the expected kernel $K(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}, \mathbf{z}\sim\mathbb{Q}}[k(\mathbf{x}, \mathbf{z})]$. This kernel can be computed in closed-form for certain classes of distributions and kernels k. Examples are given in Table 4.1.

Alternatively, one can approximate the kernel $K(\mathbb{P}, \mathbb{Q})$ by the empirical estimate:

$$K_{\text{emp}}(\widehat{\mathbb{P}}_n, \widehat{\mathbb{Q}}_m) = \frac{1}{n \cdot m} \sum_{i=1}^n \sum_{j=1}^m k(\mathbf{x}_i, \mathbf{z}_j) \tag{4.11}$$

where $\widehat{\mathbb{P}}_n$ and $\widehat{\mathbb{Q}}_m$ are empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$ given random samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{z}_j\}_{j=1}^m$, respectively. A finite sample of size m from a distribution $\mathbb{P}$ suffices (with high probability) to compute an approximation within an error of $O(m^{-\frac{1}{2}})$. Instead, if the sample set is sufficiently large, one may choose to approximate the true distribution by simpler probabilistic models, *e.g.*, a mixture of Gaussians model, and choose a kernel k whose expected value admits an analytic form. Storing only the parameters of probabilistic models may save some space compared to storing all data points.

Note that the standard SVM feature map $\phi(\mathbf{x})$ is usually nonlinear in $\mathbf{x}$, whereas $\boldsymbol{\mu}_{\mathbb{P}}$ is *linear* in $\mathbb{P}$. Thus, for an SMM, the first level kernel k is used to obtain a vectorial representation of the measures, and the second level kernel K allows for a nonlinear algorithm on distributions. For clarity, we will refer to k and K as the **embedding kernel** and the **level-2 kernel**, respectively. Table 4.2 gives some examples of level-2 kernels that can be applied in this framework.

4.5.1 Kernels on Probability Distributions

As the map (4.7) is linear in $\mathscr{P}$, optimizing the functional (4.10) amounts to finding a function in $\mathscr{H}$ that approximate well functions from $\mathscr{P}$ to $\mathbb{R}$ in the function class

$$\mathcal{F} \triangleq \{\mathbb{P} \to \int_{\mathcal{X}} g \, \mathrm{d}\mathbb{P} \,|\, \mathbb{P} \in \mathscr{P},\, g \in C_b(\mathcal{X})\}$$

where $C_b(\mathcal{X})$ is a class of bounded continuous functions on $\mathcal{X}$. Since $\delta_{\mathbf{x}} \in \mathscr{P}$ for any $\mathbf{x} \in \mathcal{X}$, it follows that $C_b(\mathcal{X}) \subset \mathcal{F} \subset C_b(\mathscr{P})$ where $C_b(\mathscr{P})$ is a class of bounded continuous functions

Table 4.1: The analytic forms of expected kernels for different choices of kernels and distributions.

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

on $\mathscr{P}$ endowed with the topology of weak convergence and the associated Borel σ-algebra. The following lemma states the relation between the RKHS $\mathscr{H}$ induced by the kernel k and the function class $\mathcal{F}$.

Lemma 4.2. *Assuming that $\mathcal{X}$ is compact, the RKHS $\mathscr{H}$ induced by a kernel k is dense in $\mathcal{F}$ if k is universal,* i.e.*, for every function $F \in \mathcal{F}$ and every $\varepsilon > 0$ there exists a function $g \in \mathscr{H}$ with* $\sup_{\mathbb{P} \in \mathscr{P}} |F(\mathbb{P}) - \int g \, \mathrm{d}\mathbb{P}| \leq \varepsilon$.

Proof. Assume that k is universal. Then, for every function $f \in C_b(\mathcal{X})$ and every $\varepsilon > 0$ there exists a function $g \in \mathscr{H}$ induced by k with $\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - g(\mathbf{x})| \leq \varepsilon$ (Steinwart 2002). Hence, by linearity of $\mathcal{F}$, for every $F \in \mathcal{F}$ and every $\varepsilon > 0$ there exists a function $h \in \mathscr{H}$ such that $\sup_{\mathbb{P} \in \mathscr{P}} |F(\mathbb{P}) - \int h \, \mathrm{d}\mathbb{P}| \leq \varepsilon$. ■

Nonlinear kernels on $\mathscr{P}$ can be defined in an analogous way to nonlinear kernels on $\mathcal{X}$, by treating mean embeddings $\boldsymbol{\mu}_{\mathbb{P}}$ of $\mathbb{P} \in \mathscr{P}$ as its feature representation. First, assume that the map (4.7) is injective and let $\langle \cdot, \cdot \rangle_{\mathscr{P}}$ be an inner product on $\mathscr{P}$. By linearity, we have $\langle \mathbb{P}, \mathbb{Q} \rangle_{\mathscr{P}} = \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ (cf. Berlinet and Thomas-Agnan (2004) for more details). Then, the nonlinear kernels on $\mathscr{P}$ can be defined as

$$K(\mathbb{P}, \mathbb{Q}) = \kappa(\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}}) = \langle \psi(\boldsymbol{\mu}_{\mathbb{P}}), \psi(\boldsymbol{\mu}_{\mathbb{Q}}) \rangle_{\mathscr{G}}$$

where κ is a positive definite kernel and $\mathscr{G}$ denotes the correspond RKHS of functions from $\mathscr{H}$ to $\mathbb{R}$. As a result, many standard nonlinear kernels on $\mathcal{X}$ can be used to define nonlinear kernels on $\mathscr{P}$ as long as the kernel evaluation depends entirely on the inner product $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$, *e.g.*, $K(\mathbb{P}, \mathbb{Q}) = (\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} + c)^d$ (see Table 4.2 for more examples). Although requiring more computational effort, their practical use is simple and flexible. Specifically, the notion of p.d. kernels on distributions proposed in this work is so generic that standard kernel functions can be reused to derive kernels on distributions that are different from many other kernel functions proposed specifically for certain distributions.

It has been recently proved that the Gaussian RBF kernel given by $K(\mathbb{P}, \mathbb{Q}) = \exp(-\frac{\gamma}{2}\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|^2_{\mathscr{H}})$, $\forall \mathbb{P}, \mathbb{Q} \in \mathscr{P}$ is universal w.r.t $C_b(\mathscr{P})$ given that $\mathcal{X}$ is compact and the map $\boldsymbol{\mu}$ is injective (Christmann and Steinwart 2010). Despite its success in real-world applications, the theory of kernel-based classifiers beyond the input space $\mathcal{X} \subset \mathbb{R}^d$, as also mentioned by Christmann and Steinwart (2010), is still incomplete. It is therefore of theoretical interest to consider more general classes of universal kernels on probability distributions.

4.5.2 Flexible Support Vector Machines

It turns out that, for certain choices of distributions $\mathbb{P}$, the linear SMM trained using $\{(\mathbb{P}_i, y_i)\}_{i=1}^m$ is equivalent to an SVM trained using some samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ with an appropriate choice of kernel function.

Table 4.2: Examples of some well-known kernel functions that can be used as inducing kernels.

Kernel Function	**The Level-2 Kernel** $K(\mathbb{P},\mathbb{Q})$
Gaussian kernel	$\exp\left(-0.5\gamma\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2\right)$
Exponential kernel	$\exp\left(-\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{2\sigma^2}\right)$
Laplacian kernel	$\exp\left(-\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\sigma}\right)$
Hyperbolic Tangent kernel	$\tanh(\alpha\langle\boldsymbol{\mu}_{\mathbb{P}},\boldsymbol{\mu}_{\mathbb{Q}}\rangle_{\mathscr{H}}+c)$
Rational Quadratic kernel	$1-\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2}{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2+c}$
Multiquadratic kernel	$\sqrt{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2+c}$
Inverse multiquadric kernel	$\dfrac{1}{\sqrt{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2+c}}$
Circular kernel	$\dfrac{2}{\pi}\arccos\left(-\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2}{\sigma}\right)-\dfrac{2}{\pi}\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\sigma}\sqrt{1-\left(\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\sigma}\right)^2}$ if $\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|<\sigma$, zero otherwise.
Spherical kernel	$1-\dfrac{3}{2}\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\sigma}+\dfrac{1}{2}\left(\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\sigma}\right)^3$ if $\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|<\sigma$, zero otherwise.
Wave kernel	$\dfrac{\theta}{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}\sin\left(\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|}{\theta}\right)$
Power kernel	$-\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^d$
Log kernel	$-\log\left(\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^d+1\right)$
Cauchy kernel	$\dfrac{1}{1+\dfrac{\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^2}{\sigma}}$
Generalized T-student kernel	$\dfrac{1}{1+\|\boldsymbol{\mu}_{\mathbb{P}}-\boldsymbol{\mu}_{\mathbb{Q}}\|^d}$

Lemma 4.3. *Let $k(\mathbf{x},\mathbf{z})$ be a bounded positive definite kernel on a measure space such that $\iint k(\mathbf{x},\mathbf{z})^2\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z}<\infty$, and $g(\mathbf{x},\tilde{\mathbf{x}})$ be a square integrable function such that $\int g(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\tilde{\mathbf{x}}<\infty$ for all $\mathbf{x}$. Given a sample $\{(\mathbb{P}_i,y_i)\}_{i=1}^m$ where each $\mathbb{P}_i$ is assumed to have a density given by $g(\mathbf{x}_i,\mathbf{x})$, the linear SMM is equivalent to the SVM on the training sample $\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ with kernel*

$$K_g(\mathbf{x},\mathbf{z})=\iint k(\tilde{\mathbf{x}},\tilde{\mathbf{z}})g(\mathbf{x},\tilde{\mathbf{x}})g(\mathbf{z},\tilde{\mathbf{z}})\,\mathrm{d}\tilde{\mathbf{x}}\,\mathrm{d}\tilde{\mathbf{z}}.$$

Proof. For a training sample $\{(\mathbf{x}_i,y_i)\}_{i=1}^m$, the SVM with kernel K_g minimizes

$$\ell(\{\mathbf{x}_i,y_i,f(\mathbf{x}_i)+b\}_{i=1}^m)+\lambda\|f\|^2_{\mathscr{H}_{K_g}}\ .$$

By the representer theorem, $f(\mathbf{x})=\sum_{i=1}^m\alpha_iK_g(\mathbf{x},\mathbf{x}_j)$ with some $\alpha_i\in\mathbb{R}$, hence this is equivalent to

$$\ell(\{\mathbf{x}_i,y_i,\sum_{j=1}^m\alpha_jK_g(\mathbf{x}_i,\mathbf{x}_j)+b\}_{i=1}^m)+\lambda\sum_{i,j=1}^m\alpha_i\alpha_jK_g(\mathbf{x}_i,\mathbf{x}_j)\ .$$

Next, consider the kernel mean of the probability measure $g(\mathbf{x}_i,\mathbf{x})\,\mathrm{d}\mathbf{x}$ given by $\boldsymbol{\mu}_i=\int k(\cdot,\tilde{\mathbf{x}})g(\mathbf{x}_i,\tilde{\mathbf{x}})\,\mathrm{d}\tilde{\mathbf{x}}$ and note that $\langle\boldsymbol{\mu}_i,f\rangle_{\mathscr{H}_k}=\int f(\tilde{\mathbf{x}})g(\mathbf{x}_i,\tilde{\mathbf{x}})\,\mathrm{d}\tilde{\mathbf{x}}$ for any $f\in\mathscr{H}_k$. The linear SMM with loss ℓ and kernel k minimizes

$$\ell(\{\mathbb{P}_i,y_i,\langle\boldsymbol{\mu}_i,f\rangle_{\mathscr{H}_k}+b\}_{i=1}^m)+\lambda\|f\|^2_{\mathscr{H}_k}\ .$$

By Theorem 4.1, each minimizer f admits a representation of the form

$$f = \sum_{j=1}^{m} \alpha_j \boldsymbol{\mu}_j = \sum_{j=1}^{m} \alpha_j \int k(\cdot, \tilde{\mathbf{x}}) g(\mathbf{x}_j, \tilde{\mathbf{x}}) \, \mathrm{d}\tilde{\mathbf{x}} \ .$$

Thus, for this f we have

$$\langle \boldsymbol{\mu}_i, f \rangle_{\mathscr{H}_k} = \sum_{j=1}^{m} \alpha_j \iint k(\tilde{\mathbf{z}}, \tilde{\mathbf{x}}) g(\mathbf{x}_i, \tilde{\mathbf{x}}) g(\mathbf{x}_j, \tilde{\mathbf{z}}) \, \mathrm{d}\tilde{\mathbf{x}} \, \mathrm{d}\tilde{\mathbf{z}} = \sum_{j=1}^{m} \alpha_j K_g(\mathbf{x}_i, \mathbf{x}_j)$$

and

$$\|f\|^2_{\mathscr{H}_k} = \sum_{i,j=1}^{m} \alpha_i \alpha_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j K_g(\mathbf{x}_i, \mathbf{x}_j)$$

, as above. This completes the proof. ■

Note that the important assumption for this equivalence is that the distributions $\mathbb{P}_i$ differ only in their location in the parameter space. This need not be the case in all possible applications of SMMs. Furthermore, we have $K_g(\mathbf{x}, \mathbf{z}) = \left\langle \int k(\tilde{\mathbf{x}}, \cdot) g(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\tilde{\mathbf{x}}, \int k(\tilde{\mathbf{z}}, \cdot) g(\mathbf{z}, \tilde{\mathbf{z}}) \, \mathrm{d}\tilde{\mathbf{z}} \right\rangle_{\mathscr{H}}$. Thus, it is clear that the feature map of $\mathbf{x}$ depends not only on the kernel k, but also on the density $g(\mathbf{x}, \tilde{\mathbf{x}})$. Consequently, by virtue of Lemma 4.3, the kernel K_g allows the SVM to place different kernels at each data point. We call this algorithm a *flexible SVM* (Flex-SVM).

Consider the linear SMM with Gaussian distributions $\mathcal{N}(\mathbf{x}_1; \sigma_1^2 \cdot \mathbf{I}), \ldots, \mathcal{N}(\mathbf{x}_m; \sigma_m^2 \cdot \mathbf{I})$ and Gaussian RBF kernel k_{σ^2} with bandwidth parameter σ. The convolution theorem of Gaussian distributions implies that this SMM is equivalent to a flexible SVM that places a data-dependent kernel $k_{\sigma^2 + 2\sigma_i^2}(\mathbf{x}_i, \cdot)$ on training example $\mathbf{x}_i$, *i.e.*, a Gaussian RBF kernel with larger bandwidth.

4.5.3 A Unifying View: SVM and Parzen Window Classifier

Although having been introduced independently, the proposed framework has an intrinsic connection to two well-known existing learning algorithms, namely a support vector machine (SVM) and Parzen window classifier (PWC).

Regularization on Distributions. Recall that a regularization problem on probability distributions can be formulated as follow. Given training examples $(\mathbb{P}_i, y_i) \in \mathscr{P} \times \mathbb{R}$, $i = 1, \ldots, m$, a strictly monotonically increasing function $\Omega : [0, +\infty) \to \mathbb{R}$, and a loss function $\ell : (\mathscr{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, we find $f \in \mathscr{H}$ such that the regularization functional

$$\ell \left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f] \right) + \Omega \left(\|f\|_{\mathscr{H}} \right) \tag{4.12}$$

is minimized where $\mathcal{H}$ is an RKHS with a reproducing kernel k. By representer theorem on distributions, any solution f admits a representation of the form $f = \sum_{i=1}^{m} \alpha_i \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_i}[k(\mathbf{x}, \cdot)]$ for some $\boldsymbol{\alpha} \in \mathbb{R}^m$. Given training examples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \{+1, -1\}$, we will show that both SVM and PWC can be recovered as a solution to the regularization functional (4.12).

Support Vector Machines. The connection between SVM and SMM is quite straightforward. In this case, we replace each training sample $\mathbf{x}_i$ by a Dirac measure $\delta_{\mathbf{x}_i}$ centered at that sample. Note that this reparameterization does not alter the problem as the map $\mathbf{x}_i \mapsto k(\mathbf{x}_i, \cdot)$ and $\delta_{\mathbf{x}_i} \mapsto \int k(\mathbf{x}, \cdot) \delta_{\mathbf{x}_i}(\mathbf{x})$ are equivalent. Replacing $\mathbb{P}_i$ in (4.10) by $\delta_{\mathbf{x}_i}$ yields

$$\ell_H \left(\mathbf{x}_1, y_1, f(\mathbf{x}_1), \ldots, \mathbf{x}_m, y_m, f(\mathbf{x}_m) \right) + \Omega \left(\|f\|_{\mathscr{H}} \right)$$

which corresponds to the SVM (Schölkopf and Smola 2001). In summary, the SVM treats every samples as a probability distribution that captures the *local* information of the sample.

Parzen Window Classifiers. As opposed to the SVM, the PWC approaches the problem from a completely opposite direction. In general, it begins by estimating the class conditional distribution

$$\mathbb{P}(\mathbf{x}|y) = \frac{1}{|\{i|y_i = y\}|} \sum_{i,y_i=y} k(\mathbf{x}, \mathbf{x}_i)$$

and by Bayes rule the posterior can be computed by

$$\mathbb{P}(y|\mathbf{x}) = \frac{\sum_{i,y_i=y} k(\mathbf{x}, \mathbf{x}_i)}{\sum_{i=1}^{n} k(\mathbf{x}, \mathbf{x}_i)}. \tag{4.13}$$

Consequently, for binary classification problem, the estimated class conditional $\mathbb{P}(\cdot|y = +1)$ and $\mathbb{P}(\cdot|y = -1)$ can be seen as class-specific mean functions in $\mathscr{H}$, *i.e.*,

$$M^+ = \frac{1}{|\{i|y_i = +1\}|} \sum_{y_i=+1} k(\mathbf{x}_i, \cdot), \qquad M^- = \frac{1}{|\{i|y_i = -1\}|} \sum_{y_i=-1} k(\mathbf{x}_i, \cdot)$$

where $M^+ = \mathbb{P}(\cdot|y = +1)$ and $M^- = \mathbb{P}(\cdot|y = -1)$. Hence, the classification function based on the posterior (4.13) can be equivalently written as

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{x}, M^+ \rangle_{\mathscr{H}} - \langle \mathbf{x}, M^- \rangle_{\mathscr{H}}) \tag{4.14}$$

It is not difficult to see that for a positive semi-definite kernel k the mean functions M^+ and M^- are two distinct functions in the RKHS $\mathscr{H}$. As a result, the classification function (4.14) can be obtained by solving a classification problem on training examples $(M^+, +1)$ and $(M^-, -1)$, which is exactly equivalent to minimizing

$$\ell_H(\widehat{\mathbb{P}}_+, +1, \mathbb{E}_{\widehat{\mathbb{P}}_+}[f], \widehat{\mathbb{P}}_-, -1, \mathbb{E}_{\widehat{\mathbb{P}}_-}[f]) + \Omega\left(\|f\|_{\mathscr{H}}\right)$$

As opposed to the SVM, the PWC treats class conditional distributions as training samples, emphasizing more on the *global* information of the training data.

Intuitively, we may consider both SVM and PWC as the extreme ends of the spectrum of learning algorithms: one that look locally at the training data and another one that consider its global properties. Consequently, It is natural to ask what constitute various learning algorithms along this spectrum?

From the distributional point of view, these two learning algorithms employ different learning strategy. Roughly speaking, the SVM performs a learning at the most fine-grained level for the best accuracy at the expense of training time. On the other hand, the Parzen window classifier trade-off the accuracy with the training time (no training time) in order to obtain the solution very quickly. These two learning approaches are also different in term of estimation time. A good learning strategy therefore should trade-off these quantities.

4.5.4 Extensions to Other Algorithms

Algorithmically, one of the advantages of the proposed framework is that other algorithms can be generalized as long as those algorithms rely only on the evaluation of the kernel between probability distributions. Algorithms reviewed in Section 2.2.3 such as kernel ridge regression and Gaussian process can be generalized to a space of probability distributions. For example, Szabó et al. (2015) studies the regression problem on distributions using kernel ridge regression.

One may also consider (5.22) as a covariance kernel for GP regression and classification which allows for uncertain inputs. That is, we may consider the following covariance function

$$\mathrm{Cov}(f(\mathbb{P}), f(\mathbb{Q})) := \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} = \iint k(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(y) \, .$$

The kernel mean representation is in general infinite dimensional when the RKHS associated with the kernel has infinite dimension, *e.g.*, Gaussian RBF kernel. Nevertheless, we may resort to a finite dimensional RKHS. In which case the mean embedding becomes a finite dimensional object which we can think of as a feature vector representing the distribution. As a result, we can apply any learning algorithm on this representation. However, for infinite dimensional RKHS, most algorithms need to operate in their dual forms which can be prohibitive for large-scale problems. In this case, one may resort to a finite approximation of the kernel means (cf. Section 2.3.6) which then allows the algorithms to work in their primal form. The approximation also allows for a wider class of learning algorithms on distributions, *e.g.*, ensemble algorithms, random forest, *etc*. For instance, Lopez-Paz et al. (2015a) and Lopez-Paz et al. (2015b) approximates the kernel mean representation using the random Fourier feature (Rahimi and Recht 2007) and then apply random forest classifiers on these approximations.

4.6 Theoretical Analysis

This section presents key theoretical aspects of the proposed framework, which reveal important connection between kernel-based learning algorithms on the space of distributions and on the input space on which they are defined.

4.6.1 Risk Deviation Bound

Given a training sample $\{(\mathbb{P}_i, y_i)\}_{i=1}^m$ drawn i.i.d. from some unknown probability distribution $\mathcal{P}$ on $\mathscr{P} \times \mathcal{Y}$, a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and a function class $\mathcal{F}$, the goal of statistical learning is to find the function $f \in \mathcal{F}$ that minimizes the expected risk functional

$$\mathcal{R}(f) = \int_{\mathscr{P}} \int_{\mathcal{X}} \ell(y, f(\mathbf{x})) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathcal{P}(\mathbb{P}, y). \tag{4.15}$$

Since $\mathcal{P}$ is unknown, the empirical risk

$$\widehat{\mathcal{R}}(f) = \frac{1}{m} \sum_{i=1}^{m} \int_{\mathcal{X}} \ell(y_i, f(\mathbf{x})) \, \mathrm{d}\mathbb{P}_i(\mathbf{x}) \tag{4.16}$$

based on the training sample is considered instead. Furthermore, the risk functional can be simplified further by considering $\frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{\mathbf{x}_{ij} \sim \mathbb{P}_i} \ell(y_i, f(\mathbf{x}_{ij}))$ based on n samples $\mathbf{x}_{ij}$ drawn from each $\mathbb{P}_i$.

Our framework, on the other hand, alleviates the problem by minimizing the risk functional

$$\mathcal{R}^{\mu}(f) = \int_{\mathscr{P}} \ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]) \, \mathrm{d}\mathcal{P}(\mathbb{P}, y) \tag{4.17}$$

for $f \in \mathscr{H}$ with corresponding empirical risk functional

$$\widehat{\mathcal{R}}^{\mu}(f) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, \mathbb{E}_{\mathbb{P}_i}[f(\mathbf{x})]) \tag{4.18}$$

(cf. the discussion at the end of Section 4.4). It is often easier to optimize $\widehat{\mathcal{R}}^{\mu}(f)$ as the expectation can be computed exactly for certain choices of $\mathbb{P}_i$ and $\mathscr{H}$. Moreover, for universal $\mathscr{H}$, this simplification preserves all information of the distributions. Nevertheless, there is still a loss of information due to the loss function ℓ.

Due to the i.i.d. assumption, the analysis of the difference between $\mathcal{R}$ and $\mathcal{R}^{\mu}$ can be simplified w.l.o.g. to the analysis of the difference between $\mathbb{E}_{\mathbb{P}}[\ell(y, f(\mathbf{x}))]$ and $\ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])$ for a particular distribution $\mathbb{P} \in \mathscr{P}$. The theorem below provides a bound on the difference between $\mathbb{E}_{\mathbb{P}}[\ell(y, f(\mathbf{x}))]$ and $\ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])$.

Theorem 4.4. *Given an arbitrary probability distribution $\mathbb{P}$ with variance σ^2, a Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$ with constant C_f, an arbitrary loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is Lipschitz continuous in the second argument with constant C_ℓ, it follows that*

$$|\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[\ell(y, f(\mathbf{x}))] - \ell(y, \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[f(\mathbf{x})])| \leq 2C_\ell C_f \sigma$$

for any $y \in \mathbb{R}$.

Proof. Assume that $\mathbf{x}$ is distributed according to $\mathbb{P}$. Let $\mathbf{m}_X$ be the mean of X in $\mathbb{R}^d$. Thus, we have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}[\ell(y, f(\mathbf{x}))] - \ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])| &\leq \int |\ell(y, f(\tilde{\mathbf{x}})) - \ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])| \, \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}}) \\
&\leq C_\ell \int |f(\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]| \, \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}}) \\
&\leq \underbrace{C_\ell \int |f(\tilde{\mathbf{x}}) - f(\mathbf{m}_X)| \, \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}})}_{A} \\
&\quad + \underbrace{C_\ell |f(\mathbf{m}_X) - \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})]|}_{B} \, .
\end{aligned}$$

Control of (A). The first term is upper bounded by

$$C_\ell \int C_f \|\tilde{\mathbf{x}} - \mathbf{m}_X\| \, \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}}) \leq C_\ell C_f \sigma \ , \tag{4.19}$$

where the last inequality is given by $\mathbb{E}_{\mathbb{P}}[\|\tilde{\mathbf{x}} - \mathbf{m}_X\|] \leq \sqrt{\mathbb{E}_{\mathbb{P}}[\|\tilde{\mathbf{x}} - \mathbf{m}_X\|^2]} = \sigma$.

Control of (B). Similarly, the second term is upper bounded by

$$C_\ell \left| \int f(\mathbf{m}_X) - f(\tilde{\mathbf{x}}) \right| \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}}) \leq C_\ell \int C_f \|\mathbf{m}_X - \tilde{\mathbf{x}}\| \, \mathrm{d}\mathbb{P}(\tilde{\mathbf{x}}) \leq C_\ell C_f \sigma \ . \tag{4.20}$$

Combining (4.19) and (4.20) yields

$$|\mathbb{E}_{\mathbb{P}}[\ell(y, f(\mathbf{x}))] - \ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])| \leq 2C_\ell C_f \sigma \ ,$$

thus completing the proof. ■

Theorem 4.4 indicates that if the random variable $\mathbf{x}$ is concentrated around its mean and the function f and ℓ are well-behaved, *i.e.*, Lipschitz continuous, then the loss deviation $|\mathbb{E}_{\mathbb{P}}[\ell(y, f(\mathbf{x}))] - \ell(y, \mathbb{E}_{\mathbb{P}}[f(\mathbf{x})])|$ will be small. As a result, if this holds for any distribution $\mathbb{P}_i$ in the training set $\{(\mathbb{P}_i, y_i)\}_{i=1}^m$, the true risk deviation $|\mathcal{R} - \mathcal{R}^{\mu}|$ is also expected to be small.

4.6.2 Rademacher Complexity and Generalization Bound

In this section, I outline some well-known results in learning theory and then give a generalization bound based on Rademacher complexity of linear SMM.

Definition 4.1. *(Rademacher complexity). Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be a set of samples drawn from some unknown distribution $\mathbb{P}$ on $\mathcal{X}$. The empirical Rademacher complexity of $\mathcal{F}$ is*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right| \mid \mathbf{x}_1, ..., \mathbf{x}_n\right],$$

where σ_i are i.i.d. uniform random variables on ± 1. The Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}[\hat{\mathfrak{R}}_n(\mathcal{F})] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right|\right].$$

The Rademacher complexity is quite popular in learning theory as a characterization of the richness of the function class $\mathcal{F}$. Intuitively, it measures how well correlated the function in $\mathcal{F}$ is to the random noise on a sample S. The following lemma shows the relation between $\mathfrak{R}_n(\mathcal{F})$ and $\hat{\mathfrak{R}}_n(\mathcal{F})$.

Lemma 4.5. *Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $[0, 1]$ and a sample $S = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. Then, with probability at least $1 - \delta$,*

$$\mathfrak{R}_n(\mathcal{F}) \leq \hat{\mathfrak{R}}_n(\mathcal{F}) + 2\sqrt{\frac{\ln 1/\delta}{2n}}.$$

Lemma 4.6 and Theorem 4.7 provide an upper bound of uniform convergence in expectation and generalization bounds for a function class $\mathcal{F}$ in term of its Rademacher complexity, respectively.

Lemma 4.6 (Koltchinskii and Panchenko (2002))**.**

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)\right|\right] \leq 2\mathfrak{R}_n(\mathcal{F}).$$

Theorem 4.7 (Bartlett and Mendelson (2003))**.** *Let $\delta \in (0, 1)$ and $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $[0, 1]$. Then with probability at least $1 - \delta$, all $f \in \mathcal{F}$ satisfy*

$$\mathbb{E}f(\mathbf{x}) \leq \mathbb{E}_n f(\mathbf{x}) + 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{2n}}.$$

By virtue of Lemma 4.5, we also have with probability at least $1 - \delta$

$$\mathbb{E}f(\mathbf{x}) \leq \mathbb{E}_n f(\mathbf{x}) + 2\hat{\mathfrak{R}}_n(\mathcal{F}) + 5\sqrt{\frac{\ln 2/\delta}{2n}}.$$

In most cases, however, we are interested in the composition the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and the hypothesis $f \in \mathcal{F}$, *e.g.*, $\ell(y, f(\mathbf{x})) = (f(\mathbf{x}) - y)^2$. Deriving a complexity measure of the function $\ell \circ f$ can be involved especially for non-linear loss function. Ledoux-Talagrand contraction, given in Lemma 4.8, allows us to bound the Rademacher complexity of such functions in terms of the Rademacher complexity of $\mathcal{F}$.

Lemma 4.8 (Ledoux-Talagrand contraction)**.** *If $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and satisfies $\mathcal{A}(0) = 0$, then $\hat{\mathfrak{R}}_n(\mathcal{A} \circ \mathcal{F}) \leq 2L\hat{\mathfrak{R}}_n(\mathcal{F})$.*

We are now in a position to derive the Rademacher complexity of linear SMM and the corresponding generalization bound.

Lemma 4.9. *Let $\mathscr{H}$ be an RKHS with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathscr{P}$ be a set of probability measures on $\mathcal{X}$, $\mathcal{S} = (\mathbb{P}_1, ..., \mathbb{P}_n)$ be a sample drawn i.i.d. according to some distribution $\mathcal{Q}$ on $\mathscr{P}$. Given a positive constant b and a class of real-valued functions*

$$\mathcal{F}_b = \left\{ g : \mathscr{P} \to \mathbb{R} \,\middle|\, \mathbf{w} \in \mathscr{H}, \|\mathbf{w}\|_{\mathscr{H}} \leq b \quad \textit{s.t.} \quad g(\mathbb{P}) = \langle \boldsymbol{\mu}_{\mathbb{P}}, \mathbf{w} \rangle_{\mathscr{H}} \right\},$$

the Rademacher complexity of $\mathcal{F}_b$ is given by

$$\mathfrak{R}_n(\mathcal{F}_b) \leq \frac{b}{\sqrt{n}} \sqrt{\mathbb{E}_{\mathbb{P}\sim\mathcal{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \mathbf{x})]}.$$

Proof of Lemma 4.9. Let $\mathcal{S} = \{\mathbb{P}_1, ..., \mathbb{P}_n\}$ be a set of samples drawn from a distribution $\mathcal{Q}$ on $\mathscr{P}$, and σ_i be i.i.d. uniform random variable on ± 1. Consequently, we have

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_b) &= \mathbb{E}_{\mathcal{Q}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{F}_b} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(\mathbb{P}_i) \right| \right] \\
&= \mathbb{E}_{\mathcal{Q}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\| \leq b} \left| \left\langle \mathbf{w}, \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_i}[k(\mathbf{x}, \cdot)] \right\rangle \right| \right] \\
&\leq \frac{b}{n} \mathbb{E}_{\mathcal{Q}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\left(\sum_{i,j=1}^{n} \sigma_i \sigma_j \mathbb{E}_{\mathbf{x}_i\sim\mathbb{P}_i, \mathbf{x}_j\sim\mathbb{P}_j}[k(\mathbf{x}_i, \mathbf{x}_j)] \right)^{\frac{1}{2}} \right] \\
&\leq \frac{b}{n} \mathbb{E}_{\mathcal{Q}} \left[\left(\sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_i}[k(\mathbf{x}, \mathbf{x})] \right)^{\frac{1}{2}} \right] \\
&\leq \frac{b}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{Q}} \mathbb{E}_{\mathbb{P}_i}[k(\mathbf{x}, \mathbf{x})] \right)^{\frac{1}{2}} \\
&= \frac{b}{\sqrt{n}} \sqrt{\mathbb{E}_{\mathbb{P}\sim\mathcal{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \mathbf{x})]}.
\end{aligned}$$

This completes the proof. ■

As an interpretation, Lemma 4.9 considers a linear SMM in an RKHS-ball of radius b. The result suggests that $\mathfrak{R}_n(\mathcal{F}_b)$ is bounded by a finite constant as long as $\sup_{x\in\mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$ and the bound vanishes as $n \to \infty$. Furthermore, using the results from Table 4.1 and a prior knowledge on $\mathcal{Q}$, one can derive an analytic form of $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \mathbf{x})]$.

Theorem 4.10. *Let $\mathcal{S} = \{(\mathbb{P}_1, y_1), ..., (\mathbb{P}_n, y_n)\}$ be i.i.d. sample drawn according to some distribution over $\mathscr{P} \times \{-1, +1\}$. For any $h(\mathbb{P}) = \operatorname{sign}(f(\mathbb{P}))$ for $f \in \mathcal{F}_b$, with probability at least $1 - \delta$ over the samples of size n,*

$$\Pr(h(\mathbb{P}) \neq y) \leq \widehat{\Pr}(h(\mathbb{P}_i) \neq y_i) + 2b \sqrt{\frac{\mathbb{E}_{\mathbb{P}\sim\mathcal{Q}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x}, \mathbf{x})]}{n}} + \sqrt{\frac{\ln 1/\delta}{2n}}.$$

Proof of Theorem 4.10. Consider the hypothesis $h(\mathbb{P}) = \text{sign}(g(\mathbb{P})) \in \{-1, 1\}$ and the loss function $\ell(\mathbb{P}, y, h(\mathbb{P})) = \Theta(-yg(\mathbb{P}))$ where Θ is the Heavyside function,

$$\Theta(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\Pr(y \neq \text{sign}(g(\mathbb{P}))) = \mathbb{E}[\Theta(-yg(\mathbb{P}))]$. Since $\eta := (1 - yg(\mathbb{P}))_+ \geq \Theta(-yg(\mathbb{P}))$ and $\eta(\cdot)$ is Lipschitz with constant 1, we have

$$\Pr(y \neq \text{sign}(g(\mathbb{P}))) = \mathbb{E}[\Theta(-yg(\mathbb{P}))] \leq \mathbb{E}[(1 - yg(\mathbb{P}))_+] \,.$$

Moreover, $\sigma_i yg(\mathbb{P})$ is symmetric around 0, so that $yg(\mathbb{P})$ has the same distribution and $\mathfrak{R}_n(Y\mathcal{F}_b) = \mathfrak{R}_n(\mathcal{F}_b)$. Furthermore, $\tilde{\eta}(\cdot) = \eta(\cdot) - 1$ is Lipschitz with constant 1 and satisfies $\tilde{\eta}(0) = 0$. Lemma 4.8 implies

$$\mathfrak{R}_n(\tilde{\eta}(Y\mathcal{F}_b)) \leq 2\mathfrak{R}_n(Y\mathcal{F}_b) = 2\mathfrak{R}_n(\mathcal{F}_b)$$

By Theorem 4.7 and Lemma 4.9, with probability at least $1 - \delta$,

$$\begin{aligned} \Pr(h(\mathbb{P}) \neq y) - 1 &\leq \mathbb{E}_n[(1 - yh(\mathbb{P}))_+ - 1] + 2\mathfrak{R}_n(\tilde{\eta}(Y\mathcal{F})) + \sqrt{\frac{\ln 1/\delta}{2n}} \\ &= \widehat{\Pr}(h(\mathbb{P}_i) \neq y_i) + 2b\sqrt{\frac{\mathbb{E}_{\mathbb{P}\sim\mathcal{Q}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}}[k(\mathbf{x},\mathbf{x})]}{n}} + \sqrt{\frac{\ln 1/\delta}{2n}}. \end{aligned}$$

This completes the proof. ■

The generalization bound in Theorem 4.10 resembles the standard result in learning theory except the second term on the r.h.s. which characterizes the Rademacher complexity of $\mathcal{F}_b$ over the space of probability distributions.

It is worth mentioning that similar result can be obtained in term of other complexity measures. For instance, we may consider a *fat-shattering dimension* of Bartlett and Shawe-Taylor (1999). Let consider the function class $\mathcal{F}_r = \{\boldsymbol{\mu}_\mathbb{P} \mapsto \langle \mathbf{w}, \boldsymbol{\mu}_\mathbb{P} \rangle : \|\mathbf{w}\|_\mathscr{H} \leq 1, \|\boldsymbol{\mu}_\mathbb{P}\|_\mathscr{H} \leq r\}$. It is not difficult to show that $\texttt{fat}_{\mathcal{F}_r}(\gamma) \leq (r/\gamma)^2$ where $\texttt{fat}_{\mathcal{F}_r}(\gamma)$ denotes the fat-shattering dimension of $\mathcal{F}_r$ which depends on the margin γ. It follows from Bartlett and Shawe-Taylor (1999) that there is a constant c such that w.p. $1 - \delta$ over n independent examples $\mathcal{S}$, if $h = \text{sign}(f) \in \text{sign}(\mathcal{F}_r)$ has margin at least γ on all the examples in $\mathcal{S}$, then the error of h is no more than $(c/n)((r^2/\gamma^2)\log^2 n + \log(1/\delta))$. Moreover, w.p. $1 - \delta$, every classifier $h \in \text{sign}(\mathcal{F}_r)$ has error no more than $s/n + \sqrt{(c/n)((r^2/\gamma^2)\log^2 n + \log(1/\delta))}$ where s is the number of laballed examples in $\mathcal{S}$ with margin less than γ. Unlike Rademacher complexity, the fat-shattering bounds also incorporate label information.

A more thorough analysis of distributional learning can be found in more recent works such as Lopez-Paz et al. (2015b) for classification setting and Szabó et al. (2015) for regression setting. In those works, the basic assumption is that we only have access to the sample sets $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)$ where $\mathbf{X}_i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i\}$, $\mathbf{x}_j^i \sim \mathbb{P}_i$.

4.7 Experimental Results

In the experiments, we primarily consider three different learning algorithms: i) **SVM** is considered as a baseline algorithm. ii) **Augmented SVM (ASVM)** is an SVM trained on augmented samples drawn according to the distributions $\{\mathbb{P}_i\}_{i=1}^m$. The same number of examples are drawn from each distribution. iii) **SMM** is distribution-based method that can be applied directly on the distributions.[1]

[1] We used the LIBSVM implementation.

Table 4.3: Accuracies (%) of SMM on synthetic data with different combinations of embedding and level-2 kernels.

		Embedding kernels				
		LIN	POLY2	POLY3	RBF	URBF
Level-2 kernels	LIN	85.20±2.20	81.04±3.11	81.10±2.76	87.74±2.19	85.39±2.56
	POLY	83.95±2.11	81.34±1.21	82.66±1.75	88.06±1.73	86.84±1.51
	RBF	87.80±1.96	73.12±3.29	78.28±2.19	**89.65±1.37**	86.86±1.88

4.7.1 Synthetic Data

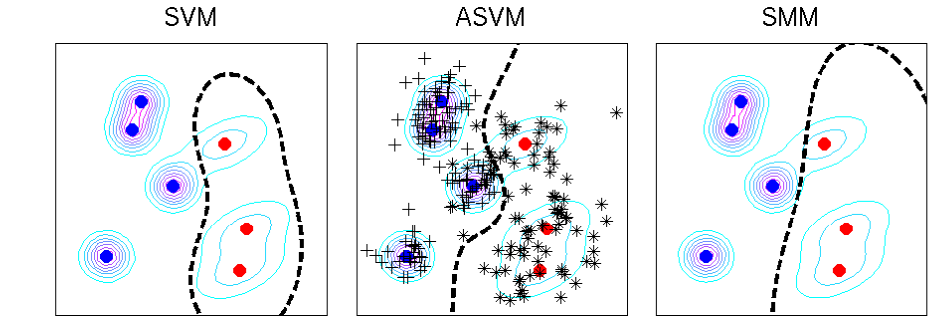
Firstly, we conducted a basic experiment that illustrates a fundamental difference between SVM, ASVM, and SMM. A binary classification problem of 7 Gaussian distributions with different means and covariances was considered. We trained the SVM using only the means of the distributions, ASVM with 30 virtual examples generated from each distribution, and SMM using distributions as training examples. A Gaussian RBF kernel with $\gamma = 0.25$ was used for all algorithms.

Figure 4.1a shows the resulting decision boundaries. Having been trained only on means of the distributions, the SVM classifier tends to overemphasize the regions with high densities and underrepresent the lower density regions. In contrast, the ASVM is more expensive and sensitive to outliers, especially when learning on heavy-tailed distributions. The SMM treats each distribution as a training example and implicitly incorporates properties of the distributions, *i.e.*, means and covariances, into the classifier. Note that the SVM can be trained to achieve a similar result to the SMM by choosing an appropriate value for γ (cf. Lemma 4.3). Nevertheless, this becomes more difficult if the training distributions are, for example, nonisotropic and have different covariance matrices.

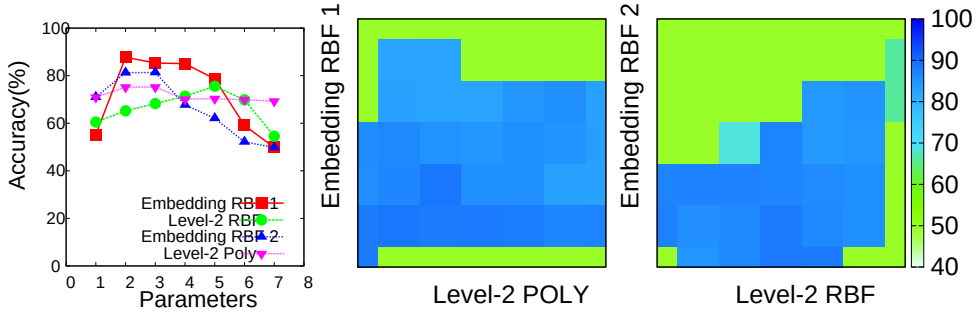
Secondly, we evaluate the performance of the SMM for different combinations of embedding and level-2 kernels. Two classes of synthetic Gaussian distributions on $\mathbb{R}^{10}$ were generated. The mean parameters of the positive and negative distributions are normally distributed with means $\mathbf{m}^+ = (1, \ldots, 1)$ and $\mathbf{m}^- = (2, \ldots, 2)$ and identical covariance matrix $\mathbf{\Sigma} = 0.5 \cdot \mathbf{I}_{10}$, respectively. The covariance matrix for each distribution is generated according to two Wishart distributions with covariance matrices given by $\mathbf{\Sigma}^+ = 0.6 \cdot \mathbf{I}_{10}$ and $\mathbf{\Sigma}^- = 1.2 \cdot \mathbf{I}_{10}$ with 10 degrees of freedom. The training set consists of 500 distributions from the positive class and 500 distributions from the negative class. The test set consists of 200 distributions with the same class proportion.

The kernels used in the experiment include linear kernel (LIN), polynomial kernel of degree 2 (POLY2), polynomial kernel of degree 3 (POLY3), unnormalized Gaussian RBF kernel (RBF), and normalized Gaussian RBF kernel (NRBF). To fix parameter values of both kernel functions and SMM, 10-fold cross-validation (10-CV) is performed on a parameter grid, $C \in \{2^{-3}, 2^{-2}, \ldots, 2^7\}$ for SMM, bandwidth parameter $\gamma \in \{10^{-3}, 10^{-2}, \ldots, 10^2\}$ for Gaussian RBF kernels, and degree parameter $d \in \{2, 3, 4, 5, 6\}$ for polynomial kernels. The average accuracy and ± 1 standard deviation for all kernel combinations over 30 repetitions are reported in Table 4.3. Moreover, we also investigate the sensitivity of kernel parameters for two kernel combinations: RBF-RBF and POLY-RBF. In this case, we consider the bandwidth parameter $\gamma = \{10^{-3}, 10^{-2}, \ldots, 10^3\}$ for Gaussian RBF kernels and degree parameter $d = \{2, 3, \ldots, 8\}$ for polynomial kernels. Figure 4.1b depicts the accuracy values and average accuracies for considered kernel functions.

Table 4.3 indicates that both embedding and level-2 kernels are important for the performance of the classifier. The embedding kernels tend to have more impact on the predictive



(a) decision boundaries.



(b) sensitivity of kernel parameters

Figure 4.1: (a) The decision boundaries of SVM, ASVM, and SMM. (b) the heatmap plots of average accuracies of SMM over 30 experiments using POLY-RBF (center) and RBF-RBF (right) kernel combinations with the plots of average accuracies at different parameter values (left).

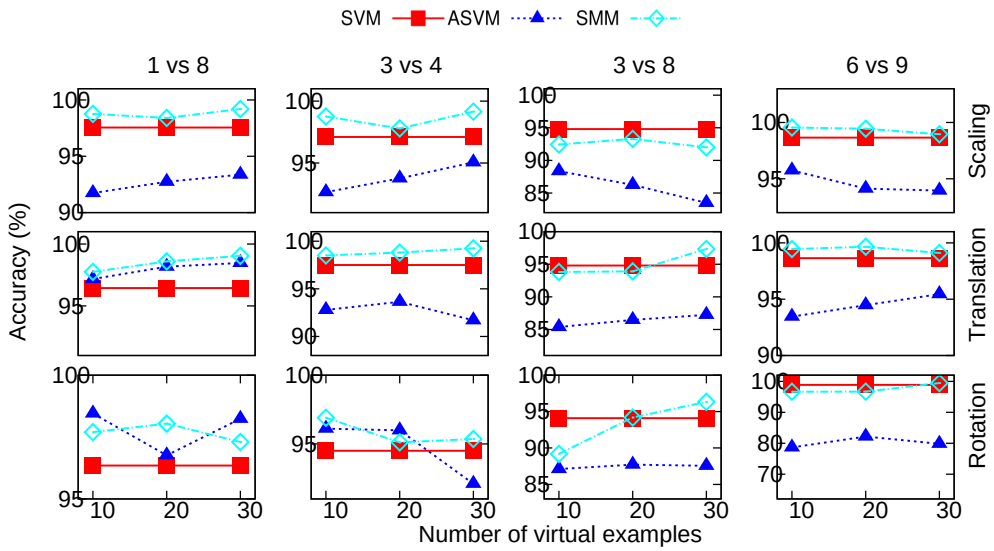


Figure 4.2: The performance of SVM, ASVM, and SMM algorithms on handwritten digits constructed using three basic transformations.

performance compared to the level-2 kernels. This conclusion also coincides with the results depicted in Figure 4.1b.

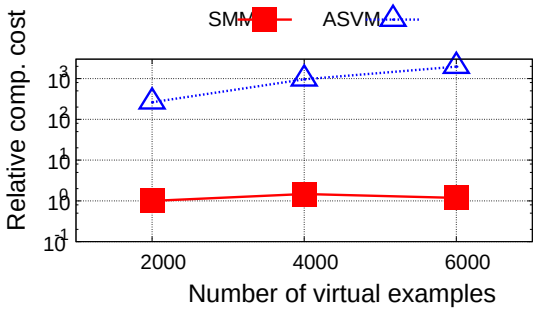


Figure 4.3: Relative computational cost of ASVM and SMM (baseline: SMM with 2000 virtual examples).

4.7.2 Handwritten Digit Recognition

In this section, the proposed framework is applied to distributions over equivalence classes of images that are invariant to basic transformations, namely, *scaling*, *translation*, and *rotation*. We consider the handwritten digits obtained from the USPS dataset. For each 16×16 image, the distribution over the equivalence class of the transformations is determined by a prior on parameters associated with such transformations. Scaling and translation are parametrized by the scale factors (s_x, s_y) and displacements (t_x, t_y) along the x and y axes, respectively. The rotation is parametrized by an angle θ. We adopt Gaussian distributions as prior distributions, including $\mathcal{N}([1,1], 0.1 \cdot \mathbf{I}_2)$, $\mathcal{N}([0,0], 5 \cdot \mathbf{I}_2)$, and $\mathcal{N}(0; \pi)$. For each image, the virtual examples are obtained by sampling parameter values from the distribution and applying the transformation accordingly.

Experiments are categorized into simple and difficult binary classification tasks. The former consists of classifying digit 1 against digit 8 and digit 3 against digit 4. The latter considers classifying digit 3 against digit 8 and digit 6 against digit 9. The initial dataset for each task is constructed by randomly selecting 100 examples from each class. Then, for each example in the initial dataset, we generate 10, 20, and 30 virtual examples using the aforementioned transformations to construct virtual data sets consisting of 2,000, 4,000, and 6,000 examples, respectively. One third of examples in the initial dataset are used as a test set. The original examples are excluded from the virtual datasets. The virtual examples are normalized such that their feature values are in $[0, 1]$. Then, to reduce computational cost, principle component analysis (PCA) is performed to reduce the dimensionality to 16. We compare the SVM on the initial dataset, the ASVM on the virtual datasets, and the SMM. For SVM and ASVM, the Gaussian RBF kernel is used. For SMM, we employ the empirical kernel (4.11) with Gaussian RBF kernel as a base kernel. The parameters of the algorithms are fixed by 10-CV over parameters $C \in \{2^{-3}, 2^{-2}, \ldots, 2^{7}\}$ and $\gamma \in \{0.01, 0.1, 1\}$.

The results depicted in Figure 4.2 clearly demonstrate the benefits of learning directly from the equivalence classes of digits under basic transformations.[2] In most cases, the SMM outperforms both the SVM and the ASVM as the number of virtual examples increases. Moreover, Figure 4.3 shows the benefit of the SMM over the ASVM in term of computational cost.[3]

[2]While the reported results were obtained using virtual examples with Gaussian parameter distributions (Sec. 4.7.2), we got similar results using uniform distributions.

[3]The evaluation was made on a 64-bit desktop computer with Intel® Core™ 2 Duo CPU E8400 at 3.00GHz×2 and 4GB of memory.

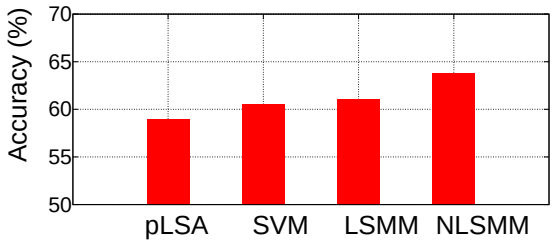


Figure 4.4: Accuracies of four different techniques for natural scene categorization.

4.7.3 Natural Scene Categorization

This section illustrates benefits of the nonlinear kernels between distributions for learning natural scene categories in which the bag-of-word (BoW) representation is used to represent images in the dataset. Each image is represented as a collection of local patches, each being a codeword from a large vocabulary of codewords called codebook. Standard BoW representations encode each image as a histogram that enumerates the occurrence probability of local patches detected in the image w.r.t. those in the codebook. On the other hand, our setting represents each image as a distribution over these codewords. Thus, images of different scenes tends to generate distinct set of patches. Based on this representation, both the histogram and the local patches can be used in our framework.

We use the dataset presented in Fei-fei (2005). According to their results, most errors occurs among the four indoor categories (830 images), namely, bedroom (174 images), living room (289 images), kitchen (151 images), and office (216 images). Therefore, we will focus on these four categories. For each category, we split the dataset randomly into two separate sets of images, 100 for training and the rest for testing.

A codebook is formed from the training images of all categories. Firstly, interesting keypoints in the image are randomly detected. Local patches are then generated accordingly. After patch detection, each patch is transformed into a 128-dim SIFT vector (Lowe 1999). Given the collection of detected patches, K-means clustering is performed over all local patches. Codewords are then defined as the centers of the learned clusters. Then, each patch in an image is mapped to a codeword and the image can be represented by the histogram of the codewords. In addition, we also have an $M \times 128$ matrix of SIFT vectors where M is the number of codewords.

We compare the performance of a Probabilistic Latent Semantic Analysis (pLSA) with the standard BoW representation, SVM, linear SMM (LSMM), and nonlinear SMM (NLSMM). For SMM, we use the empirical embedding kernel with Gaussian RBF base kernel k:

$$K(\mathbf{h}_i, \mathbf{h}_j) = \sum_{r=1}^{M} \sum_{s=1}^{M} h_i(c_r) h_j(c_s) k(c_r, c_s)$$

where $\mathbf{h}_i$ is the histogram of the ith image and c_r is the rth SIFT vector. A Gaussian RBF kernel is also used as the level-2 kernel for nonlinear SMM. For the SVM, we adopt a Gaussian RBF kernel with χ^2-distance between the histograms (Vedaldi et al. 2009), *i.e.*,

$$K(\mathbf{h}_i, \mathbf{h}_j) = \exp\left(-\gamma \chi^2(\mathbf{h}_i, \mathbf{h}_j)\right) \quad \text{where} \quad \chi^2(\mathbf{h}_i, \mathbf{h}_j) = \sum_{r=1}^{M} \frac{(h_i(c_r) - h_j(c_r))^2}{h_i(c_r) + h_j(c_r)}.$$

The parameters of the algorithms are fixed by 10-CV over parameters $C \in \{2^{-3}, 2^{-2}, \ldots, 2^7\}$ and $\gamma \in \{0.01, 0.1, 1\}$. For NLSMM, we use the best γ of LSMM in the base kernel and perform 10-CV to choose γ parameter only for the level-2 kernel. To deal with multiple categories, we

adopt the pairwise approach and voting scheme to categorize test images. The results in Figure 4.4 illustrate the benefit of the distribution-based framework. Understanding the context of a complex scene is challenging. Employing distribution-based methods provides an elegant way of utilizing higher-order statistics in natural images that could not be captured by traditional sample-based methods.

4.8 Discussions

This chapter proposes a method for kernel-based discriminative learning on probability distributions. The trick is to embed distributions into an RKHS, resulting in a simple and efficient learning algorithm on distributions. A family of linear and nonlinear kernels on distributions allows one to flexibly choose the kernel function that is suitable for the problems at hand. Our analyses provide insights into the relations between distribution-based methods and traditional sample-based methods, particularly the flexible SVM that allows the SVM to place different kernels on each training example. The experimental results illustrate the benefits of learning from a pool of distributions, compared to a pool of examples, both on synthetic and real-world data.

END OF CHAPTER 4

Chapter 5

Unsupervised Learning on Distributions

An equally important setting for machine learning is when label information are not available and we only observe unlabeled samples from distributions. In this chapter, I present an unsupervised learning framework on distributions.

5.1 Introduction

A technological advances have allowed many emerging scientific disciplines such as population genetics, flow cytometry, and astronomy to easily produce a tremendous amount of data. However, it is often the case that the acquisition of these data are subjected to different variations. For example, flow cytometry data obtained from different patients is expected to undergo biological variations. Likewise, in population genetics, populations of organisms are categorized according to their spatial distribution. In addition to the investigation of within-group variability, *e.g.*, random genetic variability, one may be interested in examining the divergence between groups in accordance with the underlying distributions, which may reveal collective behaviors of the data, *e.g.*, structured genetic variability. Consequently, there is a need for exploratory tool that is able to use data to access the properties of corresponding distributions.

Figure 5.1 illustrates the hierarchical data generating process that we typically encounter in practice. We are interested in using the sample $\mathbb{P}_1, \ldots, \mathbb{P}_\ell$ to unravel and study the properties of the distribution $\mathscr{P}^*$. Unfortunately, we only observe the samples $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$ where $\mathbf{X}_i = \{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)}\}$ and $\mathbf{x}_j^{(i)} \sim \mathbb{P}_i$. As a result, we must rely only on the sample set $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell\}$. Throughout this chapter, I argue that by using the kernel mean embedding constructed from $\mathbf{X}_i$ as a representation of distribution $\mathbb{P}_i$, we are able to make an inference on the properties of $\mathscr{P}^*$ to some extent. I demonstrate this aspect empirically via PCA on distributions, and then present some concrete applications as well as accompanying theoretical insights.

5.2 Distributional Principal Component Analysis

Principal component analysis (PCA) was first proposed by Pearson (1901) and its modern instantiation was later formalized by Hotelling (1933b). Since then, it has become an essential tool for multivariate data analysis and unsupervised dimensionality reduction (Jolliffe 1986, Burges 2010). The goal of PCA is to find a set of orthogonal components that best explains the variance of the observations. It can be shown mathematically that variance of the projected observations is maximized when these orthogonal components are the eigenvectors of covariance matrix between the observations (see *e.g.*, Section 2.2.3 and Jolliffe (1986) for a detailed treatment of PCA). Probabilistic PCA (Tipping and Bishop 1999) provides an alternative view of PCA as a maximum likelihood estimation that yields better interpretability. The success of PCA has encouraged many specialized extensions of PCA in many fields. For instance, robust PCA (De la Torre and Black 2001) was proposed to deal with outliers and becomes popular

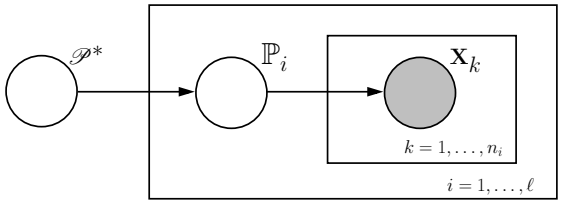


Figure 5.1: The graphical model describing the generative process of the framework considered in this work. The observations are sample sets whose members are drawn according to the the random distributions.

especially in computer vision (Torre and Black 2003). Functional PCA (Shang 2011, van der Linde 2008) was developed particularly for functional data such as time series (Ingrassia and Costanzo 2005) and functional magnetic resonance imaging (fMRI) (Viviani et al. 2005), and has become an essential tool in functional data analysis (FDA) (Ramsay and Silverman 2005). One of the disadvantages of PCA is that it can only discover the linear subspace. To deal with nonlinear features, nonlinear extensions of PCA have been introduced (Kramer 1991, Lawrence 2005, Scholz et al. 2005). In particular, Schölkopf et al. (1998) used the kernel trick for nonlinear PCA in which nonlinear features need not be computed explicitly. This advantage also leads to several specialized extension of kernel PCA (KPCA), including robust KPCA (Nguyen and la Torre 2009, Huang et al. 2009) and missing data in KPCA (Sanguinetti and Lawrence 2006).

In many situations, the data naturally fall into one of the multiple categories. In these cases, linear discriminant analysis (LDA) is often adopted (Fisher 1936, McLachlan 1992). The goal of LDA is to determine which variables discriminate between two or more naturally occurring groups. It is similar to PCA in the sense that both looks for linear combinations of variables that best explain the data. Unlike PCA, LDA is a supervised technique which also takes class memberships into account. Mathematically speaking, LDA finds a linear subspace which maximizes the between-class scattering of projected data, while minimizing their within-class scattering. For binary categories, LDA often relies on the assumption that class conditional probability density functions are both normally distributed and class covariances are identical, *i.e.*, homoscedastic assumption. In multiclass LDA, the sample covariance of the class means is used to measure between-class scattering. Therefore, most higher-order statistics, *e.g.*, class covariance and tensors, arising from the collective behaviors of the data are neglected by LDA. We, on the other hand, aim to incorporate these higher-order statistics by mean of PCA on probability distributions. I call it a *distributional PCA* (DPCA) which amounts to solving KPCA with inner product $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$. In addition to an exploratory aspect, understanding the consequences of DPCA will offer a new spectrum of algorithms in machine learning for multi-source learning (patient-hospital), domain adaptation (finding a low-dimensional subspace of domains), and multi-task or transfer learning.

5.2.1 Analysis of Kernel Mean Representation

In exploratory data analysis, it is important to understand what kind of information is captured by the proposed representation. To simplify the analysis, I will focus on a family of *shift-invariant kernels* $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x} - \mathbf{x}')$. By Bochner's theorem (Theorem 2.2), we know that ψ is a Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$. I assume that the measure $\mathrm{d}\Lambda(\boldsymbol{\omega})$ can be represented by a density $\widehat{\Psi}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega}$. The function $\widehat{\Psi}(\boldsymbol{\omega})$ is called the *spectrum* and can be computed as the inverse Fourier transform of $\psi(\mathbf{x})$ given by

$$\widehat{\Psi}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle \mathbf{x}, \boldsymbol{\omega} \rangle} \psi(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

If the kernel k is properly scaled, Bochner's theorem guarantees that its Fourier transform is a proper probability distribution.

The following theorem related the kernel $\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ to the inner product between their characteristic functions $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$ in $L^2(\mathbb{R}^d, \Lambda)$.

Theorem 5.1. *Let $\mathscr{P}$ denote the set of all Borel probability measures on $\mathbb{R}^d$ and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded, measurable, and shift-invariant kernel endowed with a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$. For any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$ whose characteristic functions are $\varphi_{\mathbb{P}}$ and $\varphi_{\mathbb{Q}}$, respectively, the following equality holds:*

$$K(\mathbb{P}, \mathbb{Q}) = \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} = \langle \varphi_{\mathbb{P}}, \varphi_{\mathbb{Q}} \rangle_{L^2(\mathbb{R}^d, \Lambda)} \tag{5.1}$$

for some finite nonnegative Borel measure Λ on $\mathbb{R}^d$.

Proof of Theorem 5.1. Since k is bounded, $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[k(\mathbf{x}, \cdot)] < \infty$ and is well-defined for any $\mathbb{P} \in \mathscr{P}$. Then, by definition, we have

$$\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} = \iint k(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}')$$

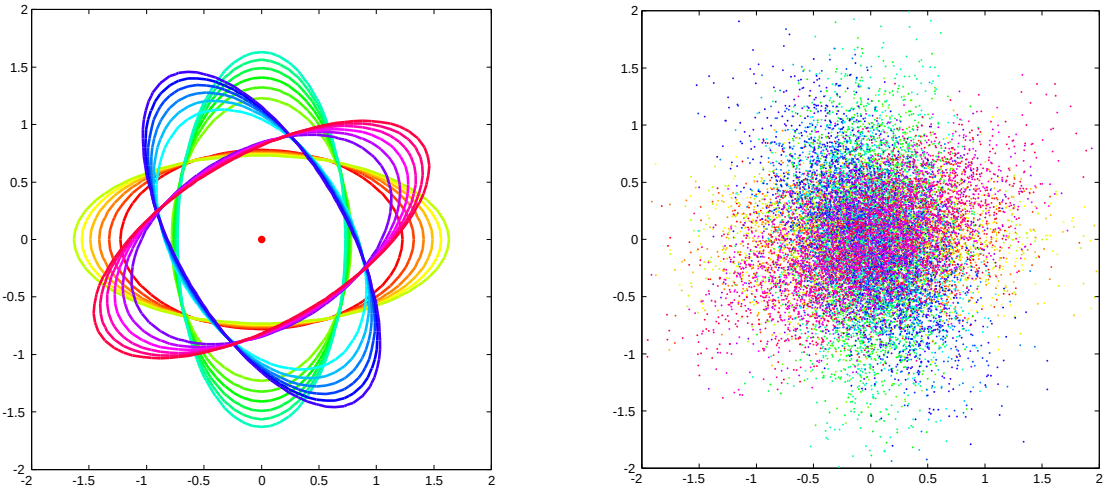
Thus, it is trivial that $K : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ is a positive semidefinite kernel on $\mathscr{P}$. Since k is shift-invariant, it follows from Bochner's theorem (cf. Theorem 2.2) that,

$$\begin{aligned}
\iint k(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') &= \iint \psi(\mathbf{x} - \mathbf{x}') \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \\
&= \iiint e^{-i\langle \mathbf{x} - \mathbf{x}', \omega \rangle} \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \\
&= \iiint e^{-i\langle \mathbf{x}, \boldsymbol{\omega} \rangle} \cdot e^{i\langle \mathbf{x}', \boldsymbol{\omega} \rangle} \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \\
&= \iiint e^{-i\langle \mathbf{x}, \boldsymbol{\omega} \rangle} \cdot e^{i\langle \mathbf{x}', \boldsymbol{\omega} \rangle} \, \mathrm{d}\mathbb{P}(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= \int \left[\int e^{-i\langle \mathbf{x}, \boldsymbol{\omega} \rangle} \, \mathrm{d}\mathbb{P}(\mathbf{x}) \right] \cdot \left[\int e^{i\langle \mathbf{x}', \boldsymbol{\omega} \rangle} \, \mathrm{d}\mathbb{Q}(\mathbf{x}') \right] \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= \int \overline{\varphi_{\mathbb{P}}(\boldsymbol{\omega})} \cdot \varphi_{\mathbb{Q}}(\boldsymbol{\omega}) \, \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= \langle \varphi_{\mathbb{P}}, \varphi_{\mathbb{Q}} \rangle_{L^2(\mathbb{R}^d, \Lambda)}
\end{aligned}$$

thus completing the proof. ∎

Theorem 5.1 implies that $K(\mathbb{P}, \mathbb{Q}) = \langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ can be equivalently written as the L^2-inner product between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the nonnegative finite Borel measure Λ which is the Fourier transform of ψ. Strictly speaking, the kernel $K(\mathbb{P}, \mathbb{Q})$ is the generalization of the Hermitian inner product w.r.t. a measure Λ such that the integral matters only on sets with positive measure.

If $\operatorname{supp}(\Lambda)$ is $\mathbb{R}^d$, then the integral is defined everywhere and there is no loss of information. As a result, the map (4.7) is injective (Sriperumbudur et al. 2008; 2010). On the other hand, we may consider the non-characteristic kernel functions. In such cases, L^2 space consists of equivalence classes of characteristic functions. Two characteristic functions represent the same L^2 function if the set where they differ has measure zero. Consequently, it is possible for DPCA to completely neglect certain properties of the distributions by choosing the appropriate kernel k accordingly. Moreover, instead of working with individual probability distributions, we can



(a) **(b)**

Figure 5.2: (a) The synthetic Gaussian distributions with identical mean and varying covariance matrices. b the sample drawn according to the synthetic Gaussian distributions.

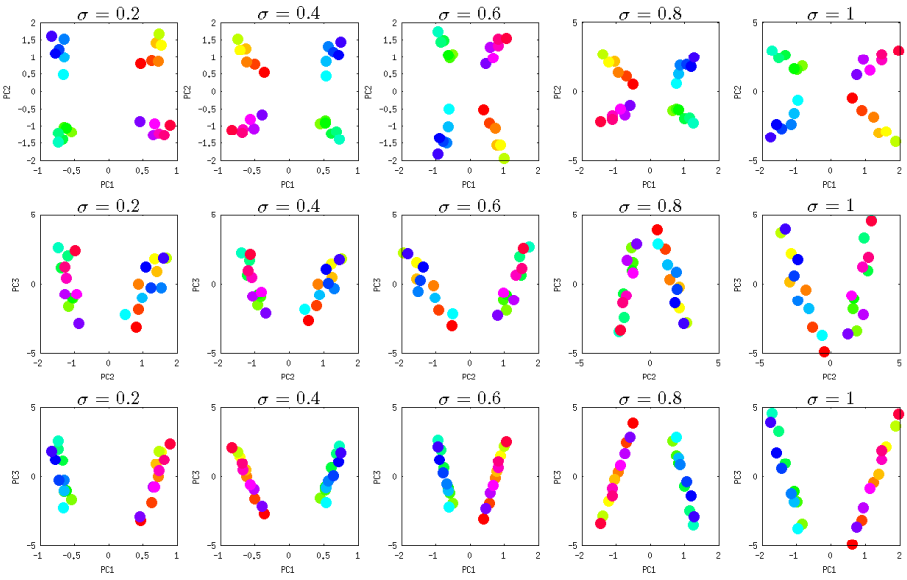


Figure 5.3: The projection of data onto the first three principal components. Each point and its color in the plot corresponds to the distribution shown in Figure 5.2a.

deal with the equivalence classes of probability distributions whose characteristic functions are the same on the set where Λ has positive measure.

To illustrate this, I conduct a simple experiment which involves principal component analysis on distributions. Figure 5.2 depicts synthetic data. The data set consists of 24 zero-mean Gaussian distributions whose covariance matrices are different as illustrated in Figure 5.2a. The samples from these distributions are depicted in Figure 5.2b. Based on these samples, I then perform the KPCA using the inner product $\langle \hat{\boldsymbol{\mu}}_{\mathbb{P}}, \hat{\boldsymbol{\mu}}_{\mathbb{Q}} \rangle_{\mathscr{H}}$ with Gaussian RBF kernel. The results are shown in Figure 5.3 and 5.4 using different bandwidth parameter σ. As we can see, the results reflect the similarity between distributions observed in Figure 5.2a.

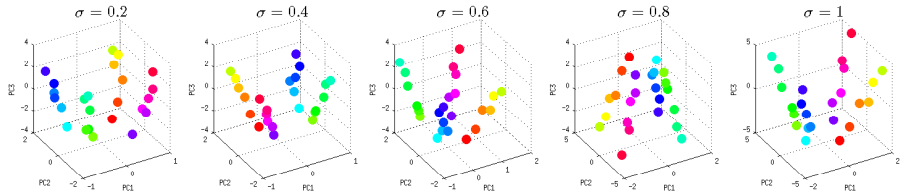


Figure 5.4: Same as Figure 5.3, but visualize the projection on the first three principal components simultaneously.

In the subsequent sections, I present some concrete examples of unsupervised learning algorithms on distributions.

5.3 One-Class Support Measure Machines

Anomaly detection is one of the most important tools in all data-driven scientific disciplines. Data that do not conform to the expected behaviors often bear some interesting characteristics and can help domain experts better understand the problem at hand. However, in the era of data explosion, the anomaly may appear not only in the data themselves, but also as a result of their interactions. The main objective of this paper is to investigate the latter type of anomalies. To be consistent with the previous works (Póczos et al. 2011, Xiong et al. 2011b;a), we will refer to this problem as a group anomaly detection, as opposed to a traditional point anomaly detection.

Like traditional point anomaly detection, the group anomaly detection refers to a problem of finding patterns in groups of data that do not conform to expected behaviors (Póczos et al. 2011, Xiong et al. 2011b;a). That is, an ultimate goal is to detect interesting aggregate behaviors of data points among several groups. In principle, anomalous groups may consist of individually anomalous points, which are relatively easy to detect. On the other hand, anomalous groups of relatively normal points, whose behavior as a group is unusual, is much more difficult to detect. In this work, we are interested in the latter type of group anomalies. Figure 5.5 illustrates this scenario.

Group anomaly detection may shed light in a wide range of applications. For example, a Sloan Digital Sky Survey (SDSS) has produced a tremendous amount of astronomical data. It is therefore very crucial to detect rare objects such as stars, galaxies, or quasars that might lead to a scientific discovery. In addition to individual celestial objects, investigating groups of them may help astronomers understand the universe on larger scales. For instance, the anomalous group of galaxies, which is the smallest aggregates of galaxies, may reveal interesting phenomena, *e.g.*, the gravitational interactions of galaxies.

Likewise, a new physical phenomena in high energy particle physics such as Higgs boson appear as a tiny excesses of certain types of collision events among a vast background of known physics in particle detectors (Bhat 2011, Vatanen et al. 2012). Investigating each collision event individually is no longer sufficient as the individual events may not be anomalies by themselves, but their occurrence together as a group is anomalous. Hence, we need a powerful algorithm to detect such a rare and highly structured anomaly.

Lastly, the algorithm proposed in this paper can be applied to point anomaly detection with substantial and heterogeneous uncertainties. For example, it is often costly and time-consuming to obtain the full spectra of astronomical objects. Instead, relatively noisier measurements are usually made. In addition, the estimated uncertainty which represents the uncertainty one would obtain from multiple observations is also available. Incorporating these uncertainties has been shown to improve the performance of the learning systems (Kirkpatrick et al. 2011, Bovy et al.

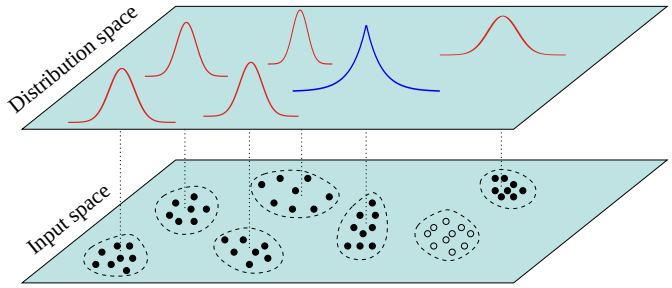


Figure 5.5: An illustration of two types of group anomalies. An anomalous group may be a group of anomalous samples which is easy to detect (unfilled points). In this paper, we are interested in detecting anomalous groups of normal samples (filled points) which is more difficult to detect because of the higher-order statistics. Note that group anomaly we are interested in can only be observed in the space of distributions.

2011, Ross et al. 2012).

The anomaly detection has been intensively studied (Chandola et al. (2009) and references therein). However, few attempts have been made on developing successful group anomaly detection algorithms. For example, a straightforward approach is to define a set of features for each group and apply standard point anomaly detection (Chan and Mahoney 2005). Despite its simplicity, this approach requires a specific domain knowledge to construct appropriate sets of features. Another possibility is to first identify the individually anomalous points and then find their aggregations (Das et al. 2008). Again, this approach relies only on the detection of anomalous points and thus cannot find the anomalous groups in which their members are perfectly normal. Successful group anomaly detectors should be able to incorporate the higher-order statistics of the groups.

Recently, a family of hierarchical probabilistic models based on a Latent Dirichlet Allocation (LDA) (Blei et al. 2001) has been proposed to cope with both types of group anomalies (Xiong et al. 2011b;a). In these models, the data points in each group are assumed to be one of the K different types and generated by a mixture of K Gaussian distributions. Although the distributions over these K types can vary across M groups, they share common generator. The groups that have small probabilities under the model are marked as anomalies using scoring criteria defined as a combination of a point-based anomaly score and a group-based anomaly score. The Flexible Genre Model (FGM) recently extends this idea to model more complex group structures (Xiong et al. 2011a).

Instead of employing a generative approach, we propose a simple and efficient discriminative way of detecting group anomaly. In this work, M groups of data points are represented by a set of M probability distributions assumed to be i.i.d. realization of some unknown distribution $\mathscr{P}$. In practice, only i.i.d samples from these distributions are observed. Hence, we can treat group anomaly detection as detecting the anomalous distributions based on their empirical samples. To allow for a practical algorithm, the distributions are mapped into the RKHS using the kernel mean embedding. By working directly with the distributions, the higher-order information arising from the aggregate behaviors of the data points can be incorporated efficiently.

5.3.1 Quantile Estimation on Probability Distributions

Let $\mathcal{X}$ denote a non-empty input space with associated σ-algebra $\mathcal{A}$, $\mathbb{P}$ denote the probability distribution on $(\mathcal{X}, \mathcal{A})$, and $\mathscr{P}$ denote the set of all probability distributions on $(\mathcal{X}, \mathcal{A})$. The space $\mathscr{P}$ is endowed with the topology of weak convergence and the associated Borel σ-algebra.

We assume that there exists a distribution $\mathcal{P}$ on $\mathscr{P}$, where $\mathbb{P}_1, \ldots, \mathbb{P}_\ell$ are i.i.d. realizations

from $\mathcal{P}$, and the sample S_i is made of n_i i.i.d. samples distributed according to the distribution $\mathbb{P}_i$. In this work, we observe ℓ samples $S_i = \{\mathbf{x}_k^{(i)}\}_{1 \leq k \leq n_i}$ for $i = 1, \ldots, \ell$. For each sample S_i, $\widehat{\mathbb{P}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{\mathbf{x}_j^{(i)}}$ is the associated empirical distribution of $\mathbb{P}_i$.

In this section, we formulate a group anomaly detection problem as learning quantile function $q : \mathscr{P} \to \mathbb{R}$ to estimate the support of $\mathcal{P}$. Let $\mathcal{C}$ be a class of measurable subsets of $\mathscr{P}$ and λ be a real-valued function defined on $\mathcal{C}$, the quantile function w.r.t. $(\mathcal{P}, \mathcal{C}, \lambda)$ is

$$q(\beta) = \inf\{\lambda(C) : \mathcal{P}(C) \geq \beta, C \in \mathcal{C}\} \ ,$$

where $0 < \beta \leq 1$. In this paper, we consider when λ is Lebesgue measure, in which case $C(\beta)$ is the minimum volume $C \in \mathcal{C}$ that contains at least a fraction β of the probability mass of $\mathcal{P}$. Thus, the function q can be used to test if any test distribution $\mathbb{P}_t$ is anomalous w.r.t. the training distributions.

Rather than estimating $C(\beta)$ in the space of distributions directly, we first map the distributions into a feature space via a positive semi-definite kernel k. Our class $\mathcal{C}$ is then implicitly defined as the set of half-spaces in the feature space. Specifically, $C_{\mathbf{w}} = \{\mathbb{P} \mid f_{\mathbf{w}}(\mathbb{P}) \geq \rho\}$ where $(\mathbf{w}, \rho)$ are respectively a weight vector and an offset parametrizing a hyperplane in the feature space associated with the kernel k. The optimal $(\mathbf{w}, \rho)$ is obtained by minimizing a regularizer which controls the smoothness of the estimated function describing C.

5.3.2 OCSMM Formulation

Our approach is in line with previous attempts in group anomaly detection that find a set of appropriate features for each group. On the one hand, however, the mean embedding approach captures all necessary information about the groups without relying heavily on a specific domain knowledge. On the other hand, it is flexible to choose the feature representation that is suitable to the problem at hand via the choice of the kernel k.

Using the mean embedding representation (4.7), the primal optimization problem for one-class SMM can be subsequently formulated in an analogous way to the one-class SVM (Schölkopf et al. 2001b) as follow:

$$\underset{\mathbf{w}, b, \xi, \rho}{\text{minimize}} \qquad \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle_{\mathscr{H}} - \rho + \frac{1}{\nu \ell} \sum_{i=1}^{\ell} \xi_i \tag{5.2a}$$

$$\text{subject to} \qquad \langle \mathbf{w}, \boldsymbol{\mu}_{\mathbb{P}_i} \rangle_{\mathscr{H}} \geq \rho - \xi_i, \xi_i \geq 0 \tag{5.2b}$$

where ξ_i denote slack variables and $\nu \in (0, 1]$ is a trade-off parameter corresponding to an expected fraction of outliers within the feature space. The trade-off ν is an upper bound on the fraction of outliers and lower bound on the fraction of support measures (Schölkopf et al. 2001b).

The trade-off parameter ν plays an important role in group anomaly detection. Small ν implies that anomalous groups are rare compared to the normal groups. Too small ν leads to some anomalous groups being rejected. On the other hand, large ν implies that anomalous groups are common. Too large ν leads to some normal groups being accepted as anomaly. As group anomaly is subtle, one need to choose ν very carefully to reduce the effort in the interpretation of the results.

By introducing Lagrange multipliers $\boldsymbol{\alpha}$, we have $\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\mu}_{\mathbb{P}_i} = \sum_{i=1}^{\ell} \alpha_i \mathbb{E}_{\mathbb{P}_i}[k(\mathbf{x}, \cdot)]$ and the dual form of (5.2) can be written as

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \qquad \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j \langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle_{\mathscr{H}} \tag{5.3a}$$

$$\text{subject to} \qquad 0 \leq \alpha_i \leq \frac{1}{\nu\ell},\ \sum_{i=1}^{\ell} \alpha_i = 1\,. \tag{5.3b}$$

Note that the dual form is a quadratic programming and depends on the inner product $\langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle_{\mathscr{H}}$. Given that we can compute $\langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle_{\mathscr{H}}$, we can employ the standard QP solvers to solve (5.3).

From (5.3), we can see that $\boldsymbol{\mu}_{\mathbb{P}}$ is a feature map associated with the kernel $K : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$, defined as $K(\mathbb{P}_i, \mathbb{P}_j) = \langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle_{\mathscr{H}}$. It follows from Fubini's theorem and reproducing property of $\mathscr{H}$ that

$$\begin{aligned} \langle \boldsymbol{\mu}_{\mathbb{P}_i}, \boldsymbol{\mu}_{\mathbb{P}_j} \rangle_{\mathscr{H}} &= \iint \langle k(\mathbf{x}, \cdot), k(\mathbf{y}, \cdot) \rangle_{\mathscr{H}} \, \mathrm{d}\mathbb{P}_i(\mathbf{x}) \, \mathrm{d}\mathbb{P}_j(\mathbf{y}) \\ &= \iint k(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbb{P}_i(\mathbf{x}) \, \mathrm{d}\mathbb{P}_j(\mathbf{y}) \ . \end{aligned} \tag{5.4}$$

Hence, K is a positive definite kernel on $\mathscr{P}$. Given the sample sets $S_1, \ldots, S_\ell$, one can estimate (5.10) by

$$K(\widehat{\mathbb{P}}_i, \widehat{\mathbb{P}}_j) = \frac{1}{n_i \cdot n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} k(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) \tag{5.5}$$

where $\mathbf{x}_k^{(i)} \in S_i$, $\mathbf{x}_l^{(j)} \in S_j$, and n_i is the number of samples in S_i for $i = 1, \ldots, \ell$.

Previous works in kernel-based anomaly detection have shown that the Gaussian RBF kernel is more suitable than some other kernels such as polynomial kernels (Hoffmann 2007). Thus we will focus primarily on the Gaussian RBF kernel given by

$$k_\sigma(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{5.6}$$

where $\sigma > 0$ is a bandwidth parameter. In the sequel, we denote the RKHS associated with kernel k_σ by $\mathscr{H}_\sigma$. Also, let $\phi : \mathcal{X} \to \mathscr{H}_\sigma$ be a feature map such that $k_\sigma(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathscr{H}_\sigma}$.

In group anomaly detection, we always observe the i.i.d. samples from the distribution underlying the group. Thus, it is natural to use the empirical kernel (5.5). However, one may relax this assumption and apply the kernel (5.10) directly. For instance, if we have a Gaussian distribution $\mathbb{P}_i = \mathcal{N}(\mathbf{m}_i, \boldsymbol{\Sigma}_i)$ and a Gaussian RBF kernel k_σ, we can compute the kernel analytically by

$$K(\mathbb{P}_i, \mathbb{P}_j) = \frac{\exp\left(-\frac{1}{2}(\mathbf{m}_i - \mathbf{m}_j)^\mathsf{T} \mathbf{B}^{-1} (\mathbf{m}_i - \mathbf{m}_j)\right)}{|\frac{1}{\sigma^2}\boldsymbol{\Sigma}_i + \frac{1}{\sigma^2}\boldsymbol{\Sigma}_j + \mathbf{I}|^{\frac{1}{2}}} \tag{5.7}$$

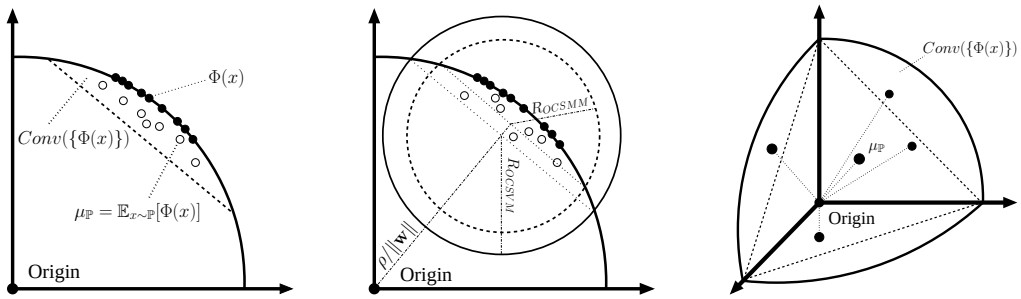
where $\mathbf{B} = \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j + \sigma^2 \mathbf{I}$. This kernel is particularly useful when one want to incorporate the point-wise uncertainty of the observation into the learning algorithm (Muandet et al. 2012).

5.3.3 Geometric Interpretation

For translation-invariant kernel, $k(\mathbf{x}, \mathbf{x})$ is constant for all $\mathbf{x} \in \mathcal{X}$. That is, $\|\phi(\mathbf{x})\|_{\mathscr{H}} = \tau$ for some constant ρ. This implies that all of the images $\phi(\mathbf{x})$ lie on the sphere in the feature space (cf. Figure 5.6a). Consequently, the following inequality holds

$$\|\boldsymbol{\mu}_{\mathbb{P}}\|_{\mathscr{H}} = \left\| \int k(\mathbf{x}, \cdot) \, \mathrm{d}\mathbb{P}(\mathbf{x}) \right\|_{\mathscr{H}} \leq \int \|k(\mathbf{x}, \cdot)\|_{\mathscr{H}} \, \mathrm{d}\mathbb{P}(\mathbf{x}) = \tau\,,$$

which shows that all mean embeddings lie inside the sphere (cf. Figure 5.6a). As a result, we can establish the existence and uniqueness of the separating hyperplane $\mathbf{w}$ in (5.2) through the following theorem.



(a) feature map and mean map **(b)** minimum enclosing sphere **(c)** spherical normalization

Figure 5.6: (a) The two dimensional representation of the RKHS of Gaussian RBF kernels. Since the kernels depend only on $\mathbf{x} - \mathbf{x}'$, $k(\mathbf{x}, \mathbf{x})$ is constant. Therefore, all feature maps $\phi(\mathbf{x})$ (black dots) lie on a sphere in feature space. Hence, for any probability distribution $\mathbb{P}$, its mean embedding $\boldsymbol{\mu}_{\mathbb{P}}$ always lies in the convex hull of the feature maps, which in this case, forms a segment of the sphere. (b) In general, the solution of OCSMM is different from the minimum enclosing sphere. (c) Three dimensional sphere in the feature space. For the Gaussian RBF kernel, the kernel mean embeddings of all distributions always lie inside the segment of the sphere. In addition, the angle between any pair of mean embeddings is always greater than zero. Consequently, the mean embeddings can be scaled, *e.g.*, to lie on the sphere, and the map is still injective.

Theorem 5.2. *There exists a unique separating hyperplane* $\boldsymbol{w}$ *as a solution to* (5.2) *that separates* $\boldsymbol{\mu}_{\mathbb{P}_1}, \boldsymbol{\mu}_{\mathbb{P}_2}, \ldots, \boldsymbol{\mu}_{\mathbb{P}_\ell}$ *from the origin.*

Proof. Due to the separability of the feature maps $\phi(\mathbf{x})$, the convex hull of the mean embeddings $\boldsymbol{\mu}_{\mathbb{P}_1}, \boldsymbol{\mu}_{\mathbb{P}_2}, \ldots, \boldsymbol{\mu}_{\mathbb{P}_\ell}$ does not contain the origin. The existence and uniqueness of the hyperplane then follows from the supporting hyperplane theorem (Schölkopf and Smola 2001). ■

By Theorem 5.2, the OCSMM is a simple generalization of OCSVM to the space of probability distributions. Furthermore, the straightforward generalization will allow for a direct application of an efficient learning algorithm as well as existing theoretical results.

There is a well-known connection between the solution of OCSVM with translation invariant kernels and the center of the minimum enclosing sphere (MES) (Tax and Duin 1999; 2004). Intuitively, this is not the case for OCSMM, even when the kernel k is translation-invariant, as illustrated in Figure 5.6b. Fortunately, the connection between OCSMM and MES can be made precise by applying the spherical normalization

$$\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}} \longmapsto \frac{\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}}{\sqrt{\langle \boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{P}} \rangle_{\mathscr{H}} \langle \boldsymbol{\mu}_{\mathbb{Q}}, \boldsymbol{\mu}_{\mathbb{Q}} \rangle_{\mathscr{H}}}} \tag{5.8}$$

After the normalization, $\|\boldsymbol{\mu}_{\mathbb{P}}\|_{\mathscr{H}} = 1$ for all $\mathbb{P} \in \mathscr{P}$. That is, all mean embeddings lie on the unit sphere in the feature space. Consequently, the OCSMM and MES are equivalent after the normalization.

Given the equivalence between OCSMM and MES, it is natural to ask if the spherical normalization (5.8) preserves the injectivity of the Hilbert space embedding. In other words, is there an information loss after the normalization? The following theorem answers this question for kernel k that satisfies some reasonable assumptions.

Theorem 5.3. *Assume that* k *is characteristic and the samples are linearly independent in the feature space* $\mathscr{H}$*. Then, the spherical normalization preserves the injectivity of the mapping* $\boldsymbol{\mu} : \mathscr{P} \to \mathscr{H}$*.*

Proof. Let us assume the normalization does not preserve the injectivity of the mapping. Thus, there exist two distinct probability distributions $\mathbb{P}$ and $\mathbb{Q}$ for which

$$\begin{aligned} \boldsymbol{\mu}_{\mathbb{P}} &= \boldsymbol{\mu}_{\mathbb{Q}} \\ \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{P}(\mathbf{x}) &= \int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbb{Q}(\mathbf{x}) \\ \int k(\mathbf{x}, \cdot)\, \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{x}) &= 0 \ . \end{aligned}$$

As $\mathbb{P} \neq \mathbb{Q}$, the last equality holds if and only if there exists $\mathbf{x} \in \mathcal{X}$ for which $k(\mathbf{x}, \cdot)$ are linearly dependent, which contradicts the assumption. Consequently, the spherical normalization must preserve the injectivity of the mapping. ■

The Gaussian RBF kernel satisfies the assumption given in Theorem 5.3 as the kernel matrix will be full-rank and thereby the samples are linearly independent in the feature space. Figure 5.6c depicts an effect of the spherical normalization.

It is important to note that the spherical normalization does not necessarily improve the performance of the OCSMM. It ensures that all the information about the distributions are preserved.

5.3.4 OCSMM and Kernel Density Estimation

In this section we make a connection between the OCSMM and kernel density estimation (KDE). First, we give a definition of the KDE. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be an i.i.d. samples from some distribution F with unknown density f, the KDE of f is defined as

$$\hat{f}(\mathbf{y}) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{\mathbf{y} - \mathbf{x}_i}{h}\right) \tag{5.9}$$

For $\hat{f}$ to be a density, we require that the kernel satisfies $k(\cdot, \cdot) \geq 0$ and $\int k(\mathbf{x}, \cdot)\, \mathrm{d}\mathbf{x} = 1$, which includes, for example, the Gaussian kernel, the multivariate Student kernel, and the Laplacian kernel.

When $\nu = 1$, it is well-known that, under some technical assumptions, the OCSVM corresponds exactly to the KDE (Schölkopf et al. 2001b). That is, the solution $\mathbf{w}$ of (5.2) can be written as a uniform sum over training samples similar to (5.9). Moreover, setting $\nu < 1$ yields a sparse representation where the summand consists of only support vectors of the OCSVM.

Interestingly, we can make a similar correspondence between the KDE and the OCSMM. It follows from Lemma 4.3 (cf. Muandet et al. (2012; Lemma 4)) that for certain classes of training probability distributions, the OCSMM on these distributions corresponds to the OCSVM on some training samples equipped with an appropriate kernel function. To understand this connection, consider the OCSMM with the Gaussian RBF kernel k_σ and isotropic Gaussian distributions $\mathcal{N}(m_1; \sigma_1^2), \mathcal{N}(m_2; \sigma_2^2), \ldots, \mathcal{N}(m_n; \sigma_n^2)$.[1] We analyze this scenario under two conditions:

(C1) Identical bandwidth. If $\sigma_i = \sigma_j$ for all $1 \leq i, j \leq n$, the OCSMM is equivalent to the OCSVM on the training samples $m_1, m_2, \ldots, m_n$ with Gaussian RBF kernel $k_{\sigma^2 + \sigma_i^2}$ (cf. the kernel (5.7)). Hence, the OCSMM corresponds to the OCSVM on the means of the distributions with kernel of larger bandwidth.

[1] We adopt the Gaussian distributions here for the sake of simplicity. More general statement for non-Gaussian distributions follows straightforwardly.

(C2) Variable bandwidth. Similarly, if $\sigma_i \neq \sigma_j$ for some $1 \leq i, j \leq n$, the OCSMM is equivalent to the OCSVM on the training samples $m_1, m_2, \ldots, m_n$ with Gaussian RBF kernel $k_{\sigma^2+\sigma_i^2}$. Note that the kernel bandwidth may be different at each training samples. Thus, OCSMM in this case corresponds to the OCSVM with variable bandwidth parameters.

On the one hand, the above scenario allows the OCSVM to cope with noisy/uncertain inputs, leading to more robust point anomaly detection algorithm. That is, we can treat the means as the measurements and the covariances as the measurement uncertainties (cf. Section 5.3.5). On the other hand, one can also interpret the OCSMM when $\nu = 1$ as a generalization of traditional KDE, where we have a data-dependent bandwidth at each data point. This type of KDE is known in the statistics as variable kernel density estimators (VKDEs) (Breiman et al. 1977, Abramson 1982, Terrell and Scott 1992). For $\nu < 1$, the OCSMM gives a sparse representation of the VKDE.

Formally, the VKDE is characterized by (5.9) with an adaptive bandwidth $h(\mathbf{x}_i)$. For example, the bandwidth is adapted to be larger where the data are less dense, with the aim to reduce the bias. There are basically two different views of VKDE. The first is known as a *balloon estimator* (Terrell and Scott 1992). Essentially, its bandwidth may depend only on the point at which the estimate is taken, *i.e.*, the bandwidth in (5.9) may be written as $h(\mathbf{y})$. The second type of VKDE is a *sample smoothing estimator* (Terrell and Scott 1992). As opposed to the balloon estimator, it is a mixture of individually scaled kernels centered at each observation, *i.e.*, the bandwidth is $h(\mathbf{x}_i)$. The advantage of balloon estimator is that it has a straightforward asymptotic analysis, but the final estimator may not be a density. The sample smoothing estimator is a density if k is a density, but exhibits *non-locality*.

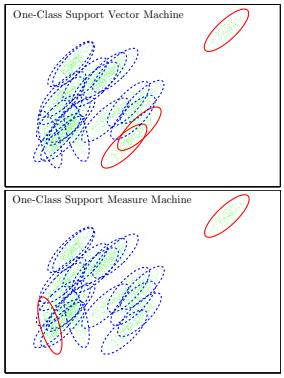
Both types of the VKDEs may be seen from the OCSMM point of view. Firstly, under the condition **(C1)**, the balloon estimator can be recovered by considering different test distribution $\mathbb{P}_t = \mathcal{N}(m_t; \sigma_t)$. As $\sigma_t \to 0$, one obtain the standard KDE on m_t. Similarly, the OCSMM under the condition **(C2)** with $\mathbb{P}_t = \delta_{m_t}$ gives the sample smoothing estimator. Interestingly, the OCSMM under the condition **(C2)** with $\mathbb{P}_t = \mathcal{N}(m_t; \sigma_t)$ results in a combination of these two types of the VKDEs.

In summary, we show that many variants of KDE can be seen as solutions to the regularization functional (5.2), and thereby provides an insight into a connection between large-margin approach and kernel density estimation.

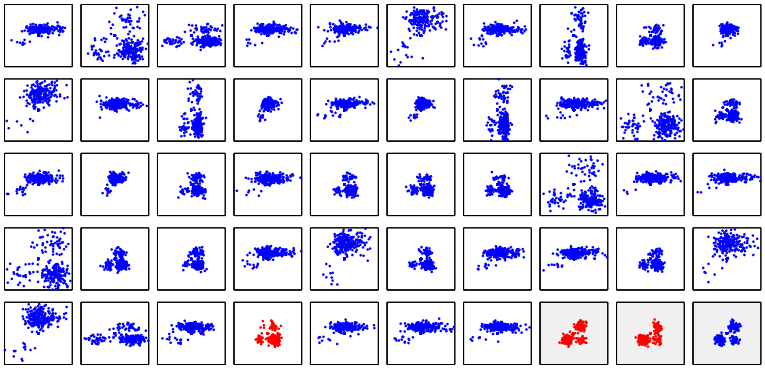
5.3.5 Experimental Results

We firstly illustrate a fundamental difference between point and group anomaly detection problems. Then, we demonstrate an advantage of OCSMM on uncertain data when the noise is observed explicitly. Lastly, we compare the OCSMM with existing group anomaly detection techniques, namely, K-nearest neighbor (KNN) based anomaly detection (Zhao and Saligrama 2009) with NP-L_2 divergence and NP-Renyi divergence (Póczos et al. 2011), and Multinomial Genre Model (MGM) (Xiong et al. 2011b) on Sloan Digital Sky Survey (SDSS) dataset and High Energy Particle Physics dataset.

Model Selection and Setup. One of the long-standing problems of one-class algorithms is model selection. Since no labeled data is available during training, we cannot perform cross validation. To encourage a fair comparison of different algorithms in our experiments, we will try out different parameter settings and report the best performance of each algorithm. We believe this simple approach should serve its purpose at reflecting the relative performance of different algorithms. We will employ the Gaussian RBF kernel (5.6) throughout the experiments. For the OCSVM and the OCSMM, the bandwidth parameter σ^2 is fixed at $\text{median}\{\|\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}\|^2\}$



(a) OCSVM vs OCSMM



(b) The results of the OCSMM on the mixture of Gaussian dataset

Figure 5.7: (a) The results of group anomaly detection on synthetic data obtained from the OCSVM and the OCSMM. Blue dashed ovals represent the normal groups, whereas red ovals represent the detected anomalous groups. The OCSVM is only able to detect the anomalous groups that are spatially far from the rest in the dataset, whereas the OCSMM also takes into account other higher-order statistics and therefore can also detect anomalous groups which possess distinctive properties. (b) The results of the OCSMM on the synthetic data of the mixture of Gaussian. The shaded boxes represent the anomalous groups that have different mixing proportion to the rest of the dataset. The OCSMM is able to detects the anomalous groups although they look reasonably normal and cannot be easily distinguished from other groups in the data set based only on an inspection.

for all i, j, k, l where $\mathbf{x}_k^{(i)}$ denotes the k-th data point in the i-th group, and we consider $\nu = (0.1, 0.2, \ldots, 0.9)$. The OCSVM treats group means as training samples. For synthetic experiments with OCSMM, we use the empirical kernel (5.5), whereas the non-linear kernel $K(\mathbb{P}_i, \mathbb{P}_j) = \exp(\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\mathbb{P}_j}\|^2_{\mathscr{H}}/2\gamma^2)$ will be used for real data where we set $\gamma = \sigma$. Our experiments suggest that these choices of parameters usually work well in practice. For KNN-L_2 and KNN-Renyi (α=0.99), we consider when there are 3,5,7,9, and 11 nearest neighbors. For MGM, we follow the same experimental setup as in Xiong et al. (2011b).

Synthetic Data

To illustrate the difference between point anomaly and group anomaly, we represent the group of data points by the 2-dimensional Gaussian distribution. We generate 20 normal groups with the covariance $\boldsymbol{\Sigma} = [0.01, 0.008; 0.008, 0.01]$. The means of these groups are drawn uniformly from $[0, 1]$. Then, we generate 2 anomalous groups of Gaussian distributions whose covariances are rotated by 60 degree from the covariance $\boldsymbol{\Sigma}$. Furthermore, we perturb one of the normal groups to make it relatively far from the rest of the dataset to introduce an additional degree of anomaly (cf. Figure 5.7a). Lastly, we generate 100 samples from each of these distributions to form the training set.

For the OCSVM, we represent each group by its empirical average. Since the expected proportion of outliers in the dataset is approximately 10%, we use $\nu = 0.1$ accordingly for both OCSVM and OCSMM. Figure 5.7a depicts the result which demonstrates that the OCSMM can detect anomalous aggregate patterns undetected by the OCSVM.

Then, we conduct similar experiment as that in Xiong et al. (2011b). That is, the groups are represented as a mixture of four 2-dimensional Gaussian distributions. The means of the mixture components are $[-1, -1], [1, -1], [0, 1], [1, 1]$ and the covariances are all $\boldsymbol{\Sigma} = 0.15 \times \mathbf{I}_2$, where $\mathbf{I}_2$ denotes the 2D identity matrix. Then, we design two types of normal groups, which are specified by two mixing proportions $[0.22, 0.64, 0.03, 0.11]$ and $[0.22, 0.03, 0.64, 0.11]$, respectively. To generate a normal group, we first decide with probability $[0.48, 0.52]$ which

mixing proportion will be used. Then, the data points are generated from mixture of Gaussian using the specified mixing proportion. The mixing proportion of the anomalous group is $[0.61, 0.1, 0.06, 0.23]$.

We generated 47 normal groups with $n_i \sim \text{Poisson}(300)$ instances in each group. Note that the individual samples in each group are perfectly normal compared to other samples. To test the performance of our technique, we inject the group anomalies, where the individual points are normal, but they together as a group look anomalous. In this anomalous group the individual points are samples from one of the $K = 4$ normal topics, but the mixing proportion was different from both of the normal mixing proportions. We inject 3 anomalous groups into the data set. The OCSMM is trained using the same setting as in the previous experiment. The results are depicted in Figure 5.7b.

Noisy Data

The OCSMM may be adopted to learn from data points whose uncertainties are observed explicitly. To illustrate this claim, we generate samples from the unit circle using $x = \cos\theta + \varepsilon$ and $y = \sin\theta + \varepsilon$ where $\theta \sim (-\pi, \pi]$ and ε is a zero-mean isotropic Gaussian noise $\mathcal{N}(0, 0.05)$. A different point-wise Gaussian noise $\mathcal{N}(0, \omega_i)$ where $\omega_i \in (0.2, 0.3)$ is further added to each point to simulate the random measurement corruption. In this experiment, we assume that ω_i is available during training. This situation is often encountered in many applications such as astronomy and computational biology. Both OCSVM and OCSMM are trained on the corrupted data. As opposed to the OCSVM that considers only the observed data points, the OCSMM also uses ω_i for every point via the kernel (5.7). Then, we consider a slightly more complicate data generated by $x = r \cdot \cos(\theta)$ and $y = r \cdot \sin(\theta)$ where $r = \sin(4\theta) + 2$ and $\theta \in (0, 2\pi]$. The data used in both examples are illustrated in Figure 5.8.

As illustrated by Figure 5.8, the density function estimated by the OCSMM is relatively less susceptible to the additional corruption than that estimated by the OCSVM, and tends to estimate the true density more accurately. This is not surprising because we also take into account an additional information about the uncertainty. However, this experiment suggests that when dealing with uncertain data, it might be beneficial to also estimate the uncertainty, as commonly performed in astronomy, and incorporate it into the model. This scenario has not been fully investigated in AI and machine learning communities. Our framework provides one possible way to deal with such a scenario.

Sloan Digital Sky Survey

Sloan Digital Sky Survey (SDSS)[2] consists of a series of massive spectroscopic surveys of the distant universe, the milky way galaxies, and extrasolar planetary systems. The SDSS datasets contain images and spectra of more than 930,000 galaxies and more than 120,000 quasars.

In this experiment, we are interested in identifying anomalous groups of galaxies, as previously studied in Póczos et al. (2011) and Xiong et al. (2011b;a). To replicate the experiments conducted in Xiong et al. (2011b), we use the same dataset which consists of 505 spatial clusters of galaxies. Each of which contains about 10-15 galaxies. The data were preprocessed by PCA to reduce the 1000-dimensional features to 4-dimensional vectors.

To evaluate the performance of different algorithms to detect group anomaly, we consider artificially random injections. Each anomalous group is constructed by randomly selecting galaxies. There are 50 anomalous groups of galaxies in total. Note that although these groups of

[2]See http://www.sdss.org for the detail of the surveys.

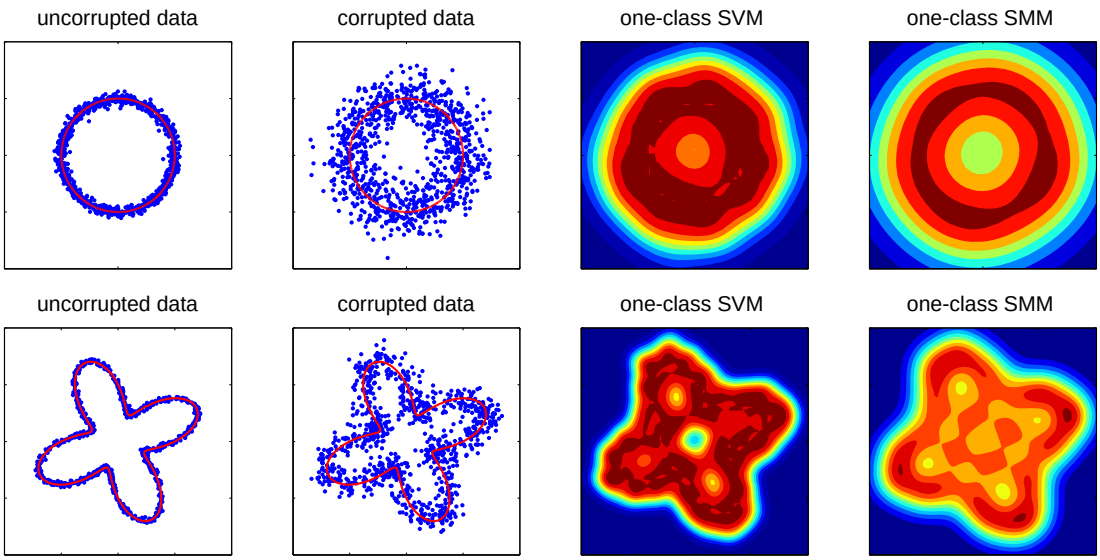


Figure 5.8: The density functions estimated by the OCSVM and the OCSMM using the corrupted data.

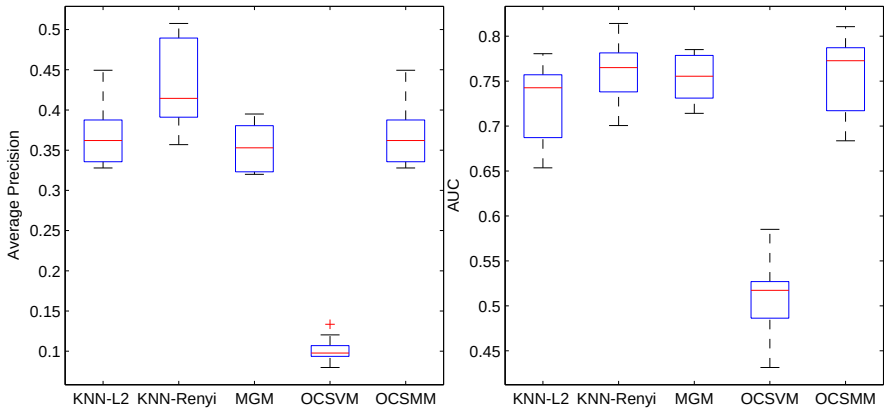


Figure 5.9: The average precision (AP) and area under the ROC curve (AUC) of different group anomaly detection algorithms on the SDSS dataset.

galaxies contain usual galaxies, their aggregations are anomalous due to the way the groups are constructed.

The average precision (AP) and area under the ROC curve (AUC) from 10 random repetitions are shown in Figure 5.9. Based on the average precision, KNN-L2, MGM, and OCSMM achieve similar results on this dataset and KNN-Renyi outperforms all other algorithms. On the other hand, the OCSMM and KNN-Renyi achieve highest AUC scores on this dataset. Moreover, it is clear that point anomaly detection using the OCSVM fails to detect group anomalies.

High Energy Particle Physics

In this section, we demonstrate our group anomaly detection algorithm in high energy particle physics, which is largely the study of fundamental particles, *e.g.*, neutrinos, and their interactions. Essentially, all particles and their dynamics can be described by a quantum field theory called the *Standard Model*. Hence, given massive datasets from high-energy physics experiments, one is interested in discovering deviations from known Standard Model physics.

Searching for the Higgs boson, for example, has recently received much attention in particle

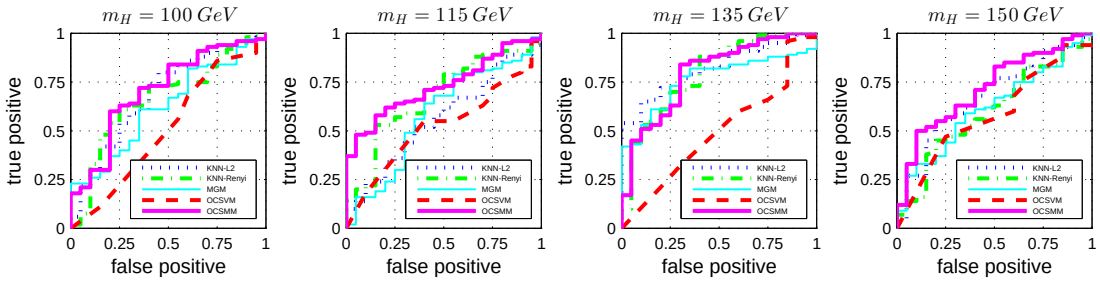


Figure 5.10: The ROC of different group anomaly detection algorithms on the Higgs boson datasets with various Higgs masses m_H.

Table 5.1: The AUC scores for different settings shown in Figure 5.10.

m_H	KNN-L2	KNN-Renyi	MGM	OCSVM	OCSMM
100 GeV	0.6835	0.6655	0.6350	0.5125	**0.7085**
115 GeV	0.5645	0.6783	0.5860	0.5263	**0.7305**
135 GeV	**0.8190**	0.7925	0.7630	0.4958	0.7950
150 GeV	0.6713	0.6027	0.6165	0.5862	**0.7200**

physics and machine learning communities (see *e.g.*, Bhat (2011), Vatanen et al. (2012) and references therein). A new physical phenomena usually manifest themselves as tiny excesses of certain types of collision events among a vast background of known physics in particle detectors.

Anomalies occur as a cluster among the background data. The background data distribution contaminated by these anomalies will therefore be different from the true background distribution. It is very difficult to detect this difference in general because the contamination can be considerably small. In this experiment, we consider similar condition as in Vatanen et al. (2012) and generate data using the standard HEP Monte Carlo generators such as PYTHIA[3]. In particular, we consider a Monte Carlo simulated events where the Higgs is produced in association with the W boson and decays into two bottom quarks.

The data vector consists of 5 variables (p_x, p_y, p_z, e, m) corresponding to different characteristics of the topology of a collision event. The variables p_x, p_y, p_z, e represents the momentum four-vector in units of GeV with $c = 1$. The variable m is the particle mass in the same unit. The signal looks slightly different for different Higgs masses m_H, which is an unknown free parameter in the Standard Model. In this experiment, we consider $m_H = 100, 115, 135$, and 150 GeV. We generate 120 groups of collision events, 100 of which contain only background signals, whereas the rest also contain the Higgs boson collision events. For each group, the number of observable particles ranges from 200 to 500 particles. The goal is to detect the anomalous groups of signals which might contain the Higgs boson without prior knowledge of m_H.

Figure 5.10 depicts the ROC of different group anomaly detection algorithms. The associated AUC scores for different settings are reported in Table 5.1. The OCSMM and KNN-based group anomaly detection algorithms tend to achieve competitive performance and outperform the MGM algorithm. Moreover, it is clear that traditional point anomaly detection algorithm fails to detect high-level anomalous structures.

[3] http://home.thep.lu.se/∼torbjorn/Pythia.html

5.3.6 Discussions

To conclude, we propose a simple and efficient algorithm for detecting group anomalies called one-class support measure machine (OCSMM). To handle aggregate behaviors of data points, groups are represented as probability distributions which account for higher-order information arising from those behaviors. The set of distributions are represented as mean functions in the RKHS via the kernel mean embedding. We also extend the relationship between the OCSVM and the KDE to the OCSMM in the context of variable kernel density estimation, bridging the gap between large-margin approach and kernel density estimation. We demonstrate the proposed algorithm on both synthetic and real-world datasets, which achieve competitive results compared to existing group anomaly detection techniques.

It is vital to note the differences between the OCSMM and hierarchical probabilistic models such as MGM and FGM. Firstly, the probabilistic models assume that data are generated according to some parametric distributions, *i.e.*, mixture of Gaussian, whereas the OCSMM is nonparametric in the sense that no assumption is made about the distributions. It is therefore applicable to a wider range of applications. Secondly, the probabilistic models follow a bottom-up approach. That is, detecting group-based anomalies requires point-based anomaly detection. Thus, the performance also depends on how well anomalous points can be detected. Furthermore, it is computational expensive and may not be suitable for large-scale datasets. On the other hand, the OCSMM adopts the top-down approach by detecting the group-based anomalies directly. If one is interested in finding anomalous points, this can be done subsequently in a group-wise manner. As a result, the top-down approach is generally less computational expensive and can be used efficiently for online applications and large-scale datasets.

5.4 Domain Generalization

Domain generalization considers how to take knowledge acquired from an arbitrary number of related domains, and apply it to previously unseen domains. To illustrate the problem, consider an example taken from Blanchard et al. (2011b) which studied automatic gating of flow cytometry data. For each of N patients, a set of n_i cells are obtained from peripheral blood samples using a flow cytometer. The cells are then labeled by an expert into different subpopulations, *e.g.*, as a lymphocyte or not. Correctly identifying cell subpopulations is vital for diagnosing the health of patients. However, manual gating is very time consuming. To automate gating, we need to construct a classifier that generalizes well to previously unseen patients, where the distribution of cell types may differ dramatically from the training data.

Unfortunately, we cannot apply standard machine learning techniques directly because the data violates the basic assumption that training data and test data come from the same distribution. Moreover, the training set consists of heterogeneous samples from several distributions, *i.e.*, gated cells from several patients. In this case, the data exhibits covariate (or dataset) shift (Widmer and Kurat 1996, Quionero-Candela et al. 2009, Bickel et al. 2009): although the marginal distributions $\mathbb{P}_X$ on cell attributes vary due to biological or technical variations, the functional relationship $\mathbb{P}(Y|X)$ across different domains is largely stable (cell type is a stable function of a cell's chemical attributes).

A considerable effort has been made in domain adaptation and transfer learning to remedy this problem, see Pan and Yang (2010), Ben-David et al. (2010) and references therein. Given a test domain, *e.g.*, a cell population from a new patient, the idea of domain adaptation is to adapt a classifier trained on the training domain, *e.g.*, a cell population from another patient, such that the generalization error on the test domain is minimized. The main drawback of this approach is that one has to repeat this process for every new patient, which can be time-consuming –

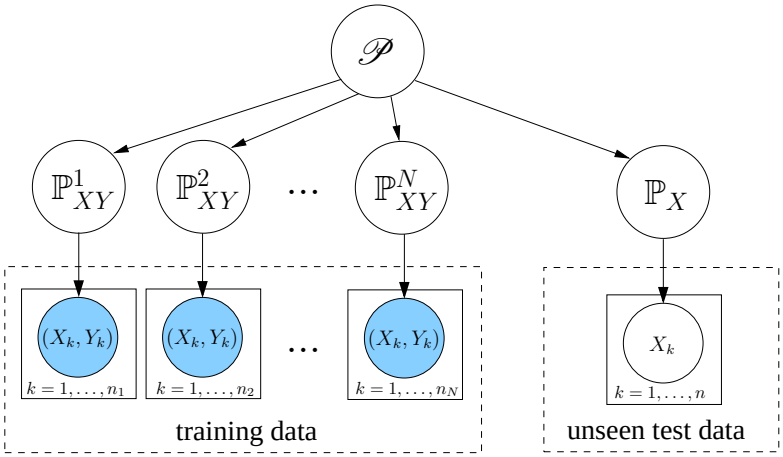


Figure 5.11: A simplified schematic diagram of the domain generalization framework. A major difference between our framework and most previous work in domain adaptation is that we do not observe the test domains during training time. See text for detailed description on how the data are generated.

especially in medical diagnosis where time is a valuable asset. In this work, across-domain information, which may be more informative than the domain-specific information, is extracted from the training data and used to generalize the classifier to new patients without retraining.

Overview. The goal of (supervised) domain generalization is to estimate a functional relationship that handles changes in the marginal $\mathbb{P}(X)$ or conditional $\mathbb{P}(Y|X)$ well, see Figure 5.11. We assume that the conditional probability $\mathbb{P}(Y|X)$ is stable or varies smoothly with the marginal $\mathbb{P}(X)$. Even if the conditional is stable, learning algorithms may still suffer from *model misspecification* due to variation in the marginal $\mathbb{P}(X)$. That is, if the learning algorithm cannot find a solution that perfectly captures the functional relationship between X and Y then its approximate solution will be sensitive to changes in $\mathbb{P}(X)$.

In this paper, we introduce Domain Invariant Component Analysis (DICA), a kernel-based algorithm that finds a transformation of the data that (i) minimizes the difference between marginal distributions $\mathbb{P}_X$ of domains as much as possible while (ii) preserving the functional relationship $\mathbb{P}(Y|X)$.

The novelty of this work is twofold. First, DICA extracts *invariants*: features that transfer across domains. It not only minimizes the divergence between marginal distributions $\mathbb{P}(X)$, but also preserves the functional relationship encoded in the posterior $\mathbb{P}(Y|X)$. The resulting learning algorithm is very simple. Second, while prior work in domain adaptation focused on using data from many different domains to specifically improve the performance on the target task, which is observed during the training time (the classifier is adapted to the specific target task), we assume access to abundant training data and are interested in the generalization ability of the invariant subspace to previously unseen domains (the classifier generalizes to new domains without retraining).

Moreover, we show that DICA generalizes or is closely related to many well-known dimension reduction algorithms including kernel principal component analysis (KPCA) (Schölkopf et al. 1998, Fukumizu et al. 2004), transfer component analysis (TCA) (Pan et al. 2011), and covariance operator inverse regression (COIR) (Kim and Pavlovic 2011).

Related work. Domain generalization is a form of transfer learning, which applies expertise acquired in source domains to improve learning of target domains (cf. Pan and Yang (2010)

and references therein). Most previous work assumes the availability of the target domain to which the knowledge will be transferred. In contrast, domain generalization focuses on the generalization ability on previously unseen domains. That is, the test data comes from domains that are not available during training.

Recently, Blanchard et al. (2011b) proposed an augmented SVM that incorporates empirical marginal distributions into the kernel. A detailed error analysis showed universal consistency of the approach. We apply methods from Blanchard et al. (2011b) to derive theoretical guarantees on the finite sample performance of DICA.

Learning a shared subspace is a common approach in settings where there is distribution mismatch. For example, a typical approach in multitask learning is to uncover a joint (latent) feature/subspace that benefits tasks individually (Argyriou et al. 2007, Gu and Zhou 2009, Passos et al. 2012). A similar idea has been adopted in domain adaptation, where the learned subspace reduces mismatch between source and target domains (Gretton et al. 2009b, Pan et al. 2011). Although these approaches have proven successful in various applications, no previous work has fully investigated the generalization ability of a subspace to unseen domains.

5.4.1 Distributional (Co-)Variance

First, we define the distributional variance, which measures the dissimilarity across domains. We decompose $\mathscr{P}$ into $\mathscr{P}_X$, which generates the marginal distribution $\mathbb{P}_X$, and $\mathscr{P}_{Y|X}$, which generates posteriors $\mathbb{P}_{Y|X}$. The data generating process begins by generating the marginal $\mathbb{P}_X$ according to $\mathscr{P}_X$. Conditioned on $\mathbb{P}_X$, it then generate conditional $\mathbb{P}_{Y|X}$ according to $\mathscr{P}_{Y|X}$. The data point $(\mathbf{x}, \mathbf{y})$ is generated according to $\mathbb{P}_X$ and $\mathbb{P}_{Y|X}$, respectively. Given set of distributions $\mathcal{P} = \{\mathbb{P}^1, \mathbb{P}^2 \ldots, \mathbb{P}^N\}$ drawn according to $\mathscr{P}_X$, define $N \times N$ Gram matrix $\mathbf{G}$ with entries

$$\mathbf{G}_{ij} := \langle \boldsymbol{\mu}_{\mathbb{P}^i}, \boldsymbol{\mu}_{\mathbb{P}^j} \rangle_{\mathscr{H}} = \iint k(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbb{P}^i(\mathbf{x}) \, \mathrm{d}\mathbb{P}^j(\mathbf{z}), \tag{5.10}$$

for $i, j = 1, \ldots, N$. Note that $\mathbf{G}_{ij}$ is the inner product between kernel mean embeddings of $\mathbb{P}^i$ and $\mathbb{P}^j$ in $\mathscr{H}$. Based on (5.10), we define the distributional variance, which estimates the variance of the distribution $\mathscr{P}_X$:

Definition 5.1. *Introduce probability distribution $\mathcal{P}$ on $\mathscr{H}$ with $\mathcal{P}(\boldsymbol{\mu}_{\mathbb{P}^i}) = \frac{1}{N}$ and center $\mathbf{G}$ to obtain the covariance operator of $\mathcal{P}$, denoted as $\boldsymbol{\Sigma} := \mathbf{G} - \mathbf{1}_N\mathbf{G} - \mathbf{G}\mathbf{1}_N + \mathbf{1}_N\mathbf{G}\mathbf{1}_N$. The* ***distributional variance*** *is*

$$\mathbb{V}_{\mathscr{H}}(\mathcal{P}) := \frac{1}{N}\mathrm{tr}(\boldsymbol{\Sigma}) = \frac{1}{N}\mathrm{tr}(\mathbf{G}) - \frac{1}{N^2}\sum_{i,j=1}^{N} \mathbf{G}_{ij}. \tag{5.11}$$

The following theorem shows that the distributional variance is suitable as a measure of divergence between domains.

Theorem 5.4. *Let $\bar{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}^i$. If k is a characteristic kernel, then $\mathbb{V}_{\mathscr{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{\mu}_{\mathbb{P}^i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}} = 0$ if and only if $\mathbb{P}^1 = \mathbb{P}^2 = \cdots = \mathbb{P}^N$.*

To estimate $\mathbb{V}_{\mathscr{H}}(\mathcal{P})$ from N sample sets $\mathcal{S} = \{S^i\}_{i=1}^N$ drawn from $\mathbb{P}^1, \ldots, \mathbb{P}^N$, we define block kernel and coefficient matrices

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{1,1} & \cdots & \mathbf{K}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{N,1} & \cdots & \mathbf{K}_{N,N} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{1,1} & \cdots & \mathbf{Q}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{Q}_{N,1} & \cdots & \mathbf{Q}_{N,N} \end{pmatrix} \in \mathbb{R}^{n \times n} \ ,$$

where $n = \sum_{i=1}^{N} n_i$ and $[\mathbf{K}_{i,j}]_{k,l} = k(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)})$ is the Gram matrix evaluated between the sample S^i and S^j. Following (5.11), elements of the coefficient matrix $\mathbf{Q}_{i,j} \in \mathbb{R}^{n_i \times n_j}$ equal $(N-1)/(N^2 n_i^2)$ if $i = j$, and $-1/(N^2 n_i n_j)$ otherwise. Hence, the empirical distributional variance is

$$\widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{KQ}) \ . \tag{5.12}$$

Theorem 5.5. *The empirical estimator* $\widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{KQ})$ *obtained from Gram matrix*

$$\widehat{\mathbf{G}}_{ij} := \frac{1}{n_i \cdot n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} k(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)})$$

is a consistent estimator of $\mathbb{V}_{\mathscr{H}}(\mathcal{P})$.

5.4.2 Domain-Invariant Component Analysis

Let $\mathcal{X}$ denote a nonempty input space and $\mathcal{Y}$ an arbitrary output space. We define a **domain** to be a joint distribution $\mathbb{P}_{XY}$ on $\mathcal{X} \times \mathcal{Y}$, and let $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ denote the set of all domains. Let $\mathfrak{P}_{\mathcal{X}}$ and $\mathfrak{P}_{\mathcal{Y}|\mathcal{X}}$ denote the set of probability distributions $\mathbb{P}_X$ on X and $\mathbb{P}_{Y|X}$ on Y given X respectively.

We assume domains are sampled from probability distribution $\mathscr{P}$ on $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ which has a bounded second moment, *i.e.*, the variance is well-defined. Domains are not observed directly. Instead, we observe N samples $\mathcal{S} = \{S^i\}_{i=1}^N$, where $S^i = \{(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})\}_{k=1}^{n_i}$ is sampled from $\mathbb{P}_{XY}^i$ and each $\mathbb{P}_{XY}^1, \ldots, \mathbb{P}_{XY}^N$ is sampled from $\mathscr{P}$. Since in general $\mathbb{P}_{XY}^i \neq \mathbb{P}_{XY}^j$, the samples in $\mathcal{S}$ are not i.i.d. Let $\widehat{\mathbb{P}}^i$ denote empirical distribution associated with each sample S^i. For brevity, we use $\mathbb{P}$ and $\mathbb{P}_X$ interchangeably to denote the marginal distribution.

Let $\mathscr{H}$ and $\mathscr{F}$ denote RKHS on $\mathcal{X}$ and $\mathcal{Y}$ with kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, respectively. Associated with $\mathscr{H}$ and $\mathscr{F}$ are mappings $\mathbf{x} \to \phi(\mathbf{x}) \in \mathscr{H}$ and $\mathbf{y} \to \varphi(\mathbf{y}) \in \mathscr{F}$ induced by the kernels $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$. Without loss of generality, we assume the feature maps of X and Y have zero means, *i.e.*, $\sum_{k=1}^n \phi(\mathbf{x}_k) = 0 = \sum_{k=1}^n \varphi(\mathbf{y}_k)$. Let $\mathbf{C}_{XX}$, $\mathbf{C}_{YY}$, $\mathbf{C}_{XY}$, and $\mathbf{C}_{YX}$ be the covariance operators in and between the RKHS of X and Y.

Objective. Using the samples $\mathcal{S}$, our goal is to produce an estimate $f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ that generalizes well to test samples $S^t = \{\mathbf{x}_k^{(t)}\}_{k=1}^{n_t}$ drawn according to some unknown distribution $\mathbb{P}^t \in \mathfrak{P}_{\mathcal{X}}$ (Blanchard et al. 2011b). Since the performance of f depends in part on how dissimilar the test distribution $\mathbb{P}^t$ is from those in the training samples, we propose to preprocess the data to actively reduce the dissimilarity between domains. Intuitively, we want to find transformation $\mathcal{B}$ in $\mathscr{H}$ that (i) minimizes the distance between empirical distributions of the transformed samples $\mathcal{B}(S^i)$ and (ii) preserves the functional relationship between X and Y, *i.e.*, $Y \perp\!\!\!\perp X \,|\, \mathcal{B}(X)$. We formulate an optimization problem capturing these constraints below.

DICA finds an orthogonal transform $\mathcal{B}$ onto a low-dimensional subspace ($m \ll n$) that minimizes the distributional variance $\mathbb{V}_{\mathscr{H}}(\mathcal{S})$ between samples from $\mathcal{S}$, *i.e.*the *dissimilarity across domains*. Simultaneously, we require that $\mathcal{B}$ preserves the functional relationship between X and Y, *i.e.*$Y \perp\!\!\!\perp X \,|\, \mathcal{B}(X)$.

Minimizing distributional variance. To simplify notation, we "flatten" $\{(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})_{k=1}^{n_i}\}_{i=1}^N$ to $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^n$ where $n = \sum_{i=1}^N n_i$. Let $\boldsymbol{\beta}_k = \sum_{i=1}^n \beta_k^i \phi(\mathbf{x}_i) = \Phi_{\mathbf{x}}\boldsymbol{\beta}_k$ be the k^{th} basis function of $\mathcal{B}$ where $\Phi_{\mathbf{x}} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]$ and $\boldsymbol{\beta}_k$ are n-dimensional coefficient vectors. Let $B = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m]$ and $\tilde{\Phi}_{\mathbf{x}}$ denote the projection of $\Phi_{\mathbf{x}}$ onto $\boldsymbol{\beta}_k$, *i.e.*, $\tilde{\Phi}_{\mathbf{x}} =$

$\boldsymbol{\beta}_k^\top \Phi_{\mathbf{x}} = \boldsymbol{\beta}_k^\top \Phi_{\mathbf{x}}^\top \Phi_{\mathbf{x}} = \boldsymbol{\beta}_k^\top \mathbf{K}$. The kernel on the $\mathcal{B}$-projection of X is

$$\widetilde{\mathbf{K}} := \tilde{\Phi}_{\mathbf{x}}^\top \tilde{\Phi}_{\mathbf{x}} = \mathbf{K}\mathbf{B}\mathbf{B}^\top\mathbf{K} \ . \tag{5.13}$$

After applying transformation $\mathcal{B}$, the empirical distributional variance between sample distributions is

$$\widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{B}\mathcal{S}) = \mathrm{tr}(\widetilde{\mathbf{K}}\mathbf{Q}) = \mathrm{tr}(\mathbf{B}^\top\mathbf{K}\mathbf{Q}\mathbf{K}\mathbf{B}) \ . \tag{5.14}$$

Preserving the functional relationship. The central subspace C is the minimal subspace that captures the functional relationship between X and Y, *i.e.*, $Y \perp\!\!\!\perp X|\mathbf{C}^\top X$. Note that in this work we generalize a linear transformation $\mathbf{C}^\top X$ to nonlinear one $\mathcal{B}(X)$. To find the central subspace we use the inverse regression framework, (Li 1991):

Theorem 5.6. *If there exists a* ***central subspace*** $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_m]$ *satisfying* $Y \perp\!\!\!\perp X|\mathbf{C}^\top X$, *and for any* $\mathbf{a} \in \mathbb{R}^d$, $\mathbb{E}[\mathbf{a}^\top X|\mathbf{C}^\top X]$ *is linear in* $\{\mathbf{c}_i^\top X\}_{i=1}^m$, *then* $\mathbb{E}[X|Y] \subset$ *span*$\{\mathbf{C}_{XX}\mathbf{c}_i\}_{i=1}^m$.

It follows that the bases $\mathbf{C}$ of the central subspace coincide with the m largest eigenvectors of $\mathbb{V}(\mathbb{E}[X|Y])$ premultiplied by $\mathbf{C}_{XX}^{-1}$. Thus, the basis $\mathbf{c}$ is the solution to the eigenvalue problem $\mathbb{V}(\mathbb{E}[X|Y])\mathbf{C}_{XX}\mathbf{c} = \gamma\mathbf{C}_{XX}\mathbf{c}$. Alternatively, for each $\mathbf{c}_k$ one may solve

$$\max_{\mathbf{c}_k \in \mathbb{R}^d} \frac{\mathbf{c}_k^\top \mathbf{C}_{XX}^{-1}\mathbb{V}(\mathbb{E}[X|Y])\mathbf{C}_{XX}\mathbf{c}_k}{\mathbf{c}_k^\top \mathbf{c}_k}$$

under the condition that $\mathbf{c}_k$ is chosen to not be in the span of the previously chosen $\mathbf{c}_k$. In our case, $\mathbf{x}$ is mapped to $\phi(\mathbf{x}) \in \mathscr{H}$ induced by the kernel k and $\mathcal{B}$ has nonlinear basis functions $\mathbf{c}_k \in \mathscr{H}$, $k = 1, \ldots, m$. This nonlinear extension implies that $\mathbb{E}[X|Y]$ lies on a function space spanned by $\{\mathbf{C}_{XX}\mathbf{c}_k\}_{k=1}^m$, which coincide with the eigenfunctions of the operator $\mathbb{V}(\mathbb{E}[X|Y])$ (Wu 2008, Kim and Pavlovic 2011). Since we always work in $\mathscr{H}$, we drop ϕ from the notation below.

To avoid slicing the output space explicitly (Li 1991, Wu 2008), we exploit its kernel structure when estimating the covariance of the inverse regressor. The following result from Kim and Pavlovic (2011) states that, under a mild assumption, $\mathbb{V}(\mathbb{E}[X|Y])$ can be expressed in terms of covariance operators:

Theorem 5.7. *If for all* $f \in \mathscr{H}$, *there exists* $g \in \mathscr{F}$ *such that* $\mathbb{E}[f(X)|\mathbf{y}] = g(\mathbf{y})$ *for almost every* y, *then*

$$\mathbb{V}(\mathbb{E}[X|Y]) = \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX} \ . \tag{5.15}$$

Let $\Phi_{\mathbf{y}} = [\varphi(\mathbf{y}_1), \ldots, \varphi(\mathbf{y}_n)]$ and $\mathbf{L} = \Phi_{\mathbf{y}}^\top\Phi_{\mathbf{y}}$. The covariance of inverse regressor (5.15) is estimated from the samples $\mathcal{S}$ as $\widehat{\mathbb{V}}(\mathbb{E}[X|Y]) = \widehat{\mathbf{C}}_{XY}\widehat{\mathbf{C}}_{YY}^{-1}\widehat{\mathbf{C}}_{YX} = \frac{1}{n}\Phi_{\mathbf{x}}\mathbf{L}(\mathbf{L} + n\varepsilon\mathbf{I}_n)^{-1}\Phi_{\mathbf{x}}^\top$ where $\widehat{\mathbf{C}}_{XY} = \frac{1}{n}\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^\top$ and $\widehat{\mathbf{C}}_{YY} = \frac{1}{n}\Phi_{\mathbf{y}}\Phi_{\mathbf{y}}^\top$. Assuming inverses $\widehat{\mathbf{C}}_{YY}^{-1}$ and $\widehat{\mathbf{C}}_{XX}^{-1}$ exist, a straightforward computation (see Appendix C.8) shows

$$\begin{aligned} \boldsymbol{\beta}_k^\top \widehat{\Sigma}_{\mathrm{xx}}^{-1}\widehat{\mathbb{V}}(\mathbb{E}[X|Y])\widehat{\Sigma}_{\mathrm{xx}}\boldsymbol{\beta}_k &= \frac{1}{n}\boldsymbol{\beta}_k^\top \mathbf{L}(\mathbf{L} + n\varepsilon\mathbf{I})^{-1}\mathbf{K}^2\boldsymbol{\beta}_k \\ \boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k &= \boldsymbol{\beta}_k^\top \mathbf{K}\boldsymbol{\beta}_k, \end{aligned} \tag{5.16}$$

where ε smoothes the affinity structure of the output space Y, thus acting as a kernel regularizer. Since we are interested in the projection of $\phi(\mathbf{x})$ onto the basis functions $\boldsymbol{\beta}_k$, we formulate the optimization in terms of $\boldsymbol{\beta}_k$. For a new test sample $\mathbf{x}_t$, the projection onto basis function $\boldsymbol{\beta}_k$ is $\mathbf{k}_t\boldsymbol{\beta}_k$, where $\mathbf{k}_t = [k(\mathbf{x}_1, \mathbf{x}_t), \ldots, k(\mathbf{x}_n, \mathbf{x}_t)]$.

The optimization problem. Combining (5.14) and (5.16), DICA finds $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m]$ that solves

$$\max_{\mathbf{B} \in \mathbb{R}^{n \times m}} \frac{\frac{1}{n}\mathrm{tr}\left(\mathbf{B}^\top \mathbf{L}(\mathbf{L} + n\varepsilon \mathbf{I}_n)^{-1}\mathbf{K}^2\mathbf{B}\right)}{\mathrm{tr}\left(\mathbf{B}^\top \mathbf{K}\mathbf{Q}\mathbf{K}\mathbf{B} + \mathbf{B}\mathbf{K}\mathbf{B}\right)} \tag{5.17}$$

The numerator requires that $\mathbf{B}$ aligns with the bases of the central subspace. The denominator forces both dissimilarity across domains and the complexity of $\mathbf{B}$ to be small, thereby tightening generalization bounds, see §3.4.3. Rewriting (5.17) as a constrained optimization (see Appendix C.9) yields Lagrangian

$$\begin{aligned}\mathcal{L} =& \frac{1}{n}\mathrm{tr}\left(\mathbf{B}^\top \mathbf{L}(\mathbf{L} + n\varepsilon \mathbf{I}_n)^{-1}\mathbf{K}^2\mathbf{B}\right) \\ & - \mathrm{tr}\left(\left(\mathbf{B}^\top \mathbf{K}\mathbf{Q}\mathbf{K}\mathbf{B} + \mathbf{B}\mathbf{K}\mathbf{B} - \mathbf{I}_m\right)\Gamma\right) \quad ,\end{aligned} \tag{5.18}$$

where Γ is a diagonal matrix containing the Lagrange multipliers. Setting the derivative of (5.18) w.r.t. $\mathbf{B}$ to zero yields the generalized eigenvalue problem:

$$\frac{1}{n}\mathbf{L}(\mathbf{L} + n\varepsilon \mathbf{I}_n)^{-1}\mathbf{K}^2\mathbf{B} = (\mathbf{K}\mathbf{Q}\mathbf{K} + \mathbf{K})\mathbf{B}\Gamma \quad . \tag{5.19}$$

Transformation $\mathbf{B}$ corresponds to the m leading eigenvectors of the generalized eigenvalue problem (5.19).[4]

The inverse regression framework based on covariance operators has two benefits. First, it avoids explicitly slicing the output space, which makes it suitable for high-dimensional output. Second, it allows for structured outputs on which explicit slicing may be impossible, *e.g.*, trees and sequences. Since our framework is based entirely on kernels, it is applicable to any type of input and output variables, as long as the corresponding kernels can be defined.

Unsupervised DICA

In some application domains, such as image denoising, information about the target may not be available. We therefore derive an unsupervised version of DICA. Instead of preserving the central subspace, unsupervised DICA (UDICA) maximizes the variance of X in the feature space, which is estimated as $\frac{1}{n}\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^2\mathbf{B})$. Thus, UDICA solves

$$\max_{\mathbf{B} \in \mathbb{R}^{n \times m}} \frac{\frac{1}{n}\mathrm{tr}(\mathbf{B}^\top \mathbf{K}^2\mathbf{B})}{\mathrm{tr}(\mathbf{B}^\top \mathbf{K}\mathbf{Q}\mathbf{K}\mathbf{B} + \mathbf{B}^\top \mathbf{K}\mathbf{B})} \quad . \tag{5.20}$$

Similar to DICA, the solution of (5.20) is obtained by solving the generalized eigenvalue problem

$$\frac{1}{n}\mathbf{K}^2\mathbf{B} = (\mathbf{K}\mathbf{Q}\mathbf{K} + \mathbf{K})\mathbf{B}\Gamma \quad . \tag{5.21}$$

UDICA is a special case of DICA where $\mathbf{L} = \frac{1}{n}\mathbf{I}$ and $\varepsilon \to 0$. Algorithm 1 summarizes supervised and unsupervised domain-invariant component analysis.

5.4.3 Relations to Other Methods

The DICA and UDICA algorithms generalize many well-known dimension reduction techniques. In the supervised setting, if dataset $\mathcal{S}$ contains samples drawn from a single distribution $\mathbb{P}_{XY}$ then we have $\mathbf{K}\mathbf{Q}\mathbf{K} = \mathbf{0}$. Substituting $\alpha := \mathbf{K}\mathbf{B}$ gives the eigenvalue problem

[4] In practice, it is more numerically stable to solve the generalized eigenvalue problem $\frac{1}{n}\mathbf{L}(\mathbf{L} + n\varepsilon \mathbf{I}_n)^{-1}\mathbf{K}^2\mathbf{B} = (\mathbf{K}\mathbf{Q}\mathbf{K} + \mathbf{K} + \lambda \mathbf{I})\mathbf{B}\Gamma$, where λ is a small constant.

Algorithm 1 Domain-Invariant Component Analysis

Input: Parameters λ, ε, and $m \ll n$.
Sample $\mathcal{S} = \{S^i = \{(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})\}_{k=1}^{n_i}\}_{i=1}^{N}$.
Output: Projection $\mathbf{B}_{n\times m}$ and kernel $\widetilde{\mathbf{K}}_{n\times n}$.

1: Calculate gram matrix $[\mathbf{K}_{ij}]_{kl} = k(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)})$ and $[\mathbf{L}_{ij}]_{kl} = l(\mathbf{y}_k^{(i)}, \mathbf{y}_l^{(j)})$.
2: **Supervised:** $C = \mathbf{L}(\mathbf{L} + n\varepsilon\mathbf{I})^{-1}\mathbf{K}^2$.
3: **Unsupervised:** $C = \mathbf{K}^2$.
4: Solve $\frac{1}{n}C\mathbf{B} = (\mathbf{KQK} + \mathbf{K} + \lambda\mathbf{I})\mathbf{B}\Gamma$ for $\mathbf{B}$.
5: Output $\mathbf{B}$ and $\widetilde{\mathbf{K}} \leftarrow \mathbf{KBB}^\top\mathbf{K}$.
6: The test kernel $\widetilde{\mathbf{K}}^t \leftarrow \mathbf{K}^t\mathbf{BB}^\top\mathbf{K}$ where $\mathbf{K}^t_{n_t\times n}$ is the joint kernel between test and training data.

$\frac{1}{n}\mathbf{L}(\mathbf{L} + n\varepsilon\mathbf{I})^{-1}\mathbf{K}\alpha = \mathbf{K}\alpha\Gamma$, which corresponds to covariance operator inverse regression (COIR) (Kim and Pavlovic 2011).

If there is only a single distribution then unsupervised DICA reduces to KPCA since $\mathbf{KQK} = \mathbf{0}$ and finding $\mathbf{B}$ requires solving the eigensystem $\mathbf{KB} = \mathbf{B}\Gamma$ which recovers KPCA (Schölkopf et al. 1998). If there are two domains, source $\mathbb{P}_S$ and target $\mathbb{P}_T$, then UDICA is closely related – though not identical to – Transfer Component Analysis (Pan et al. 2011). This follows from the observation that $\mathbb{V}_{\mathscr{H}}(\{\mathbb{P}_S, \mathbb{P}_T\}) = \|\boldsymbol{\mu}_{\mathbb{P}_S} - \boldsymbol{\mu}_{\mathbb{P}_T}\|^2$, see proof of Theorem 5.4.

5.4.4 A Learning-Theoretic Bound

We bound the generalization error of a classifier trained after DICA-preprocessing. The main complication is that samples are not identically distributed. We adapt an approach to this problem developed in Blanchard et al. (2011b) to prove a generalization bound that applies after transforming the empirical sample using $\mathcal{B}$. Recall that $\mathcal{B} = \Phi_{\mathbf{x}}\mathbf{B}$.

Define kernel $\bar{k}$ on $\mathscr{P} \times \mathcal{X}$ as $\bar{k}((\mathbb{P}, \mathbf{x}), (\mathbb{P}', \mathbf{x}')) := k_{\mathscr{P}}(\mathbb{P}, \mathbb{P}') \cdot k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')$. Here, $k_{\mathcal{X}}$ is the kernel on $\mathscr{H}_{\mathcal{X}}$ and the kernel on distributions is $k_{\mathscr{P}}(\mathbb{P}, \mathbb{P}') := \kappa(\boldsymbol{\mu}_{\mathbb{P}}, \boldsymbol{\mu}_{\mathbb{P}'})$ where κ is a positive definite kernel (Christmann and Steinwart 2010, Muandet et al. 2012). Let $\Psi_{\mathscr{P}}$ denote the corresponding feature map.

Theorem 5.8. *Under reasonable technical assumptions, see Supplementary, it holds with probability at least* $1 - \delta$ *that,*

$$\sup_{\|f\|_{\mathscr{H}}\leq 1} \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2$$
$$\leq c_1 \frac{1}{N} \mathrm{tr}(\mathbf{B}^\intercal\mathbf{KQKB}) + \mathrm{tr}(\mathbf{B}^\top\mathbf{KB}) \left(c_2 \frac{N(\log\frac{1}{\delta} + 2\log N)}{n} + \frac{c_3 \log\frac{1}{\delta} + c_4}{N} \right).$$

The LHS is the difference between the training error and expected error (w.r.t. the distribution on domains $\mathscr{P}^*$) after applying $\mathcal{B}$.

The first term in the bound, involving $\mathrm{tr}(\mathbf{B}^\intercal\mathbf{KQKB})$, quantifies the distributional variance after applying the transform: the higher the distributional variance, the worse the guarantee, tying in with analogous results in Ben-David et al. (2007; 2010). The second term in the bound depends on the size of the distortion $\mathrm{tr}(\mathbf{B}^\intercal\mathbf{KB})$ introduced by $\mathbf{B}$: the more complicated the transform, the worse the guarantee.

The bound reveals a tradeoff between reducing the distributional variance and the complexity or size of the transform used to do so. The denominator of (5.17) is a sum of these terms, so that DICA tightens the bound in Theorem 5.8.

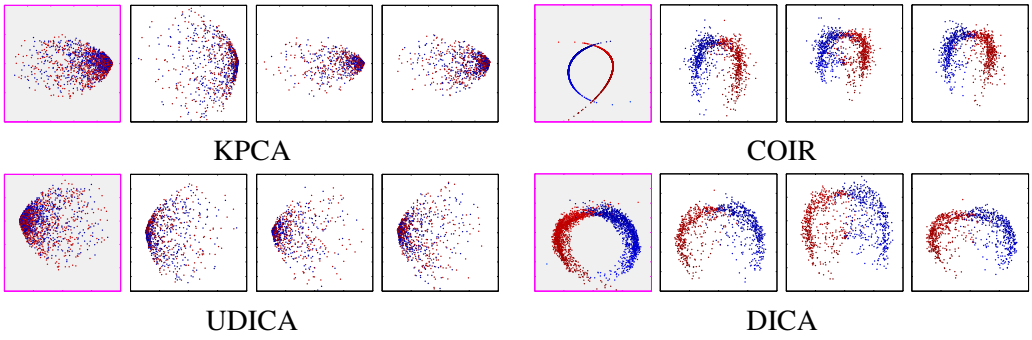


Figure 5.12: Projections of a synthetic dataset onto the first two eigenvectors obtained from the KPCA, UDICA, COIR, and DICA. The colors of data points corresponds to the output values. The shaded boxes depict the projection of training data, whereas the unshaded boxes show projections of unseen test datasets. The feature representations learnt by UDICA and DICA are more stable across test domains than those learnt by KPCA and COIR.

Preserving the functional relationship (*i.e.*central subspace) by maximizing the numerator in (5.17) should reduce the empirical risk $\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i)$. However, a rigorous demonstration has yet to be found.

5.4.5 Experimental Results

We illustrate the difference between the proposed algorithms and their single-domain counterparts using a synthetic dataset. Furthermore, we evaluate DICA in two tasks: a classification task on flow cytometry data and a regression task for Parkinson's telemonitoring.

Toy Experiments

We generate 10 collections of $n_i \sim \texttt{Poisson}(200)$ data points. The data in each collection is generated according to a five-dimensional zero-mean Gaussian distribution. For each collection, the covariance of the distribution is generated from Wishart distribution $\mathcal{W}(0.2 \times \mathbf{I}_5, 10)$. This step is to simulate different marginal distributions. The output value is $y = \text{sign}(b_1^\top \mathbf{x} + \epsilon_1) \cdot \log(|b_2^\top \mathbf{x} + c + \epsilon_2|)$, where b_1, b_2 are the weight vectors, c is a constant, and $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1)$. Note that b_1 and b_2 form a low-dimensional subspace that captures the functional relationship between X and Y. We then apply the KPCA, UDICA, COIR, and DICA algorithms on the dataset with Gaussian RBF kernels for both X and Y with bandwidth parameters $\sigma_x = \sigma_y = 1$, $\lambda = 0.1$, and $\varepsilon = 10^{-4}$.

Fig. 5.12 shows projections of the training and three previously unseen test datasets onto the first two eigenvectors. The subspaces obtained from UDICA and DICA are more stable than for KPCA and COIR. In particular, COIR shows a substantial difference between training and test data, suggesting overfitting.

Gating of Flow Cytometry Data

Graft-versus-host disease (GvHD) occurs in allogeneic hematopoietic stem cell transplant recipients when donor-immune cells in the graft recognize the recipient as "foreign" and initiate an attack on the skin, gut, liver, and other tissues. It is a significant clinical problem in the field of allogeneic blood and marrow transplantation. The GvHD dataset (Brinkman et al. 2007) consists of weekly peripheral blood samples obtained from 31 patients following allogenic blood and marrow transplant. The goal of gating is to identify $\text{CD3}^+\text{CD4}^+\text{CD8}\beta^+$ cells, which were

Table 5.2: Average accuracies over 30 random subsamples of GvHD datasets. Pooling SVM applies standard kernel function on the pooled data from multiple domains, whereas distributional SVM also considers similarity between domains using kernel (5.22). With sufficiently many samples, DICA outperforms other methods in both pooling and distributional settings. The performance of pooling SVM and distributional SVM are comparable in this case.

Methods	**Pooling SVM**			**Distributional SVM**		
	$n_i = 100$	$n_i = 500$	$n_i = 1000$	$n_i = 100$	$n_i = 500$	$n_i = 1000$
Input	91.68±.91	92.11±1.14	93.57±.77	91.53±.76	92.81±.93	92.41±.98
KPCA	91.65±.93	92.06±1.15	93.59±.77	**91.83±.60**	90.86±1.98	92.61±1.12
COIR	**91.71±.88**	92.00±1.05	92.57±.97	91.42±.95	91.54±1.14	92.61±.89
UDICA	91.20±.81	92.21±.19	93.02±.77	91.51±.79	91.74±1.08	93.02±.77
DICA	91.37±.91	**92.71±.82**	**94.16±.73**	91.51±.89	**93.42±.73**	**93.33±.86**

found to have a high correlation with the development of GvHD (Brinkman et al. 2007). We expect to find a subspace of cells that is consistent to the biological variation between patients, and is indicative of the GvHD development. For each patient, we select a dataset that contains sufficient numbers of the target cell populations. As a result, we omit one patient due to insufficient data. The corresponding flow cytometry datasets from 30 patients have sample sizes ranging from 1,000 to 10,000, and the proportion of the $CD3^+CD4^+CD8\beta^+$ cells in each dataset ranges from 10% to 30%, depending on the development of the GvHD.

To evaluate the performance of the proposed algorithms, we took data from $N = 10$ patients for training, and the remaining 20 patients for testing. We subsample the training sets and test sets to have 100, 500, and 1,000 data points (cells) each. We compare the SVM classifiers under two settings, namely, a pooling SVM and a distributional SVM. The pooling SVM disregards the inter-patient variation by combining all datasets from different patients, whereas the distributional SVM also takes the inter-patient variation into account via the kernel function (Blanchard et al. 2011b)

$$K(\tilde{\mathbf{x}}_k^{(i)}, \tilde{\mathbf{x}}_l^{(j)}) = k_1(\mathbb{P}^i, \mathbb{P}^j) \cdot k_2(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) \tag{5.22}$$

where $\tilde{\mathbf{x}}_k^{(i)} = (\mathbb{P}^i, \mathbf{x}_k^{(i)})$ and k_1 is the kernel on distributions. We use the kernels $k_1(\mathbb{P}^i, \mathbb{P}^j) = \exp\left(-\|\boldsymbol{\mu}_{\mathbb{P}^i} - \boldsymbol{\mu}_{\mathbb{P}^j}\|^2_{\mathscr{H}}/2\sigma_1^2\right)$ and $k_2(\mathbf{x}_k^{(i)}, \mathbf{x}_l^{(j)}) = \exp(-\|\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}\|^2/2\sigma_2^2)$, where $\boldsymbol{\mu}_{\mathbb{P}^i}$ is computed using k_2. For pooling SVM, the kernel $k_1(\mathbb{P}^i, \mathbb{P}^j)$ is constant for any i and j. Moreover, we use the output kernel $l(\mathbf{y}_k^{(i)}, \mathbf{y}_l^{(j)}) = \delta(\mathbf{y}_k^{(i)}, \mathbf{y}_l^{(j)})$ where $\delta(a, b)$ is 1 if $a = b$, and 0 otherwise. We compare the performance of the SVMs trained on the preprocessed datasets using the KPCA, COIR, UDICA, and DICA algorithms. It is important to note that we are not defining another kernel on top of the preprocessed data. That is, the kernel k_2 for KPCA, COIR, UDICA, and DICA is exactly (5.13). We perform 10-fold cross validation on the parameter grids to optimize for accuracy.

Table 5.2 reports average accuracies and their standard deviation over 30 repetitions of the experiments. For sufficiently large number of samples, DICA outperforms other approaches. The pooling SVM and distributional SVM achieve comparable accuracies. Figure 5.13 depicts the leave-one-out accuracies of different approaches evaluated on each subject in the dataset. Average leave-one-out accuracies are reported in Table 5.3. The distributional SVM outperforms the pooling SVM in this setting, possibly because of the relatively large number of training subjects, *i.e.*, 29 subjects. Using the invariant features learnt by DICA also gives higher accuracies than other approaches.

Table 5.3: The average leave-one-out accuracies over 30 subjects on GvHD data. The distributional SVM outperforms the pooling SVM. DICA improves classifier accuracy.

Methods	Pooling	Distributional
Input	92.03±8.21	93.19±7.20
KPCA	91.99±9.02	93.11±6.83
COIR	92.40±8.63	92.92±8.20
UDICA	92.51±5.09	92.74±5.01
DICA	**92.72±6.41**	**94.80±3.81**

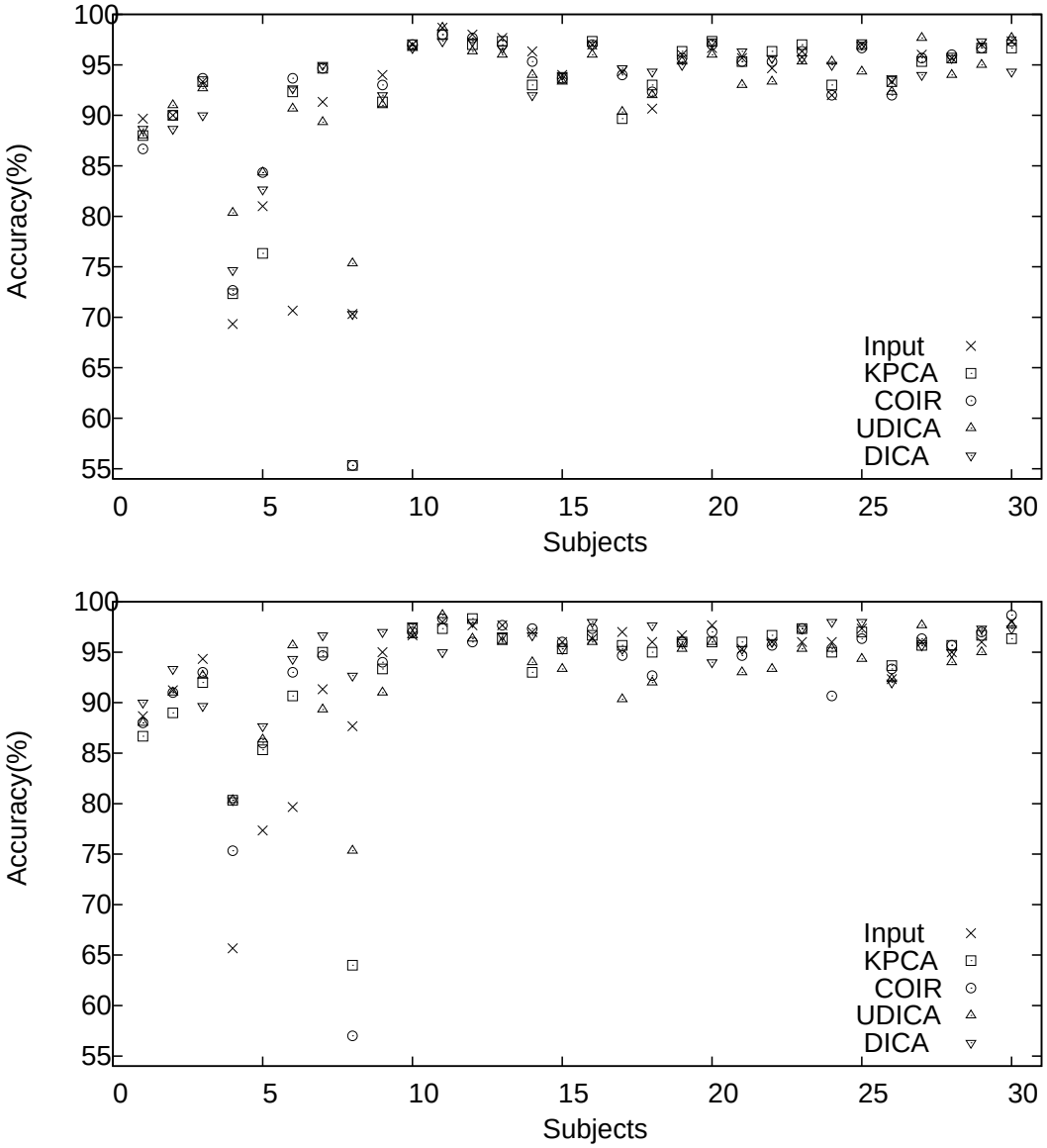


Figure 5.13: The leave-one-out accuracy of different methods evaluated on each subject in the GvHD dataset. The top figure depicts the pooling setting, whereas the bottom figure depicts the distributional setting.

Parkinson's Telemonitoring

To evaluate DICA in a regression setting, we apply it to a Parkinson's telemonitoring dataset[5]. The dataset consists of biomedical voice measurements from 42 people with early-stage Parkin-

[5] http://archive.ics.uci.edu/ml/datasets/Parkinson's+Telemonitoring

Table 5.4: Root mean square error (RMSE) of the independent Gaussian Process regression (GPR) applied to the Parkinson's telemonitoring dataset. DICA outperforms other approaches in both settings; and the distributional SVM outperforms the pooling SVM.

Methods	**Pooling GP Regression**		**Distributional GP Regression**	
	motor score	total score	motor score	total score
LLS	8.82 ± 0.77	11.80 ± 1.54	8.82 ± 0.77	11.80 ± 1.54
Input	9.58 ± 1.06	12.67 ± 1.40	8.57 ± 0.77	11.50 ± 1.56
KPCA	8.54 ± 0.89	11.20 ± 1.47	8.50 ± 0.87	11.22 ± 1.49
UDICA	8.67 ± 0.83	11.36 ± 1.43	8.75 ± 0.97	11.55 ± 1.52
COIR	9.25 ± 0.75	12.41 ± 1.63	9.23 ± 0.90	11.97 ± 2.09
DICA	$\mathbf{8.40 \pm 0.76}$	$\mathbf{11.05 \pm 1.50}$	$\mathbf{8.35 \pm 0.82}$	$\mathbf{10.02 \pm 1.01}$

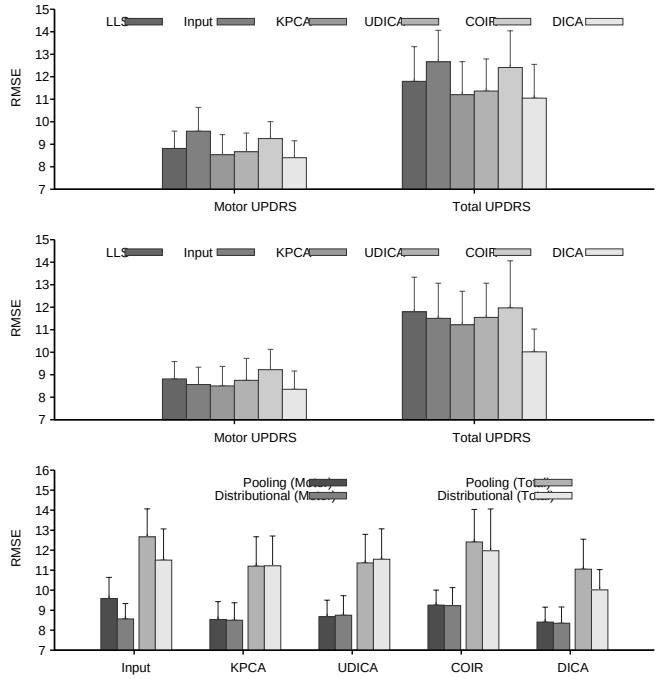


Figure 5.14: The root mean square error (RMSE) of motor and total UPDRS scores predicted by GP regression after different preprocessing methods on Parkinson's telemonitoring dataset. The top and middle rows depicts the pooling and distributional settings; the bottom row compares the two settings. Results of linear least square (LLS) are given as a baseline.

son's disease recruited for a six-month trial of a telemonitoring device for remote symptom progression monitoring. The aim is to predict the clinician's motor and total UPDRS scoring of Parkinson's disease symptoms from 16 voice measures. There are around 200 recordings per patient.

We adopt the same experimental settings as in §5.4.5, except that we employ two independent Gaussian Process (GP) regression to predict motor and total UPDRS scores. For COIR and DICA, we consider the output kernel $l(\mathbf{y}_k^{(i)}, \mathbf{y}_l^{(j)}) = \exp(-\|\mathbf{y}_k^{(i)} - \mathbf{y}_l^{(j)}\|^2/2\sigma_3^2)$ to fully account for the affinity structure of the output variable. We set σ_3 to be the median of motor and total UPDRS scores. The voice measurements from 30 patients are used for training and the rest for testing.

Fig. 5.14 depicts the results. DICA consistently, though not statistically significantly, outperforms other approaches, see Table 5.4. Inter-patient (*i.e.*, across domain) variation worsens

prediction accuracy on new patients. Reducing this variation with DICA improves the accuracy on new patients. Moreover, incorporating the inter-subject variation via distributional GP regression further improves the generalization ability, see Fig. 5.14.

5.4.6 Discussions

To conclude, we proposed a simple algorithm called Domain-Invariant Component Analysis (DICA) for learning an invariant transformation of the data which has proven significant for domain generalization both theoretically and empirically. Theorem 5.8 shows the generalization error on previously unseen domains grows with the distributional variance. We also showed that DICA generalizes KPCA and COIR, and is closely related to TCA. Finally, experimental results on both synthetic and real-world datasets show DICA performs well in practice. Interestingly, the results also suggest that the distributional SVM, which takes into account inter-domain variation, outperforms the pooling SVM which ignores it.

The motivating assumption in this work is that the functional relationship is stable or varies smoothly across domains. This is a reasonable assumption for automatic gating of flow cytometry data because the inter-subject variation of cell population makes it impossible for domain expert to apply the same gating on all subjects, and similarly makes sense for Parkinson's telemonitoring data. Nevertheless, the assumption does not hold in many applications where the conditional distributions are substantially different. It remains unclear how to develop techniques that generalize to previously unseen domains in these scenarios.

DICA can be adapted to novel applications by equipping the optimization problem with appropriate constraints. For example, one can formulate a semi-supervised extension of DICA by forcing the invariant basis functions to lie on a manifold or preserve a neighborhood structure. Moreover, by incorporating the distributional variance as a regularizer in the objective function, the invariant features and classifier can be optimized simultaneously.

END OF CHAPTER 5

Chapter 6

Conclusions and Future Research

The thesis introduces kernel-based frameworks for learning when the inputs are not just points, but probability measures. As demonstrated in this thesis, many real-world problems can in fact be viewed as learning problems on probability distributions. Probability distributions, as opposed to data points, constitute information at a higher level such as aggregate behavior of data points, how the underlying process evolves over time and domains, and a complex concept that cannot be described merely by data points. Most intelligent organisms have the ability to recognize and exploit such information naturally. Therefore, learning successfully on distributions can potentially shed light on future development of intelligent machines, and most importantly, may provide clues on the true meaning of intelligence.

The use of kernel mean embedding as a basic representation allows us to generalize many of the classical algorithms and establishes connections to existing frameworks. Through a comprehensive review in Chapter 2, it is evident that kernel mean embedding is a powerful representation of distributions. The dependence on kernel function makes it a flexible representation that can be adapted to any domains. It also permits one to model the underlying distribution without making any parametric assumption. Finally, its simplicity eases theoretical analysis and lends itself to better computational efficiency. These characteristics render kernel mean embedding increasingly appealing in the community compared to existing approaches based on density estimation, divergence measures, and information geometry, for example. Nevertheless, the review not only demonstrates the success of kernel mean embedding, but also reveals some limitations which could potentially lead to new research directions. Some of which have been investigated in this thesis. To the best of my knowledge, this is the first comprehensive review in this research area.

Kernel mean estimation is an essential step in modern applications of kernel mean embedding as well as many classical kernel-based algorithms. Chapter 3 shows that the standard kernel mean estimator can be improved, in particular, by the so-called *shrinkage estimator*. Motivated by James-Stein estimator, we propose a new family of estimators called *kernel mean shrinkage estimators* (KMSEs) which enjoy both theoretical guarantees and encouraging empirical results. Unlike James-Stein estimator, we provide some extensions using spectral filtering algorithms which are quite popular in the theory of inverse problems and regularization. This allows the estimators to take the geometrical property of the Hilbert space into account. Interestingly, the proposed idea can also be used to estimate other quantities such as (cross-) covariance operators which have been used in a wide range of applications. Our finding also provides a crucial clue to estimation in a "large d, small n" paradigm for RKHS when prior knowledge is not available. Last but not least, I believe this may eventually lead to a better understanding of the fundamental relationship between Tikhonov regularization and Stein shrinkage estimation in RKHS.

Chapter 4 provides a generalization of the empirical risk minimization (ERM) to a space of probability measures, *i.e.*, when we observe a sample $(\mathbb{P}_1, y_1), \ldots, (\mathbb{P}_n, y_n)$ rather than $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. I provide a representer theorem for distributions and show that the resulting framework amounts to constructing a kernel-based learning framework over a set of

distributions, each of which is represented by the kernel mean embedding. In particular, this chapter provides an extension of well-known support vector machine (SVM) to a space of probability distributions which we call a support measure machine (SMM). Connections to classical learning algorithms, possible extensions, and potential future directions are also discussed in this chapter. Chapter 5 then demonstrates the proposed framework in unsupervised setting, *i.e.*, when we only observe a sample $\mathbb{P}_1, \ldots, \mathbb{P}_n$. As mentioned earlier, representing data as probability distributions often leads to better performance compared to classical setting.

Last but not least, it has been pointed out that learning from distributions has potential applications in statistics. Many problems in statistics such as hypothesis testing involve finding a function of the empirical distribution to a certain set of outputs called *statistic*, *e.g.*, $\{-1, +1\}$ indicating whether or not to reject the null hypothesis. Conventional approach is to use *plug-in* estimators. On a contrary, if training data is available, we may *learn* such an estimator automatically from the data using the proposed frameworks. Preliminary results have demonstrated the effectiveness of this approach in real-world applications, *e.g.*, see Szabó et al. (2015), Lopez-Paz et al. (2015b).

Despite the success of kernel mean embedding, in my opinion, there are several open questions and possibilities for future research directions:

Kernel Choice and Interpretability. A kernel choice problem remains an ultimate open problem in kernel methods that is inherited by the kernel mean embedding. Despite some efforts to resolve this issue, *e.g.*, Gretton et al. (2012b), the kernel choice problem remains a key challenge. It is widely agreed that problem-specific knowledge should be taken into account when choosing the best kernel, but in some application domains, it may not be clear how to incorporate such knowledge. Moreover, it is also not easy to interpret the kernel mean representation and the true meaning of features remain obscure.

Bayesian Interpretation. What is 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.

Scalability. In the era of "big data", it is imparative 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.

High-dimensional Inference. In Chapter 3, we observe that the improvement of shrinkage estimator tends to increase as the data dimensionality increases. What is an underlying explanation? Kernel-based methods are known to be less prone to the *curse of dimensionality* compared to classical approaches such as density estimation, but little is known about underlying theoretical insight. Are we being too optimistic about learning with kernels in high-dimensional space? I believe this is one of the promising research directions in the kernel community.

Invariant Representation. 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).

Causality. Causal inference involves the investigation of how the distribution of outcome changes as we intervene on some other variable. Can we develop the specific framework for causal inference using kernel mean embeddding? 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. 2015b), but it remains a challenging problem. For example, Lopez-Paz et al. (2015b) 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. Can we use the kernel mean embedding to represent the counterfactual distribution?

END OF CHAPTER 6

Bibliography

I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.

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.

Y. Altun and A. J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *The 19th Annual Conference on Learning Theory (COLT)*, volume 4005, pages 139–153. Springer, 2006.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, Mar. 2012. ISSN 1935-8237. doi: 10.1561/2200000036.

S.-i. Amari. Information geometry in optimization, machine learning and statistical inference. *Frontiers of Electrical and Electronic Engineering in China*, 5(3):241–260, 2010. ISSN 1673-3460. doi: 10.1007/s11460-010-0101-3.

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.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*, pages 41–48. MIT Press, 2007.

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

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, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*. icml.cc / Omnipress, 2012.

C. R. Baker. Mutual information for gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970. doi: 10.1137/0119044.

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

P. Bartlett and J. Shawe-Taylor. Advances in kernel methods. chapter Generalization Performance of Support Vector Machines and Other Pattern Classifiers, pages 43–54. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003. ISSN 1532-4435.

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. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966. doi: 10.1214/aoms/1177699147.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

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.

J. O. Berger. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Annals of Statistics*, 4(1):223–226, 1976.

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

P. C. Bhat. Multivariate Analysis Methods in Particle Physics. *Ann.Rev.Nucl.Part.Sci.*, 61: 281–309, 2011.

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.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, pages 2137–2155, 2009.

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. 2011a.

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 24*, pages 2178–2186, 2011b.

D. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2001.

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. doi: 10.1007/BF01452844.

B. Boots, A. Gretton, and G. J. Gordon. Hilbert Space Embeddings of Predictive State Representations. In *Proc. 29th Intl. Conf. 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):49–57, 2006.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 11 2005. ISSN 1262-3318. doi: 10.1051/ps:2005018.

J. Bovy, J. F. Hennawi, D. W. Hogg, A. D. Myers, J. A. Kirkpatrick, D. J.Schlegel, N. P. Ross, E. S. Sheldon, I. D. McGreer, D. P. Schneider, and B. A. Weaver. Think outside the color box: Probabilistic target selection and the sdss-xdqso quasar targeting catalog. *The Astrophysical Journal*, 729(2):141, 2011.

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

R. R. Brinkman, M. Gasparetto, S.-J. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biol Blood Marrow Transplant*, 13(6):691–700, 2007. ISSN 1083-8791.

C. J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2, 2010.

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. doi: 10.1007/s10208-006-0196-8.

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.

P. K. Chan and M. V. Mahoney. Modeling multiple time series for anomaly detection. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)*, pages 90–97. IEEE Computer Society, 2005.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):1–58, July 2009.

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.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sept. 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411.

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. doi: 10.1145/1102351.1102371.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT '08, pages 38–53, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87986-2. doi: 10.1007/978-3-540-87987-9_8.

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.

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.

K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection in categorical datasets. In *ACM-SIGKDD*, pages 169–176. ACM, 2008.

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. doi: 10.1002/rsa.10073.

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

F. De la Torre and M. Black. Robust principal component analysis for computer vision. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 362–369, 2001.

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.

F. Dinuzzo and B. Schölkopf. The representer theorem for Hilbert spaces: a necessary and sufficient condition. In *Advances in Neural Information Processing Systems 25*, pages 189–196. 2012.

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. ISBN 3-540-79227-9, 978-3-540-79227-7.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

L. Fei-fei. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2005.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x.

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 J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in neural information processing systems 20*, pages 489–496, Red Hook, NY, USA, 9 2008. Curran. ISBN 978-1-605-60352-0.

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. doi: 10.1145/959242.959248.

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

L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

G. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

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. 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, 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 2009a.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. *Dataset Shift in Machine Learning*, chapter Covariate Shift by Kernel Mean Matching, pages 131–160. MIT Press, 2009b.

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.

M. Gruber. *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Statistics Textbooks and Monographs. Marcel Dekker, 1998.

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.

Q. Gu and J. Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 159–168. IEEE Computer Society, 2009.

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

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.

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. doi: 10.1214/aoms/1177728261.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948. doi: 10.1214/aoms/1177730196.

H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 06 2008. doi: 10.1214/009053607000000677.

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

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

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.

S.-Y. Huang, Y.-R. Yeh, and S. Eguchi. Robust kernel principal component analysis. *Neural Comput.*, 21:3179–3213, 2009.

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.

S. Ingrassia and G. D. Costanzo. Functional principal component analysis of financial time series. *New Developments in Classification and Data Analysis*, pages 351–358, 2005.

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 F. G. Cozman and A. Pfeffer, editors, *UAI*, pages 383–391. AUAI Press, 2011.

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

T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004b.

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. doi: 10.1007/978-3-642-04617-9_19.

D. M. Ji Zhao. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test, 2014.

I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

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.

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*, 13:2529âĹŠ–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. doi: 10.1109/TPAMI.2005.181.

M. Kim and V. Pavlovic. Central subspace dimensionality reduction using covariance operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):657–670, 2011.

J. A. Kirkpatrick, D. J. Schlegel, N. P. Ross, A. D. Myers, J. F. Hennawi, E. S.Sheldon, D. P. Schneider, and B. A. Weaver. A simple likelihood method for quasar target selection. *The Astrophysical Journal*, 743(2):125, 2011.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 02 2002.

S. Kpotufe. k-nn regression adapts to local intrinsic dimension. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 729–737. Curran Associates, Inc., 2011.

M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 37(2):233–243, 1991.

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. doi: 10.1109/tnn.2004.837781.

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 of *JMLR Workshop and Conference Proceedings*, San Diego, United States, May 2015.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

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

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, Berlin, May 1991.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, third edition, 1998.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwith selectors. *Annals of Statistics*, 25:929–947, 1997.

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

D. Lopez-Paz, K. Muandet, and B. Recht. The randomized causation coefficient. *Journal of Machine Learning*, 2015a.

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 (To Appear)*, 2015b.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, pages 1150–1157, Washington, DC, USA, 1999.

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. doi: 10.1109/ICRA.2013.6630971.

G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1992. ISBN 0-471-61531-5. A Wiley-Interscience Publication.

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. doi: 10.1098/rsta.1909.0016.

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.

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 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 Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1–9. Curran Associates, Inc., 2014b.

K. Muandet*, B. Sriperumburdur*, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 2015.

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

M. H. Nguyen and F. D. la Torre. Robust kernel principal component analysis. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1185–1192. 2009.

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, 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.

B. W. S. P. J. Green. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, 1994.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

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.

A. Passos, P. Rai, J. Wainer, and H. D. III. Flexible modeling of latent task structures in multitask learning. In *Proceedings of the 29th international conference on Machine learning*, Edinburgh, UK, 2012.

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.

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.

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.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS 2007 - Advances in Neural Information Processing Systems*, Dec. 2007.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition, 2005.

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. *Functional Analysis*. Academic Press, New York, 1972.

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.

N. P. Ross, A. D. Myers, E. S. Sheldon, C. Yéche, M. A. Strauss, J. Bovy, J. A. Kirkpatrick, G. T. Richards, È. Aubourg, M. R. Blanton, W. N. Brandt, W. C. Carithers, R. A. C. Croft, R. da Silva, K. Dawson, D. J. Eisenstein, J. F. Hennawi, S. Ho, D. W. Hogg, K.-G. Lee, B. Lundgren, R. G. McMahon, J. Miralda-Escudé, N. Palanque-Delabrouille, I. Pâris, P. Petitjean, M. M. Pieri, J. Rich, N. A. Roe, D. Schiminovich, D. J. Schlegel, D. P. Schneider, A. Slosar, N. Suzuki, J. L. Tinker, D. H. Weinberg, A. Weyant, M. White, and W. M. Wood-Vasey. The sdss-iii baryon oscillation spectroscopic survey: Quasar target selection for data release nine. *The Astrophysical Journal Supplement Series*, 199(1):3, 2012.

G. Sanguinetti and N. D. Lawrence. Missing data in kernel PCA. In *European Conference on Machine Learning*, pages 751–758, 2006.

Z. Sasvári. *Multivariate Characteristic and Correlation Functions*. De Gruyter, Berlin, Germany, 2013.

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, A. J. Smola, and K. R. Müller. Kernel principal component analysis. *Advances in kernel methods: support vector learning*, pages 327–352, 1999.

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, 2001a. Springer-Verlag. ISBN 3-540-42343-5.

B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001b.

B. Schölkopf, K. Muandet, K. Fukumizu, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 2015.

M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig. Non-linear PCA: a missing data approach. *Bioinformatics*, 21:3887–3895, 2005.

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, 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 2013. doi: 10.1214/13-AOS1140.

D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton. Kernel adaptive Metropolis-Hastings, Feb. 2014. URL http://arxiv.org/abs/1307.5302.

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.

H. L. Shang. A survey of functional principal component analysis. Monash econometrics and business statistics working papers, Monash University, Department of Econometrics and Business Statistics, 2011.

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.

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, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273599.

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.

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. Max-Planck-Gesellschaft, Curran. ISBN 978-1-615-67911-9.

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. doi: 10.1214/12-EJS722.

B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. Density estimation in infinite dimensional exponential families. 2013. http://arxiv.org/pdf/1312.3516.

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.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602. University of California Press, 1972.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002. ISSN 1532-4435. doi: 10.1162/153244302760185252.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

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. doi: 10.1007/s00365-012-9153-3.

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. doi: 10.1145/1831407.1831431.

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. doi: http://dx.doi.org/10.1016/j.jmva.2003.12.002.

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *Ann. Appl. Stat.*, 3(4):1236–1265, 12 2009. doi: 10.1214/09-AOAS312.

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. doi: 10.1214/009053607000000505.

D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199, 1999.

D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20 (3):1236–1265, 1992.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.

F. D. L. Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54:2003, 2003.

A. van der Linde. Variational bayesian functional PCA. *Comput. Stat. Data Anal.*, 53:517–533, 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

H. van Hoof, J. Peters, and G. Neumann. Learning of non-parametric control policies with high-dimensional state features. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

V. N. Vapnik. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4:831–838, 1992.

T. Vatanen, M. Kuusela, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai. Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *IJCNN*, pages 1–8. IEEE, 2012.

A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the International Conference on Computer Vision*, pages 606–613, 2009.

E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005.

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

E. D. Vito, L. Rosasco, and A. Toigo. Learning sets with separating kernels. *http://arxiv.org/abs/1204.3573*, April 2012.

R. Viviani, G. Groen, and M. Spitzer. Functional principal component analysis of fMRI data. *Human Brain Mapping*, 24:109–129, 2005.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundation and Trends in Machine Learning*, 1(1-2):1–305, 2008.

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.

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. doi: 10.1145/1553374.1553517.

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.

G. Widmer and M. Kurat. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23:69–101, 1996.

H.-M. Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

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.

Y. H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.*, 3 (8):579–588, 2002.

V. Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

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.

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.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning, W&CP 28 (3)*, pages 819–827. JMLR, 2013.

K. Zhang, B. Schölkopf, K. Muandet, Z. Wang, Z. Zhou, and C. Persello. *Single-Source Domain Adaptation with Target and Conditional Shift*, chapter 19, pages 427–456. Chapman & Hall/CRC Machine Learning & Pattern Recognition. Chapman and Hall/CRC, Boca Raton, USA, 2014.

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.

M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2250–2258, 2009.

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.

Appendix A

Oracle Inequalities for Kernel Mean Estimation

In this appendix, I provide an alternative way of computing the shrinkage parameter α of the shrinkage estimator $\hat{\boldsymbol{\mu}}_\alpha = \alpha f^* + (1-\alpha)\hat{\boldsymbol{\mu}}$ using an idea of unbiased estimator of the risk (Lehmann and Casella 1998; Chapter 5). To ease the analysis, I assume throughout that $f^* = 0$.

Recall that for some unknown distribution $\mathbb{P}$, our goal is to find an estimate $\hat{\boldsymbol{\mu}}_\alpha$ that minimizes the risk $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha)$. Hence, an optimal value of α can be obtained as a solution to the following minimization problem:

$$\alpha_* := \arg\min_\alpha R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha).$$

As a result, we may view $\hat{\boldsymbol{\mu}}_{\alpha_*}$ as an *oracle estimator* that outputs the best *linear* estimate of the true kernel mean. Since the underlying distribution $\mathbb{P}$ is unknown, the oracle value $\hat{\boldsymbol{\mu}}_{\alpha_*}$ is not an estimator and cannot be computed in practice. Based on the sample, we are interested in constructing a data-dependent estimator whose risk would converge to the risk of the oracle. To this end, we will first rely on the following assumption:

Assumption 1. *The risk of the standard kernel mean estimator $\hat{\boldsymbol{\mu}}$ is known and is given by $\Delta = \mathbb{E}_\mathbb{P}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2_\mathscr{H}\right]$.*

Assumption 1 is a weaker form of that in Theorem 3.3 that the true kernel mean $\boldsymbol{\mu}$ is known. That is, knowing only Δ is not sufficient to estimate the oracle value α_*. In other words, we are restricting the class of distributions that we are considering. For example, in Stein's setting it is assumed that $X \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ where σ is known. Consequently, the risk Δ is constant (see Example A.1 below). Note that in addition our setting involves a non-linear kernel k and for a certain class of distributions and kernels, we can compute Δ analytically (see Example A.2). Later, we will relax this assumption.

Next, we employ the idea of unbiased estimation of Δ_α. First, we observe that

$$\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|^2_\mathscr{H} = (1-\alpha)^2\|\hat{\boldsymbol{\mu}}\|^2_\mathscr{H} - 2(1-\alpha)\langle\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}\rangle_\mathscr{H} + \|\boldsymbol{\mu}\|^2_\mathscr{H}.$$

We then define a loss function

$$\mathcal{J}(\alpha) := (1-\alpha)^2\|\hat{\boldsymbol{\mu}}\|^2_\mathscr{H} - 2(1-\alpha)(\|\hat{\boldsymbol{\mu}}\|^2_\mathscr{H} - \Delta),$$

which under Assumption 1 is independent of the true $\boldsymbol{\mu}$. It is easy to show that

$$\mathbb{E}_\mathbb{P}[\mathcal{J}(\alpha)] \;=\; \mathbb{E}_\mathbb{P}\left[\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|^2_\mathscr{H}\right] - \|\boldsymbol{\mu}\|^2_\mathscr{H} = R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha) - \|\boldsymbol{\mu}\|^2_\mathscr{H}.$$

In other words, $\mathcal{J}(\alpha)$ is an unbiased estimator of the risk $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha)$ up to the additive term $\|\boldsymbol{\mu}\|^2_\mathscr{H}$ that is independent of α. Consequently, the minimizer of $\mathcal{J}(\alpha)$ should be close to the minimizer in α of $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha)$, *i.e.*, α_*. Since $\mathcal{J}(\alpha)$ does not depend on $\boldsymbol{\mu}$, we can define the

data-dependent shrinkage parameter as $\tilde{\alpha} \triangleq \arg\min_\alpha \mathcal{J}(\alpha)$ and the corresponding estimator $\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} = (1-\tilde{\alpha})\hat{\boldsymbol{\mu}}$.

The minimum of the risk among all estimators of the form $(1-\alpha)\hat{\boldsymbol{\mu}}$ is equal to

$$\min_\alpha \mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] = \min_\alpha \left[(1-\alpha)^2\Delta + \alpha^2\|\boldsymbol{\mu}\|_{\mathscr{H}}^2\right] = \frac{\Delta\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2},$$

and the value of oracle α that achieves this minimum is

$$\alpha_* = \frac{\Delta}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2}.$$

Likewise, the unbiased estimator of the risk has the form $\mathcal{J}(\alpha) = (\alpha^2-1)\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 + (2-2\alpha)\Delta$ and the minimizer of this expression is

$$\tilde{\alpha} = \frac{\Delta}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}. \tag{A.1}$$

Note that (A.1) is slightly different from (3.20). Under Assumption 1, the shrinkage parameter computed from (A.1) does not depend on the true $\boldsymbol{\mu}$ whatsoever and can be obtained directly from the sample. Below we give the oracle inequality of the shrinkage estimator obtained via (A.1).

Theorem A.1. *Under Assumption 1, the oracle inequality*

$$\mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] \leq \min_\alpha \mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] + 2\Delta\mathbb{E}_{\mathbb{P}}\left[\frac{\Delta - \langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right]$$

holds for all distributions $\mathbb{P}$ and kernel k.

Proof. It is not difficult to show that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] &= \Delta + \Delta^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\Delta\mathbb{E}_{\mathbb{P}}\left[\frac{\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \\
&= \Delta - \Delta^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] + 2\Delta^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\Delta\mathbb{E}_{\mathbb{P}}\left[\frac{\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right].
\end{aligned}$$

By Jensen's inequality,

$$\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \geq \frac{1}{\mathbb{E}_{\mathbb{P}}[\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2]} = \frac{1}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2}.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] &\leq \Delta - \frac{\Delta^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + 2\Delta^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\Delta\mathbb{E}_{\mathbb{P}}\left[\frac{\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \\
&= \frac{\Delta\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + 2\Delta\mathbb{E}_{\mathbb{P}}\left[\frac{\Delta - \langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right].
\end{aligned}$$

The oracle inequality follows from the fact that $\min_\alpha \mathbb{E}_{\mathbb{P}}[\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|_{\mathscr{H}}^2] = \frac{\Delta\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2}$. This completes the proof. ■

Theorem A.1 shows that the risk of estimator obtained via (A.1) is equal to the minimal risk of the oracle estimator up to a summand that does not depend on α. Next, we address the question of how to estimate Δ. As mentioned earlier, we can compute Δ analytically for a certain class of distributions and kernels. We give some concrete examples below.

Example A.1. *Consider a class of Gaussian distributions $\mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d)$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ and a linear kernel $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$. In this case we have*

$$\Delta = \mathbb{E}_{\mathbb{P}} \left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \right] = d\sigma^2,$$

which only requires the knowledge of the variance σ^2 (or $\operatorname{trace}(\boldsymbol{\Sigma})$ for more general covariance structure) of the distribution.

Example A.2. *Consider a class of Gaussian distributions $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-0.5\gamma \|\mathbf{x} - \mathbf{y}\|^2)$. Recall that*

$$\Delta = \mathbb{E}_{\mathbb{P}} \left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \right] = \frac{1}{n} (\mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}) - \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}})),$$

where $\tilde{\mathbf{x}}$ is an independent copy of $\mathbf{x}$. We have that $\mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}) = \exp(0) = 1$ and it is not difficult to show that (Muandet et al. 2012; Table 1)

$$\mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}}} k(\mathbf{x}, \tilde{\mathbf{x}}) = \left\langle \int_{\mathcal{X}} k(\mathbf{x}, \cdot) \, \mathrm{d}\mathcal{N}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\Sigma}), \int_{\mathcal{X}} k(\tilde{\mathbf{x}}, \cdot) \, \mathrm{d}\mathcal{N}(\tilde{\mathbf{x}}; \boldsymbol{\theta}, \boldsymbol{\Sigma}) \right\rangle_{\mathscr{H}} = \frac{1}{\sqrt{|2\gamma\boldsymbol{\Sigma} + \mathbf{I}|}}.$$

As a result, we get

$$\Delta = \frac{1}{n} \left(1 - \frac{1}{\sqrt{|2\gamma\boldsymbol{\Sigma} + \mathbf{I}|}} \right).$$

Similarly, this only requires the knowledge of the covariance matrix $\boldsymbol{\Sigma}$ associated to the class of distributions.

Unfortunately, the use of shrinkage estimator under Assumption 1 is quite restrictive. In general, the knowledge of the risk Δ may not be available at all and we only have access to the sample. In that case, we can relax Assumption 1 and resort to the unbiased estimate $\hat{\Delta}$ obtained from the sample. The empirical loss function can be defined accordingly as

$$\widehat{\mathcal{J}}(\alpha) \triangleq (1 - \alpha)^2 \|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - 2(1 - \alpha)(\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2 - \hat{\Delta}).$$

It follows that $\mathbb{E}_{\mathbb{P}}[\widehat{\mathcal{J}}(\alpha)] = \mathbb{E}_{\mathbb{P}}[\mathcal{J}(\alpha)] = R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha) - \|\boldsymbol{\mu}\|_{\mathscr{H}}^2$. Hence, $\widehat{\mathcal{J}}(\alpha)$ is also an unbiased estimator of the risk $R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_\alpha)$ up to a summand that does not depend on α. Moreover, $\widehat{\mathcal{J}}(\alpha) - \mathcal{J}(\alpha) = 2(1 - \alpha)(\hat{\Delta} - \Delta)$, suggesting that the quality of an estimate $\widehat{\mathcal{J}}(\alpha)$ depends on how well one can estimate $\hat{\Delta}$. We showed in the proof of Proposition 3.7 that $|\hat{\Delta} - \Delta|$ goes to zero sufficiently fast.

The following theorem, which is similar to Theorem A.1, gives the oracle inequality when the empirical version of the risk Δ is used to construct the shrinkage estimator. Assumption 1 is not needed here.

Theorem A.2. *Suppose that the kernel k satisfies $\sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) \leq R$. Define*

$$\tilde{\alpha} \triangleq \arg\min_{\alpha} \widehat{\mathcal{J}}(\alpha) = \frac{\hat{\Delta}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2} \quad \text{and} \quad R_n \triangleq \frac{R}{n}.$$

Then, the oracle inequality

$$\mathbb{E}_{\mathbb{P}} \left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \right] \leq \min_{\alpha} \mathbb{E}_{\mathbb{P}} \left[\|\hat{\boldsymbol{\mu}}_{\alpha} - \boldsymbol{\mu}\|_{\mathscr{H}}^2 \right] + \frac{\Delta^2 - R_n^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + 2\mathbb{E}_{\mathbb{P}} \left[\frac{R_n^2 - \hat{\Delta} \langle \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2} \right]$$

holds for all distributions $\mathbb{P}$.

Proof. First, it is not difficult to see that

$$\hat{\Delta} = \frac{1}{n}(\widehat{\mathbb{E}}_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) - \widehat{\mathbb{E}}_{\mathbf{x},\tilde{\mathbf{x}}} k(\mathbf{x},\tilde{\mathbf{x}})) \leq \frac{R}{n} =: R_n.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] &= \Delta + \mathbb{E}_{\mathbb{P}}\left[\frac{\hat{\Delta}^2}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\mathbb{E}_{\mathbb{P}}\left[\frac{\hat{\Delta}\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \\
&\leq \Delta + R_n^2\mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\mathbb{E}\left[\frac{\hat{\Delta}\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \\
&= \Delta - R_n^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] + 2R_n^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\mathbb{E}_{\mathbb{P}}\left[\frac{\hat{\Delta}\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right],
\end{aligned}$$

where the inequality follows from $\hat{\Delta} \leq R_n$. By Jensen's inequality,

$$\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \geq \frac{1}{\mathbb{E}_{\mathbb{P}}[\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2]} = \frac{1}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2}.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|\hat{\boldsymbol{\mu}}_{\tilde{\alpha}} - \boldsymbol{\mu}\|_{\mathscr{H}}^2\right] &\leq \Delta - \frac{R_n^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + 2R_n^2\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] - 2\mathbb{E}_{\mathbb{P}}\left[\frac{\hat{\Delta}\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right] \\
&= \frac{\Delta\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + \frac{\Delta^2 - R_n^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2} + 2\mathbb{E}_{\mathbb{P}}\left[\frac{R_n^2 - \hat{\Delta}\langle\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle_{\mathscr{H}}}{\|\hat{\boldsymbol{\mu}}\|_{\mathscr{H}}^2}\right].
\end{aligned}$$

The oracle inequality follows from the fact that $\min_\alpha \mathbb{E}[\|\hat{\boldsymbol{\mu}}_\alpha - \boldsymbol{\mu}\|_{\mathscr{H}}^2] = \frac{\Delta\|\boldsymbol{\mu}\|_{\mathscr{H}}^2}{\Delta + \|\boldsymbol{\mu}\|_{\mathscr{H}}^2}$. This completes the proof. ■

In summary, Theorem A.1 and A.2 basically show that the risk of shrinkage estimator obtained via (A.1) is close to that of the oracle estimator up to a residual term that does not depend on α and vanishes as $n \to \infty$. One of the future works in this direction is to understand how fast this residual term goes to zero compared to the risk of oracle estimator.

Appendix B

Leave-One-Out Cross Validation Score

Here I propose an alternative approach for computing the shrinkage parameter for S-KMSE. In this case, we consider a slightly different leave-one-out estimator $\hat{\boldsymbol{\mu}}_\lambda^{(-i)} = \sum_{j=1}^n \beta_j^{(-i)} \phi(\mathbf{x}_j)$ where

$$\boldsymbol{\beta}^{(-i)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \sum_{j\neq i} \left\| \phi(\mathbf{x}_j) - \sum_{k=1}^n \beta_k \phi(\mathbf{x}_i) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}\|^2.$$

By adopting an approach similar to the one presented in P. J. Green (1994; Lemma 3.1) for the ridge regression problem, we can simplify the score so that it can be evaluated efficiently. First, we show that the leave-one-out solution $\boldsymbol{\beta}^{(-i)}$ can be obtained via the standard formulation with modified target vector.

Lemma B.1. *For fixed λ and i, let $\boldsymbol{\beta}^{(-i)}$ denote the vector with components $\beta_j^{(-i)}$ for $j \neq i$. Define a vector $\Phi^* := [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_{i-1}), \hat{\boldsymbol{\mu}}_\lambda^{(-i)}, \phi(\mathbf{x}_{i+1}), \ldots, \phi(\mathbf{x}_n)]^\top$ and a matrix $\mathbf{B}^*_{ml} := \langle \phi(\mathbf{x}_m), \Phi_l^* \rangle_{\mathscr{H}}$ where $\phi(\mathbf{x}_i) := k(\cdot, \mathbf{x}_i)$. Then $\boldsymbol{\beta}^{(-i)} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{B}^* \mathbf{1}_n$.*

Proof of Lemma B.1. For any vector $\boldsymbol{\beta}$,

$$\begin{aligned}
\sum_{j=1}^n \left\| \Phi_j^* - \sum_{k=1}^n \beta_k \phi(\mathbf{x}_k) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}\|^2 &\geq \sum_{j\neq i} \left\| \Phi_j^* - \sum_{k=1}^n \beta_k \phi(\mathbf{x}_k) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}\|^2 \\
&\geq \sum_{j\neq i} \left\| \Phi_j^* - \sum_{k=1}^n \beta_k^{(-i)} \phi(\mathbf{x}_k) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}^{(-i)}\|^2 \\
&= \sum_{j=1}^n \left\| \Phi_j^* - \sum_{k=1}^n \beta_k^{(-i)} \phi(\mathbf{x}_k) \right\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}^{(-i)}\|^2
\end{aligned}$$

by the definition of $\boldsymbol{\beta}^{(-i)}$ and the fact that $\Phi_i^* = \hat{\boldsymbol{\mu}}_\lambda^{(-i)}$. It follows that $\boldsymbol{\beta}^{(-i)}$ is the minimizer of $\sum_j \|\Phi_j^* - \sum_k \beta_k \phi(\mathbf{x}_k)\|_{\mathscr{H}}^2 + \lambda \|\boldsymbol{\beta}\|^2$ so that $\boldsymbol{\beta}^{(-i)} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{B}^* \mathbf{1}_n$, as required. ■

As we can see, the resulting formulation of $\boldsymbol{\beta}^{(-i)}$ in Lemma B.1 depends on the leave-one-out solution $\hat{\boldsymbol{\mu}}_\lambda^{(-i)}$ which in turn requires a knowledge of $\boldsymbol{\beta}^{(-i)}$. As a result, we cannot use this formulation to compute $\boldsymbol{\beta}^{(-i)}$ in practice. However, it will be used as an intermediate step for deriving the LOOCV score in the following proposition.

Proposition B.2. *The LOOCV score of S-KMSE defined above is given by*

$$LOOCV(\lambda) = \frac{1}{n} \sum_{i=1}^n (\mathbf{K}\boldsymbol{\beta} - \mathbf{k}_i)^\top \mathbf{C}_\lambda (\mathbf{K}\boldsymbol{\beta} - \mathbf{k}_i)$$

where $\boldsymbol{\beta} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$ is the weight vector computed on $\{\mathbf{x}_i\}_{i=1}^n$ with shrinkage parameter λ, $\mathbf{C}_\lambda \triangleq (\mathbf{K} - \frac{1}{n}\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{K})^{-1} \mathbf{K} (\mathbf{K} - \frac{1}{n}\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{K})^{-1}$ and $\mathbf{k}_i$ is the i^{th} column of $\mathbf{K}$.

Proof of Proposition B.2. Let $\mathbf{A} := (\mathbf{K} + \lambda \mathbf{I})^{-1}$. By virtue of Lemma B.1, we can write an expression for the deleted residual $\phi(\mathbf{x}_i) - \hat{\boldsymbol{\mu}}_\lambda^{(-i)}$ as

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_\lambda^{(-i)} - \phi(\mathbf{x}_i) &= \sum_{j=1}^{n} \beta_j^{(-i)} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m=1}^{n} \left\{ \mathbf{A}\mathbf{B}^* \right\}_{jm} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m \neq i} \left\{ \mathbf{A}\mathbf{K} \right\}_{jm} \phi(\mathbf{x}_j) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \mathbf{B}^*_{li} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m \neq i} \left\{ \mathbf{A}\mathbf{K} \right\}_{jm} \phi(\mathbf{x}_j) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \hat{\boldsymbol{\mu}}_\lambda^{(-i)} \rangle \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m=1}^{n} \left\{ \mathbf{A}\mathbf{K} \right\}_{jm} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&\quad - \frac{1}{n} \sum_{j=1}^{n} \left\{ \mathbf{A}\mathbf{K} \right\}_{ji} \phi(\mathbf{x}_j) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \hat{\boldsymbol{\mu}}_\lambda^{(-i)} \rangle \phi(\mathbf{x}_j) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m=1}^{n} \left\{ \mathbf{A}\mathbf{K} \right\}_{jm} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) \\
&\quad - \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \phi(\mathbf{x}_i) \rangle \phi(\mathbf{x}_j) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \hat{\boldsymbol{\mu}}_\lambda^{(-i)} \rangle \phi(\mathbf{x}_j) \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{m=1}^{n} \left\{ \mathbf{A}\mathbf{K} \right\}_{jm} \phi(\mathbf{x}_j) - \phi(\mathbf{x}_i) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \hat{\boldsymbol{\mu}}_\lambda^{(-i)} - \phi(\mathbf{x}_i) \rangle \phi(\mathbf{x}_j) \\
&= \hat{\boldsymbol{\mu}}_\lambda - \phi(\mathbf{x}_i) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \hat{\boldsymbol{\mu}}_\lambda^{(-i)} - \phi(\mathbf{x}_i) \rangle \phi(\mathbf{x}_j).
\end{aligned}$$

Denote the deleted residual $\hat{\boldsymbol{\mu}}_\lambda^{(-i)} - \phi(\mathbf{x}_i)$ by $\Delta_\lambda^{(-i)}$. Then, the above equation can be rewritten as

$$\Delta_\lambda^{(-i)} = \hat{\boldsymbol{\mu}}_\lambda - \phi(\mathbf{x}_i) + \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbf{A}_{jl} \langle \phi(\mathbf{x}_l), \Delta_\lambda^{(-i)} \rangle \phi(\mathbf{x}_j). \tag{B.1}$$

Since $\Delta_\lambda^{(-i)}$ lies in $\mathscr{H}$, we may decompose it as $\Delta_\lambda^{(-i)} = \sum_{k=1}^{n} \xi_k \phi(\mathbf{x}_k) + h_\perp$ for some $\boldsymbol{\xi} \in \mathbb{R}^n$ where $h_\perp$ is orthogonal to the span of $\{\phi(\mathbf{x}_k)\}_{k=1}^n$. Substituting this back into (B.1) and rearranging terms yields

$$\sum_{k=1}^{n} \xi_k \phi(\mathbf{x}_k) + h_\perp = \hat{\boldsymbol{\mu}}_\lambda - \phi(\mathbf{x}_i) + \frac{1}{n} \sum_{j=1}^{n} \left\{ \mathbf{A}\mathbf{K}\boldsymbol{\xi} \right\}_j \phi(\mathbf{x}_j).$$

By taking the inner product on both sides of the equation w.r.t. the samples $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$, the optimal $\boldsymbol{\xi}$ can be obtained by solving the system of equations $\mathbf{K}\boldsymbol{\xi} = \boldsymbol{\beta}^\top \mathbf{K} - \mathbf{k}_i + \frac{1}{n}\mathbf{K}\mathbf{A}\mathbf{K}\boldsymbol{\xi}$ whose solution is $\boldsymbol{\xi} = (\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbf{A}\mathbf{K})^{-1}(\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{k}_i)$ where $\mathbf{k}_i$ denotes the ith column of matrix $\mathbf{K}$. Consequently, the leave-one-out cross-validation score for the sample $\mathbf{x}_i$ can be computed by

$$\left\| \Delta_\lambda^{(-i)} \right\|_{\mathscr{H}}^2 = \boldsymbol{\xi}^\top \mathbf{K} \boldsymbol{\xi} = (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{k}_i)^\top \mathbf{C}_\lambda (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{k}_i)$$

where $\mathbf{C}_\lambda = (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}$. Lastly, the score over the full dataset can be obtained by averaging $\|\Delta_\lambda^{(-i)}\|^2_{\mathscr{H}}$ over all i. This concludes the proof. ■

We can see that the LOOCV score in Proposition B.2 depends only on the non-leave-one-out solution $\boldsymbol{\beta}_\lambda$, which can be obtained as a by-product of the algorithm.

Appendix C

Proofs

This section contains supplementary proofs of results presented in the thesis.

C.1 Proof of Lemma 3.1

Proof. This is the standard proof of Stein's lemma (Stein 1972). Let X be a random variable distributed according to a standard normal distribution and g be an absolutely continuous function. Then, we have

$$\begin{aligned}
\mathbb{E}[g'(X)] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x) e^{-\frac{x^2}{2}} \,\mathrm{d}x \\
&\overset{(*)}{=} \frac{1}{\sqrt{2\pi}} \left\{ \int_0^{\infty} g'(x) \,\mathrm{d}x \int_x^{\infty} y e^{-\frac{y^2}{2}} \,\mathrm{d}y - \int_{-\infty}^{0} g'(x) \,\mathrm{d}x \int_{-\infty}^{x} y e^{-\frac{y^2}{2}} \,\mathrm{d}y \right\} \\
&= \frac{1}{\sqrt{2\pi}} \left\{ \int_0^{\infty} y e^{-\frac{y^2}{2}} \,\mathrm{d}y \int_0^{y} g'(x) \,\mathrm{d}x - \int_{-\infty}^{0} y e^{-\frac{y^2}{2}} \,\mathrm{d}y \int_y^{0} g'(x) \,\mathrm{d}x \right\} \\
&= \frac{1}{\sqrt{2\pi}} \left\{ \int_0^{\infty} (g(x) - g(0)) y e^{-\frac{y^2}{2}} \,\mathrm{d}y + \int_{-\infty}^{0} (g(y) - g(0)) y e^{-\frac{y^2}{2}} \,\mathrm{d}y \right\} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y g(y) e^{-\frac{y^2}{2}} \,\mathrm{d}y \\
&= \mathbb{E}[X g(X)],
\end{aligned}$$

where we invoked the Fubini's theorem in $(*)$. ∎

C.2 Proof of Theorem 3.15

Proof. Since $(e_i)_i$ is an orthonormal basis in $\mathscr{H}$, we have for any $\mathbb{P}$ and $f^* \in \mathscr{H}$

$$\boldsymbol{\mu}_{\mathbb{P}} = \sum_{i=1}^{\infty} \mu_i e_i, \quad \hat{\boldsymbol{\mu}}_{\mathbb{P}} = \sum_{i=1}^{\infty} \hat{\mu}_i e_i, \quad \text{and} \quad f^* = \sum_{i=1}^{\infty} f_i^* e_i,$$

where $\mu_i := \langle \boldsymbol{\mu}_{\mathbb{P}}, e_i \rangle$, $\hat{\mu}_i := \langle \hat{\boldsymbol{\mu}}_{\mathbb{P}}, e_i \rangle$, and $f_i^* := \langle f^*, e_i \rangle$. If follows from the Parseval's identity that

$$\begin{aligned}
\Delta &= \mathbb{E}_{\mathbb{P}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^{\infty} (\hat{\mu}_i - \mu_i)^2 \right] =: \sum_{i=1}^{\infty} \Delta_i \\
\Delta_{\boldsymbol{\alpha}} &= \mathbb{E}_{\mathbb{P}} \|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\mu}\|^2 = \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^{\infty} (\alpha_i f_i^* + (1 - \alpha_i)\hat{\mu}_i - \mu_i)^2 \right] =: \sum_{i=1}^{\infty} \Delta_{\boldsymbol{\alpha},i}.
\end{aligned}$$

Note that the problem has not changed and we are merely looking at it from a different perspective. To estimate $\boldsymbol{\mu}_{\mathbb{P}}$, we may just as well estimate its Fourier coefficient sequence μ_i with $\hat{\mu}_i$. Based on above decomposition, we may write the risk difference $\Delta_{\boldsymbol{\alpha}} - \Delta$ as $\sum_{i=1}^{\infty}(\Delta_{\boldsymbol{\alpha},i} - \Delta_i)$. We can thus ask under which conditions on $\boldsymbol{\alpha} = (\alpha_i)$ for which $\Delta_{\boldsymbol{\alpha},i} - \Delta_i < 0$ uniformly over all i.

For each coordinate i, we have

$$\begin{aligned}
\Delta_{\boldsymbol{\alpha},i} - \Delta_i &= \mathbb{E}_{\mathbb{P}}\left[(\alpha_i f_i^* + (1-\alpha_i)\hat{\mu}_i - \mu_i)^2\right] - \mathbb{E}_{\mathbb{P}}\left[(\hat{\mu}_i - \mu_i)^2\right] \\
&= \mathbb{E}_{\mathbb{P}}[\alpha_i^2 f_i^2 + 2\alpha_i f_i^*(1-\alpha_i)\hat{\mu}_i + (1-\alpha_i)^2\hat{\mu}_i^2 \\
&\quad -2\alpha_i f_i^*\mu_i - 2(1-\alpha_i)\hat{\mu}_i\mu_i + \mu_i^2] - \mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2 - 2\hat{\mu}_i\mu_i + \mu_i^2] \\
&= \alpha_i^2 f_i^2 + 2\alpha_i f_i^*\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i] - 2\alpha_i^2 f_i^*\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i] + (1-\alpha_i)^2\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] \\
&\quad -2\alpha_i f_i^*\mu_i - 2(1-\alpha_i)\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i]\mu_i + \mu_i^2 - \mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] + 2\mu_i\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i] - \mu_i^2 \\
&= \alpha_i^2 f_i^2 - 2\alpha_i^2 f_i^*\mu_i + (1-\alpha_i)^2\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] - 2(1-\alpha_i)\mu_i^2 + 2\mu_i^2 - \mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] \\
&= \alpha_i^2 f_i^2 - 2\alpha_i^2 f_i^*\mu_i + (\alpha_i^2 - 2\alpha_i)\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] + 2\alpha_i\mu_i^2.
\end{aligned}$$

Next, we substitute $\mathbb{E}_{\mathbb{P}}[\hat{\mu}_i^2] = \mathbb{E}_{\mathbb{P}}[(\hat{\mu}_i - \mu_i + \mu_i)^2] = \Delta_i + \mu_i^2$ into the last equation to obtain

$$\begin{aligned}
\Delta_{\boldsymbol{\alpha},i} - \Delta_i &= \alpha_i^2 f_i^2 - 2\alpha_i^2 f_i^*\mu_i + \alpha_i^2(\Delta_i + \mu_i^2) - 2\alpha_i(\Delta_i + \mu_i^2) + 2\alpha_i\mu_i^2 \\
&= \alpha_i^2 f_i^2 - 2\alpha_i^2 f_i^*\mu_i + \alpha_i^2\Delta_i + \alpha_i^2\mu_i^2 - 2\alpha_i\Delta_i \\
&= \alpha_i^2(f_i^2 - 2f_i^*\mu_i + \Delta_i + \mu_i^2) - 2\alpha_i\Delta_i \\
&= \alpha_i^2(\Delta_i + (f_i^* - \mu_i)^2) - 2\alpha_i\Delta_i
\end{aligned}$$

which is negative if α_i satisfies

$$0 < \alpha_i < \frac{2\Delta_i}{\Delta_i + (f_i^* - \mu_i)^2}.$$

This completes the proof. ■

C.3 Proof of Proposition 3.16

Proof. Let $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ be an eigen-decomposition of $\mathbf{K}$ where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ consists of orthogonal eigenvectors of $\mathbf{K}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{D} = \operatorname{diag}(\gamma_1, \gamma_2 \ldots, \gamma_n)$ consists of corresponding eigenvalues. As a result, the coefficients $\boldsymbol{\beta}(\lambda)$ can be written as

$$\boldsymbol{\beta}(\lambda) = g_\lambda(\mathbf{K})\mathbf{K}\mathbf{1}_n = \mathbf{U}g_\lambda(\mathbf{D})\mathbf{U}^\top\mathbf{K}\mathbf{1}_n = \sum_{i=1}^{n}\mathbf{u}_i g_\lambda(\gamma_i)\mathbf{u}_i^\top\mathbf{K}\mathbf{1}_n. \tag{C.1}$$

Using $\mathbf{K}\mathbf{1}_n = [\langle\hat{\boldsymbol{\mu}}, k(\mathbf{x}_1, \cdot)\rangle, \ldots, \langle\hat{\boldsymbol{\mu}}, k(\mathbf{x}_n, \cdot)\rangle]^\top$, we can rewrite (C.1) as

$$\begin{aligned}
\boldsymbol{\beta}(\lambda) &= \sum_{i=1}^{n}\mathbf{u}_i g_\lambda(\gamma_i)\sum_{j=1}^{n}u_{ij}\langle\hat{\boldsymbol{\mu}}, k(\mathbf{x}_j, \cdot)\rangle \\
&= \sum_{i=1}^{n}\sqrt{\gamma_i}\mathbf{u}_i g_\lambda(\gamma_i)\left\langle\hat{\boldsymbol{\mu}}, \frac{1}{\sqrt{\gamma_i}}\sum_{j=1}^{n}u_{ij}k(\mathbf{x}_j, \cdot)\right\rangle,
\end{aligned}$$

where u_{ij} is the jth component of $\mathbf{u}_i$. Next, we invoke the relation between the eigenvectors of the matrix $\mathbf{K}$ and the eigenfunctions of the empirical covariance operator $\widehat{\mathbf{C}}_{XX}$ in $\mathscr{H}$. That is, it

is known that the ith eigenfunction of $\widehat{\mathbf{C}}_{XX}$ can be expressed as $\phi_i = (1/\sqrt{\gamma_i}) \sum_{j=1}^{n} u_{ij} k(\mathbf{x}_j, \cdot)$ (Schölkopf et al. 1999). Consequently,

$$\left\langle \hat{\boldsymbol{\mu}}, \frac{1}{\sqrt{\gamma_i}} \sum_{j=1}^{n} u_{ij} k(\mathbf{x}_j, \cdot) \right\rangle = \langle \hat{\boldsymbol{\mu}}, \phi_i \rangle$$

and we can write the Spectral-KMSE as

$$\begin{aligned} \hat{\mu}_\lambda &= \sum_{j=1}^{n} \left[\sum_{i=1}^{n} u_{ij} \sqrt{\gamma_i} g_\lambda(\gamma_i) \langle \hat{\boldsymbol{\mu}}, \phi_i \rangle \right]_j k(\mathbf{x}_j, \cdot) \\ &= \sum_{i=1}^{n} \sqrt{\gamma_i} g_\lambda(\gamma_i) \langle \hat{\boldsymbol{\mu}}, \phi_i \rangle \sum_{j=1}^{n} u_{ij} k(\mathbf{x}_j, \cdot) \\ &= \sum_{i=1}^{n} g_\lambda(\gamma_i) \gamma_i \langle \hat{\boldsymbol{\mu}}, \phi_i \rangle \phi_i. \end{aligned}$$

This completes the proof. ■

C.4 Proof of Theorem 3.17

Proof. Consider the following decomposition

$$\begin{aligned} \hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_{\mathbb{P}} &= \widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) \hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}} \\ &= \widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) (\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}) + \widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) \boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}} \\ &= \widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) (\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}) + (\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I) \widehat{\mathbf{C}}_{XX}^{\beta} h \\ &\quad + (\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I)(\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}) h \end{aligned}$$

where we used the fact that there exists $h \in \mathscr{H}$ such that $\boldsymbol{\mu}_{\mathbb{P}} = \mathbf{C}_{XX}^{\beta} h$ as we assumed that $\boldsymbol{\mu}_{\mathbb{P}} \in \mathcal{R}(\mathbf{C}_{XX}^{\beta})$ for some $\beta > 0$. Therefore

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_{\mathbb{P}}\| &\leq \|\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX})\|_{op} \|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\| \\ &\quad + \|(\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I) \widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \|h\| \\ &\quad + \|\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I\|_{op} \|\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \|h\| \end{aligned}$$

where we used the fact that $\|Ab\| \leq \|A\|_{op} \|b\|$ with $A : \mathscr{H} \to \mathscr{H}$ being a bounded operator, $b \in \mathscr{H}$ and $\|\cdot\|_{op}$ denoting the operator norm defined as $\|A\|_{op} := \sup\{\|Ab\| : \|b\| = 1\}$.

By $(C1)$, $(C2)$ and $(C3)$ in Definition 3.1, we have $\|\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX})\|_{op} \leq B$, $\|\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I\|_{op} \leq C$, and $\|(\widehat{\mathbf{C}}_{XX} g_\lambda(\widehat{\mathbf{C}}_{XX}) - I)\widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \leq D\lambda^{\min\{\beta, \eta_0\}}$, respectively. Denoting $\|h\| = \|\mathbf{C}_{XX}^{-\beta} \boldsymbol{\mu}_{\mathbb{P}}\|$, we therefore have

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_{\mathbb{P}}\| \leq B\|\hat{\boldsymbol{\mu}}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}}\| + D\lambda^{\min\{\beta, \eta_0\}} \|\mathbf{C}_{XX}^{-\beta} \boldsymbol{\mu}_{\mathbb{P}}\| + C\|\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \|\mathbf{C}_{XX}^{-\beta} \boldsymbol{\mu}_{\mathbb{P}}\|. \quad \text{(C.2)}$$

For $0 \leq \beta \leq 1$, it follows from Theorem 1 in Bauer et al. (2007) that there exists a constant τ_1 such that

$$\|\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \leq \tau_1 \|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{op}^{\beta} \leq \tau_1 \|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{\text{HS}}^{\beta}.$$

On the other hand, since $\alpha \mapsto \alpha^\beta$ is Lipschitz on $[0, \kappa^2]$ for $\beta \geq 1$, the following lemma yields that

$$\|\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}\|_{op} \leq \|\mathbf{C}_{XX}^{\beta} - \widehat{\mathbf{C}}_{XX}^{\beta}\|_{\text{HS}} \leq \tau_2 \|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{\text{HS}}$$

where τ_2 is the Lipschitz constant of $\alpha \mapsto \alpha^\beta$ on $[0, \kappa^2]$. In other words,

$$\|\mathbf{C}_{XX}^\beta - \widehat{\mathbf{C}}_{XX}^\beta\|_{op} \leq \max\{\tau_1, \tau_2\} \|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{\text{HS}}^{\min\{1,\beta\}}. \tag{C.3}$$

Lemma C.1 (Contributed by Anreas Maurer, see Lemma 5 in Vito et al. (2012))**.** *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are self-adjoint Hilbert-Schmidt operators on a separable Hilbert space* $\mathscr{H}$ *with spectrum contained in the interval* $[a, b]$*, and let* $(\sigma_i)_{i\in I}$ *and* $(\tau_j)_{j\in J}$ *be the eigenvalues of* $\mathbf{A}$ *and* $\mathbf{B}$*, respectively. Given a function* $r : [a, b] \to \mathbb{R}$*, if there exists a finite constant* L *such that*

$$|r(\sigma_i) - r(\tau_j)| \leq L|\sigma_i - \tau_j|, \ \forall\, i \in I,\, j \in J,$$

then

$$\|r(\mathbf{A}) - r(\mathbf{B})\|_{HS} \leq L\|\mathbf{A} - \mathbf{B}\|_{HS}.$$

Using (C.3) in (C.2), we have

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\mathbb{P}\| \leq B\|\hat{\boldsymbol{\mu}}_\mathbb{P} - \boldsymbol{\mu}_\mathbb{P}\| + D\lambda^{\min\{\beta,\eta_0\}}\|\mathbf{C}_{XX}^{-\beta}\boldsymbol{\mu}_\mathbb{P}\| + C\tau\|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{\text{HS}}^{\min\{1,\beta\}}\|\mathbf{C}_{XX}^{-\beta}\boldsymbol{\mu}_\mathbb{P}\|, \tag{C.4}$$

where $\tau := \max\{\tau_1, \tau_2\}$. We now obtain bounds on $\|\hat{\boldsymbol{\mu}}_\mathbb{P} - \boldsymbol{\mu}_\mathbb{P}\|$ and $\|\mathbf{C}_{XX} - \widehat{\mathbf{C}}_{XX}\|_{\text{HS}}$ using the following results.

Lemma C.2 (Gretton et al. (2012a))**.** *Suppose that* $\kappa = \sup_{\mathbf{x}\in\mathcal{X}} \sqrt{k(\mathbf{x}, \mathbf{x})}$*. For any* $\delta > 0$*, the following inequality holds with probability at least* $1 - e^{-\delta}$

$$\|\hat{\boldsymbol{\mu}}_\mathbb{P} - \boldsymbol{\mu}_\mathbb{P}\| \leq \frac{2\kappa + \kappa\sqrt{2\delta}}{\sqrt{n}}.$$

Lemma C.3 (e.g., see Theorem 7 in Rosasco et al. (2010))**.** *Let* $\kappa := \sup_{\mathbf{x}\in\mathcal{X}} \sqrt{k(\mathbf{x}, \mathbf{x})}$*. For* $n \in \mathbb{N}$ *and any* $\delta > 0$*, the following inequality holds with probability at least* $1 - 2e^{-\delta}$*:*

$$\left\|\widehat{\mathbf{C}}_{XX} - \mathbf{C}_{XX}\right\|_{HS} \leq \frac{2\sqrt{2}\kappa^2\sqrt{\delta}}{\sqrt{n}}.$$

Using Lemmas C.2 and C.3 in (C.4), for any $\delta > 0$, with probability $1 - 3e^{-\delta}$, we obtain

$$\|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}_\mathbb{P}\| \leq \frac{2\kappa B + \kappa B\sqrt{2\delta}}{\sqrt{n}} + D\lambda^{\min\{\beta,\eta_0\}}\|\mathbf{C}_{XX}^{-\beta}\boldsymbol{\mu}_\mathbb{P}\| + C\tau\frac{(2\sqrt{2}\kappa^2\sqrt{\delta})^{\min\{1,\beta\}}}{n^{\min\{1/2,\beta/2\}}}\|\mathbf{C}_{XX}^{-\beta}\boldsymbol{\mu}_\mathbb{P}\|.$$

This completes the proof. ■

C.5 Proof of Theorem 5.4

Lemma C.4. *Given a set of distributions* $\mathcal{P} = \{\mathbb{P}_1, \mathbb{P}_2 \ldots, \mathbb{P}_N\}$*, the distributional variance of* $\mathcal{P}$ *is*

$$\mathbb{V}_\mathscr{H}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|_\mathscr{H}^2$$

where $\boldsymbol{\mu}_{\bar{\mathbb{P}}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\mu}_{\mathbb{P}_i}$ *and* $\bar{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^N \mathbb{P}_i$*.*

Proof. Let $\bar{\mathbb{P}}$ be the probability distribution defined as $\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i$, *i.e.*, $\bar{\mathbb{P}}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i(x)$. It follows from the linearity of the expectation that $\boldsymbol{\mu}_{\bar{\mathbb{P}}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_{\mathbb{P}_i}$. For brevity, we will denote $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ by $\langle\cdot,\cdot\rangle$. Then, expanding (5.11) gives

$$\begin{aligned}
\mathbb{V}_{\mathscr{H}}(\mathcal{P}) &= \frac{1}{N}\mathrm{tr}(\boldsymbol{\Sigma}) = \frac{1}{N}\mathrm{tr}(\mathbf{G}) - \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{G}_{ij} \\
&= \frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_i}\rangle - \frac{1}{N^2}\sum_{i,j=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_j}\rangle \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_i}\rangle - \frac{2}{N}\sum_{i,j=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_j}\rangle + \frac{1}{N}\sum_{i,j=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_j}\rangle\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_i}\rangle - 2\sum_{i=1}^{N}\left\langle\boldsymbol{\mu}_{\mathbb{P}_i},\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\mu}_{\mathbb{P}_j}\right\rangle + N\left\langle\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}_{\mathbb{P}_i},\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\mu}_{\mathbb{P}_j}\right\rangle\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_i}\rangle - 2\sum_{i=1}^{N}\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\bar{\mathbb{P}}}\rangle + N\langle\boldsymbol{\mu}_{\bar{\mathbb{P}}},\boldsymbol{\mu}_{\bar{\mathbb{P}}}\rangle\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big(\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_i}\rangle - 2\cdot\langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\bar{\mathbb{P}}}\rangle + \langle\boldsymbol{\mu}_{\bar{\mathbb{P}}},\boldsymbol{\mu}_{\bar{\mathbb{P}}}\rangle\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}} \ ,
\end{aligned}$$

which completes the proof. ■

Next, I give a proof of Theorem 5.4.

Proof of Theorem 5.4. Since k is characteristic, $\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|^2_{\mathscr{H}}$ is a metric and is zero iff $\mathbb{P} = \mathbb{Q}$ for any distributions $\mathbb{P}$ and $\mathbb{Q}$ (Sriperumbudur et al. 2010). By Lemma C.4, $\mathbb{V}_{\mathscr{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}}$. Thus, $\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}} = 0$ iff $\mathbb{P}_i = \bar{\mathbb{P}}$. Consequently, if $\mathbb{V}_{\mathscr{H}}(\mathcal{P})$ is zero, this implies that $\mathbb{P}_i = \bar{\mathbb{P}}$ for all i, meaning that $\mathbb{P}_1 = \cdots = \mathbb{P}_\ell$. Conversely, if $\mathbb{P}_1 = \cdots = \mathbb{P}_\ell$, then $\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}} = 0$ is zero for all i and thereby $\mathbb{V}_{\mathscr{H}}(\mathcal{P}) = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{\mu}_{\mathbb{P}_i} - \boldsymbol{\mu}_{\bar{\mathbb{P}}}\|^2_{\mathscr{H}}$ is zero. ■

C.6 Proof of Theorem 5.5

Proof. Recall that

$$\mathbb{V}_{\mathscr{H}}(\mathcal{P}) = \frac{1}{N}\mathrm{tr}(\mathbf{G}) - \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{G}_{ij} \quad \text{and} \quad \widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{S}) = \frac{1}{N}\mathrm{tr}(\widehat{\mathbf{G}}) - \frac{1}{N^2}\sum_{i,j=1}^{N}\widehat{\mathbf{G}}_{ij}$$

where

$$\begin{aligned}
\mathbf{G}_{ij} &= \langle\boldsymbol{\mu}_{\mathbb{P}_i},\boldsymbol{\mu}_{\mathbb{P}_j}\rangle_{\mathscr{H}} = \iint k(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbb{P}_i(\mathbf{x})\,\mathrm{d}\mathbb{P}_j(\mathbf{z}) \\
\widehat{\mathbf{G}}_{ij} &= \langle\hat{\boldsymbol{\mu}}_{\mathbb{P}_i},\hat{\boldsymbol{\mu}}_{\mathbb{P}_j}\rangle_{\mathscr{H}} = \frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}k(\mathbf{x}_k^{(i)},\mathbf{x}_l^{(j)})
\end{aligned}$$

By Theorem 15 in Altun and Smola (2006), we have a fast convergence of $\hat{\boldsymbol{\mu}}_{\mathbb{P}}$ to $\boldsymbol{\mu}_{\mathbb{P}}$. Consequently, we have $\widehat{\mathbf{G}} \to \mathbf{G}$, which implies that $\widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{S}) \to \mathbb{V}_{\mathscr{H}}(\mathcal{P})$. Hence, $\widehat{\mathbb{V}}_{\mathscr{H}}(\mathcal{S})$ is a consistent estimator of $\mathbb{V}_{\mathscr{H}}(\mathcal{P})$. ■

C.7 Proof of Theorem 5.8

We consider a scenario where distributions $\mathbb{P}^i$ are drawn according to $\mathscr{P}^*$ with probability μ_i. Introduce shorthand $\tilde{X}_{ij}$ for $(\mathbb{P}^{(i)}, X_{ij})$ for a distribution on $\mathscr{P}_{\mathcal{X}}$ and a corresponding random variable on $\mathcal{X}$.

The quantity of interest is the difference between the expected and empirical loss of a classifier $f : \mathscr{P}_{\mathcal{X}} \times \mathcal{X} \to \mathcal{Y}$ under loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$.

Assumptions. The loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is ϕ_ℓ-Lipschitz in its first variable and bounded by U_ℓ. The kernel $k_{\mathcal{X}}$ is bounded by $U_{\mathcal{X}}$. Assume that all distributions in $\mathscr{P}^*$ are mapped into a ball of size $U_{\mathscr{P}}$ by $\Psi_{\mathscr{P}}$. Finally, since $k_{\mathscr{P}}$ is a is a square exponential, there is a constant $L_{\mathscr{P}}$ such that

$$\|\Phi_{\mathscr{P}}(v) - \Phi_{\mathscr{P}}(w)\| \leq L_{\mathscr{P}}\|v - w\| \text{ for all } v, w.$$

Recall that N is the number of sampled domains, n_i is the number of samples in domain i, and $n = \sum_{i=1}^N n_i$ is the total number of samples. The proof assumes $n_i = n_j$ for all i, j.

Lemma C.5. *Recall that $\Phi_{\mathbf{x}} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]$. The composition $\mathbf{x}_t \mapsto \mathbf{k}_t \cdot \mathbf{B}$, where $\mathbf{k}_t = [k(\mathbf{x}_1, \mathbf{x}_t), \ldots, k(\mathbf{x}_n, \mathbf{x}_t)]$, can therefore be rewritten as $\phi(\mathbf{x}_t) \cdot \mathcal{B} = \phi(\mathbf{x}_t) \cdot \Phi_{\mathbf{x}} \cdot \mathbf{B}$.*

Proof. The proof modifies the approach taken in Blanchard et al. (2011b) to handle the preprocessing via transform $\mathcal{B}$, and the fact that we work with *squared* errors. Parts of the proof that pass through largely unchanged are omitted.

We repeatedly apply the inequality $|a + b|^2 \leq 2|a|^2 + 2|b|^2$. However, we only incur the multiplication-by-2 penalty once since $|a_1 + \cdots + a_n|^2 \leq 2|a_1|^2 + \cdots + 2|a_n|^2$.

Decompose

$$\begin{aligned}
&\sup_{\|f\|_{\mathscr{H}} \leq 1} \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&\leq \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^N \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&+ \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^N \left| \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\widehat{\mathbb{P}}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&+ \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^N \left| \mathbb{E}_{\widehat{\mathbb{P}}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&= (A) + (B) + (C) \ .
\end{aligned}$$

Control of (C):

$$\begin{aligned}
(C) &= \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^N \left| \mathbb{E}_{\widehat{\mathbb{P}}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\hat{\mathbb{P}}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&\leq \phi_\ell^2 \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^N \left| \mathbb{E}_{\widehat{\mathbb{P}}^i} f(\tilde{X}_{ij}\mathcal{B}) - \mathbb{E}_{\hat{\mathbb{P}}} f(\tilde{X}_{ij}\mathcal{B}) \right|^2
\end{aligned}$$

$$= \phi_\ell^2 \cdot \frac{2}{N} \sum_{i=1}^{N} \left\| \Psi_{\mathscr{P}}(\widehat{\mathbb{P}}^i) \otimes \mu_{\widehat{\mathbb{P}}^i}\mathcal{B} - \Psi_{\mathscr{P}}(\widehat{\mathbb{P}}) \otimes \mu_{\widehat{\mathbb{P}}}\mathcal{B} \right\|^2$$

Note that $\|\Psi_{\mathscr{P}}(\mu(\mathbb{P}))\|^2 \leq L_{\mathscr{P}} \cdot \|\mu_{\mathbb{P}}\|^2 \leq L_{\mathscr{P}} U_{\mathscr{P}}$. Therefore,

$$(C) \leq \phi_\ell^2 L_{\mathscr{P}} U_{\mathscr{P}} \frac{2}{N} \sum_{i=1}^{N} \left\| \mu_{\widehat{\mathbb{P}}^i}\mathcal{B} - \mu_{\widehat{\mathbb{P}}}\mathcal{B} \right\|^2 .$$

By the proof of Theorem 5.4 and since $\Phi_x^\top \mathcal{B} = \mathbf{KB}$, we have

$$(C) \leq 2\phi_\ell^2 L_{\mathscr{P}} U_{\mathscr{P}} \frac{1}{N} \mathrm{tr}(\mathbf{KBB^\intercal KL}).$$

Control of (B): Similarly,

$$\begin{aligned}
(B) &= \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^{N} \left| \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\widehat{\mathbb{P}}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2 \\
&\leq 2\phi_\ell^2 L_{\mathscr{P}} U_{\mathscr{P}} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| \mu_{\mathbb{P}^i}\mathcal{B} - \mu_{\widehat{\mathbb{P}}^i}\mathcal{B} \right\|^2 \\
&\leq 2\phi_\ell^2 L_{\mathscr{P}} U_{\mathscr{P}} \cdot \|\mathcal{B}\|_{\mathrm{HS}}^2 \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| \mu_{\mathbb{P}^i} - \mu_{\widehat{\mathbb{P}}^i} \right\|^2
\end{aligned}$$

Here we follow the strategy applied by Blanchard et al. (2011b) to control their term (I) in Theorem 5.1. Assume $n_i = n_j$ for all i, j and recall $n = \sum_{i=1}^{N} n_i$ so $n_i = n/N$ for all i.

By Hoeffding's inequality in Hilbert space, with probability greater than $1 - \delta$ the following inequality holds

$$\left\| \frac{1}{n_i} \sum_{j=1}^{n_i} \mu(\hat{X}_{ij}) - \mathbb{E}_{\mathbb{P}^{(i)}} \mu(X_{ij}) \right\|^2 \leq 9U_{\mathcal{X}} \frac{N \cdot \log 2\delta^{-1}}{n}.$$

Applying the union bound obtains

$$(Ib) \leq 18\phi_\ell^2 L_{\mathscr{P}} U_{\mathscr{P}} U_{\mathcal{X}} \cdot \|\mathcal{B}\|_{\mathrm{HS}}^2 \cdot \frac{N \cdot (\log \delta^{-1} + 2 \log N)}{n}.$$

Control of (A):

$$(A) = \sup_{\|f\|_{\mathscr{H}} \leq 1} \frac{2}{N} \sum_{i=1}^{N} \left| \mathbb{E}^*_{\mathscr{P}} \mathbb{E}_{\mathbb{P}} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) - \mathbb{E}_{\mathbb{P}^i} \ell(f(\tilde{X}_{ij}\mathcal{B}), Y_i) \right|^2$$

Following the strategy used by Blanchard et al. (2011b) to control (II) in Theorem 5.1, we obtain

$$(A) \leq c_3 \frac{\phi_\ell^2 U_{\mathcal{X}}^2 U_{\mathscr{P}} + U_\ell \log \delta^{-1}}{N} \cdot \|\mathcal{B}\|_{\mathrm{HS}}^2.$$

End of proof: We have that $\mathbf{K}$ is invertible since $\widehat{\mathbf{C}}_{XX}$ is assumed to be invertible. It follows that the trace $\mathrm{tr}(\mathbf{B}^\top \mathbf{KB})$ defines a norm which coincides with the Hilbert-Schmidt norm $\|\mathcal{B}\|_{\mathrm{HS}}^2$. Combining the three inequalities above concludes the proof. ■

C.8 Derivation of Equation (5.16)

DICA employs the covariance of inverse regressor $\mathbb{V}(\mathbb{E}[\phi(X)|Y])$, which can be written in terms of covariance operators. Let $\mathscr{H}$ and $\mathscr{F}$ be the RKHSes of X and Y endowed with reproducing kernels k and l, respectively. Let $\mathbf{C}_{XX}$, $\mathbf{C}_{YY}$, $\mathbf{C}_{XY}$, and $\mathbf{C}_{YX}$ be the covariance operators in and between the corresponding RKHSes of X and Y. We define the conditional covariance operator of X given Y, denoted by $\mathbf{\Sigma}_{\mathrm{xx|y}}$, as

$$\mathbf{\Sigma}_{\mathrm{xx|y}} \triangleq \mathbf{C}_{XX} - \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX} \ . \tag{C.5}$$

The following theorem from Fukumizu et al. (2004) states that, under mild conditions, $\mathbf{\Sigma}_{\mathrm{xx|y}}$ equals the expected conditional variance of $\phi(X)$ given Y.

Theorem C.6. *For any $f \in \mathscr{H}$, if there exists $g \in \mathscr{F}$ such that $\mathbb{E}[f(X)|Y] = g(Y)$ for almost every Y, then $\mathbf{\Sigma}_{\mathrm{xx|y}} = \mathbb{E}[\mathbb{V}(\phi(X)|Y)]$.*

Using the E-V-V-E identity[1], the covariance $\mathbb{V}(\mathbb{E}[\phi(X)|Y])$ can be expressed in terms of the conditional covariance operators as follow:

$$\mathbb{V}(\mathbb{E}[\phi(X)|Y]) = \mathbb{V}(\phi(X)) - \mathbb{E}[\mathbb{V}(\phi(X)|Y)], \tag{C.6}$$

assuming that the inverse regressor $\mathbb{E}[f(x)|y]$ is a smooth function of y for any $f \in \mathscr{H}$.

By virtue of Theorem C.6, the second term in the r.h.s. of (C.6) is $\mathbf{\Sigma}_{\mathrm{xx|y}}$. Since $\mathbb{V}(\phi(X)) = \mathrm{Cov}(\phi(x), \phi(x)) = \mathbf{C}_{XX}$, it follows from (C.5) that the covariance of the inverse regression $\mathbb{V}(\mathbb{E}[\phi(X)]|Y)$ can be expressed as

$$\mathbb{V}(\mathbb{E}[\phi(X)|Y]) = \mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX} \ . \tag{C.7}$$

The covariance (C.7) can be estimated from finite samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ by $\widehat{\mathbb{V}}(\mathbb{E}[\phi(X)|Y]) = \widehat{\mathbf{C}}_{XY}\widehat{\mathbf{C}}_{YY}^{-1}\widehat{\mathbf{C}}_{YX}$ where $\widehat{\mathbf{C}}_{XY} = \frac{1}{n}\Phi_x\Phi_y^\top$ and $\Phi_x = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)]$ and $\Phi_y = [\varphi(\mathbf{y}_1), \ldots, \varphi(\mathbf{y}_n)]$. Let $\mathbf{K}$ and $\mathbf{L}$ denote the kernel matrices computed over samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, respectively. We have

$$\begin{aligned}
\widehat{\mathbb{V}}(\mathbb{E}[\phi(X)|Y]) &= \left(\frac{1}{n}\Phi_x\Phi_y^\top\right)\left(\frac{1}{n}(\Phi_y\Phi_y^\top + n\varepsilon\mathcal{I})\right)^{-1}\left(\frac{1}{n}\Phi_y\Phi_x^\top\right) \\
&= \frac{1}{n}\Phi_x\Phi_y^\top\Phi_y\left(\Phi_y^\top\Phi_y + n\varepsilon\mathbf{I}_n\right)^{-1}\Phi_x^\top \\
&= \frac{1}{n}\Phi_x\mathbf{L}\left(\mathbf{L} + n\varepsilon\mathbf{I}_n\right)^{-1}\Phi_x^\top
\end{aligned} \tag{C.8}$$

where $\mathbf{L} = \Phi_y^\top\Phi_y$ and $\mathcal{I}$ is the identity operator. The second equation is obtained by applying the fact that $(\Phi_y\Phi_y^\top + n\varepsilon\mathcal{I})\Phi_y = \Phi_y(\Phi_y^\top\Phi_y + n\varepsilon\mathbf{I}_n)$.

Finally, using $\widehat{\mathbf{C}}_{XX} = \frac{1}{n}\Phi_x\Phi_x^\top$ and recalling that $\mathbf{K} = \Phi_x^\top\Phi_x$, we obtain

$$\begin{aligned}
\boldsymbol{\beta}_k^\top\widehat{\mathbf{C}}_{XX}^{-1}\widehat{\mathbb{V}}(\mathbb{E}[X|Y])\widehat{\mathbf{C}}_{XX}\boldsymbol{\beta}_k &= \boldsymbol{\beta}_k^\top\left(\frac{1}{n}\Phi_x\Phi_x^\top\right)^{-1}\left(\frac{1}{n}\Phi_x\mathbf{L}\left(\mathbf{L} + n\varepsilon\mathbf{I}_n\right)^{-1}\Phi_x^\top\right)\left(\frac{1}{n}\Phi_x\Phi_x^\top\right)\boldsymbol{\beta}_k \\
&= \frac{1}{n}\boldsymbol{\beta}_k^\top\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)^{-1}\Phi_x\mathbf{L}\left(\mathbf{L} + n\varepsilon\mathbf{I}_n\right)^{-1}\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)\Phi_x\boldsymbol{\beta}_k \\
&= \frac{1}{n}\boldsymbol{\beta}_k^\top\Phi_x^\top\Phi_x\left(\Phi_x^\top\Phi_x\right)^{-1}\mathbf{L}\left(\mathbf{L} + n\varepsilon\mathbf{I}_n\right)^{-1}\Phi_x^\top\left(\Phi_x\Phi_x^\top\right)\Phi_x\boldsymbol{\beta}_k
\end{aligned}$$

[1] $\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y])$ for any X, Y.

$$= \frac{1}{n}\boldsymbol{\beta}_k^\top \mathbf{L}(\mathbf{L} + n\varepsilon\mathbf{I})^{-1}\mathbf{K}^2\boldsymbol{\beta}_k$$

and

$$\boldsymbol{\beta}_k^\top \boldsymbol{\beta}_k = \boldsymbol{\beta}_k^\top \Phi_x^\top \Phi_x \boldsymbol{\beta}_k = \boldsymbol{\beta}_k^\top \mathbf{K}\boldsymbol{\beta}_k$$

as desired.

C.9 Derivation of Lagrangian (5.18)

Observe that optimization

$$\max_{\mathbf{B}\in\mathbb{R}^{n\times m}} \frac{\operatorname{tr}\left(\mathbf{B}^\top\mathbf{X}\mathbf{B}\right)}{\operatorname{tr}\left(\mathbf{B}^\top\mathbf{Y}\mathbf{B}\right)} \tag{C.9}$$

is invariant to rescaling $\mathbf{B} \mapsto \alpha \cdot \mathbf{B}$. Optimization (C.9) is therefore equivalent to

$$\begin{aligned} &\max_{\mathbf{B}\in\mathbb{R}^{n\times m}} \operatorname{tr}\left(\mathbf{B}^\top\mathbf{X}\mathbf{B}\right) \\ &\text{subject to: } \operatorname{tr}\left(\mathbf{B}^\top\mathbf{Y}\mathbf{B}\right) = 1, \end{aligned}$$

which yields Lagrangian

$$\mathcal{L} = \operatorname{tr}\left(\mathbf{B}^\top\mathbf{X}\mathbf{B}\right) - \operatorname{tr}\left(\left(\mathbf{B}^\top\mathbf{Y}\mathbf{B} - \mathbf{I}\right)\Gamma\right). \tag{C.10}$$

END OF APPENDIX