8 Categorization Based on Similarity and Features: The Reproducing Kernel Banach Space (RKBS) Approach

Jun Zhang and Haizhang Zhang

8.1 Introduction

Categorization refers to the process by which discriminably different things are judged as belonging to groups that are treated as being equivalent in some regard. The ability to categorize or classify objects, images, events, actions, etc., is a hallmark of adaptive intelligence. It allows animals and humans to robustly extract regularity and patterns in environmental inputs without much prior knowledge or assumptions about their structures and interrelations, thereby reducing the complexity of mental computation to a manageable degree. Categorization arises as a result of a flexible search for structure of things in the environment which consists of patterns of correlated features forming natural chunks (clusters). Category learning is the process of inferring the structure of categories from a set of

stimuli labeled as belonging to those categories; the knowledge acquired through this process can ultimately be used to make decisions about how to categorize new stimuli.

To an adaptive agent, categorization (as it is called in cognitive psychology) or classification and clustering (as it is called in machine learning) provides an internal model of the environment, with dimensionality of the "state space" greatly reduced compared with how inputs are specified; this solves the bottleneck problem facing reinforcement learning, which deals with the agent's optimal action planning and sequential decision-making. Solution of the categorization problem, in combination with the now well-understood solution to the reinforcement learning, will pave the way for designing artificial intelligence with autonomous and adaptive planning capabilities.

In this chapter, we review a theoretical framework of categorization that is both strongly motivated by cognitive psychology and rigorously grounded in mathematics. This is the framework of regularized learning under supervision using finite samples. We provide a unified treatment of *2-norm based reproducing kernel method for regularized learning* with *1-norm based sparsity method for feature selection*, built upon Banach space formulation of the learning problem. The mathematical tool for our exposition is the Reproducing Kernel Banach Space (RKBS) – an RKBS is equipped with an semi-imperfect discrimination operation, which specializes to the imperfect discrimination operation for the case of Hilbert spaces. The RKBS framework provides a potentially unifying framework for the analysis of similarity and feature representation in categorization.

8.1.1 Usefulness of Categories

Categories serve two primary functions (see Smith, 1995): (1) they enable efficient storage and retrieval of information, and relieve us from the burden of keeping track of every individual item we encounter; (2) they promote inferences and extend our knowledge beyond past experiences into the future, allowing us to make predictions that guide behavior. Categories serve not only to organize the knowledge we have already acquired, but also to guide our expectations through inductive reasoning. Induction is the mental capacity to extend knowledge to novel instances, for example, inferring that a newly encountered mushroom is poisonous on the basis of past encounters with other poisonous mushrooms. Inductive inference is one of the most important utilities that come with categorization.

The ability to categorize arises very early ontogenetically. Young children are able to use categories to support inductive inferences even when category membership of objects conflicts with their appearances, whereas category labels do not generally affect children's perceptual similarity judgments (Noles and Gelman, 2012). Young children's category-based inferences attained the status of what Gelman dubbed as "psychological essentialism" – children readily infer properties that concern internal features and non-visible functions from one category member to

another, and their inferences rely more on category membership information than perceptual similarity information (Gelman, 2004).

There has accumulated a large body of empirical literature on human categorization. Rosch's (e.g., Rosch, 1973; Rosch & Mervis, 1975) seminal studies of natural object categories led to the commonly held view that people learn about and use hierarchically organized categories, with entry-level categories possessing meaningful sub-structures with correlated features and super-structures that they contribute towards. Feature and category label are two sides of the coin of a hierarchical scheme – the fundamentally probabilistic nature of the environment leads to the creation of category labels for objects with overlapping features based on some similarity function defined on some feature space. However, perceptual similarity alone does not account for differences in categories, although our perceptual system clearly has evolved to facilitate categorization. As concepts and categories serve (sometimes multiple) functions and goals, the categorization system must be flexible enough to both assimilate the necessary underlying structure and discover or even create that structure.

The importance of solving categorization puzzle can be appreciated in connection with reinforcement learning. Whereas reinforcement learning (both model-based and model-free variants) enables optimal planning and sequential decision-making for an agent, category induction from finite samples will construct a compact representation of environmental state space. Categorization is both a hallmark of adaptive intelligence and the bottleneck problem affecting wider applications of the otherwise successful reinforcement learning algorithms.

8.1.2 Extant Psychological Models and Issues

There is a long tradition of empirical study and theoretical construction of mathematical models of human categorization by cognitive psychologists. Existing psychological models of categorization deal both with the process and the mechanism of various aspects of categorization. On the *normative* side, there is the rational theory of categorization (Anderson, 1990, 1991) which essentially treats category learning as non-parametric Bayesian inference about unknown probability distributions. This approach is followed by the SUSTAIN model (Love, Medin, & Gureckis, 2004; Vanpaemel & Storms, 2008), treating categories as mixture distributions over clusters, and becomes fully developed by Griffiths and associates (Austerweil and Griffiths, 2011, 2013; Sanborn, Griffiths, & Navarro, 2010) using elaborative statistical tools about sampling. On the *representational* side, there are basically two camps advocating opposing views about memory representations that support categorization. The Prototype Model (Reed, 1972; Smith & Minda, 1998) starts with the assumption that humans supposedly "average" their experience of encountering various exemplars of a category to create a prototype that is most typical for each category. This common-sense model was challenged by the class of Exemplar Model (Medin & Schaffer, 1978; Nosofsky, 1986), which claims that

humans store many exemplars of the same category in memory and those exemplars are simultaneously retrieved and compared against upon encountering a new input to be categorized. The debate between Prototype Model and Exemplar Model is how categories are represented in the memory system, whether as summary statistics (with mean and covariance structure) or as individual instances. On the *mechanistic* side, there are also different conceptualizations of how human category decisions are arrived at. The Decision-bound Model (Ashby & Gott, 1988; Ashby & Perrin, 1988) argues that humans represent stimuli in a multi-dimensional vector space that can be partitioned into various regions representing equivalent classes (with distinct category labels). Furthermore, the boundary of these partitioned regions can be explicitly learnt to form "decision boundaries." The Connectionist/Neural Network Model (Gluck & Bower, 1988; Kruschke, 1992), on the other hand, argues that input stimuli are represented in a parallel-distributed fashion that cannot be simply conceptualized as a vector space, and categorization is the result of non-linear network computation where non-linearity is crucial for both its representational power and classification accuracy. So, while the debate between prototype/exemplar models reflects contrasting theoretical assumptions about the representation and memory requirements for category structures, the debate between decision-bound/distributive network models reflects contrasting theoretical commitments about the architecture and implementation styles for the process and dynamics of category decisions.

The set of common issues to be addressed by computational models of human categorization include:

(i) The role of exemplar versus prototype in categorization: How is category structure maintained in memory? Are all exemplars equally important, or does typical exemplar (prototype) play a special role such as summary statistics?
(ii) The relationship between feature and category label: How is feature representation created? Does category label serve as just another feature?
(iii) The architecture for category decision: Are there clear boundaries in some multi-dimensional feature space that delineate one category from another? Do category decisions arise from connectionist/neural network style non-linear computation reflecting emerging properties that may not be mathematically trackable? How do rule-based versus boundary-based computation schemes reconcile with distributive representation?
(iv) The importance of similarity and generalization during categorization: Is the Sheperd's Universal Law of Generalization an adequate expression of similarity? Is similarity computation feature-based or exemplar-based?

A synthesis of the above-mentioned psychological models, all receiving various degrees of empirical support, would not only provide a satisfactory account of human categorization performance but greatly aid the search of a universal algorithm of learning-from-sample problem in machine learning. To do so, one needs to combine the strengths of exemplar-based versus feature-based approaches, and

integrate bound-based versus cluster-based approaches to explain how mechanisms (e.g., attention, similarity) at one level translate to another.

8.1.3 An Emerging Unifying Framework

Over the past decades, the field of statistical machine learning has produced a suite of power tools to tackle the problem of supervised and semi-supervised classification based on finite samples. The mathematical framework is the so-called "regularized loss minimization," which formulates the goal of classification as building an optimal model (i.e., input–output mapping) capable of generalizing beyond given sample data. The general philosophy is to treat categorization as an "ill-defined" or underconstrained inverse problem – there can be multiple schemes of category structures ("output") that can account for sample data ("input"), so there can be no unique solution of induced classification scheme that is compatible with finite sample data. Instead, one looks for an "optimal" solution, with optimality phrased in terms of the goals of categorization. A key insight in the computation approach to categorization is that the input–output mapping should be evaluated based on its performance not only on known samples but also on unseen data. Technically, when misclassification is quantified as "loss," the objective for the learning algorithm is to minimize such loss by choosing a classifier with a modest amount of complexity ("regularized") while generalizing well from seen to unseen samples.

There are two main approaches in regularized learning, however, with entirely different emphasis and mathematical underpinning: one based on *reproducing kernel methods* and the other based on *sparsity methods*. Although both adopt the same principle of regularization, i.e., balancing the competing needs for goodness-of-fit and for simplicity of the optimal classifier, they differ with respect to generalization mechanism and feature representation. On the one hand, the paradigm of kernel methods assumes that classifiers live in a function space that is a Hilbert space with a reproducing kernel. The reproducing kernel, or kernel in short, guarantees the existence of (what can be an extremely non-linear) mapping from essentially arbitrary inputs to their feature representation, in service of, say, hyperplane-based classification (e.g., maximal margin methods). Although representation of features as a high-dimensional vector space is conceptually invoked, such a paradigm does not rely on explicit construction of the feature space. Instead, a kernel function is invoked to measure similarity of objects in the input space, such that their feature representation is often "bypassed" – this is known in the research community as the "kernel trick." On the other hand, the developments in Lasso in statistics (Tibshirani, 1996) and compressive sensing in signal processing (Candès, Romberg, & Tao, 2006), emphasize the importance of sparse representation for coding and classification using properly selected bases. As is now well-understood, it is the ℓ^1-norm (rather than the L^2-norm of a Hilbert space) that enforces the sparsity solution (as surrogate to the computationally intractable ℓ^0 regularization problem) and enables optimal feature selection. However, when

regularization learning is formulated on the Banach space of functions with L^1-norm (as opposed to the sequence space ℓ^1 where bases are already chosen), no representer-type theorem is known.

The connection and interaction between similarity-based generalization (through L^2 regularization) and the sparsity-based feature selection (through ℓ^1 regularization), both of which are reasonably well-understood, lie at the crux of solving the categorization problem. From cognitive psychology, the literature is quite clear that both similarity-based generalization (captured by L^2-based kernel method) and feature-based attention (modeled as ℓ^1-based sparse feature selection) are at play during category learning. Although the relevance of the RKHS framework for exemplar and prototype or perceptron models has been discussed abstractly (Jäkel, Schölkopf, & Wichmann, 2008a,b, 2009), and its potential for neural representations of feature/object dichotomy has been suggested (Riesenhuber & Poggio, 1999; Smale *et al.*, 2010), the details have yet to be laid out.

Besides features and similarity, another important cognitive construct in categorization and category learning/induction is that of attention. Existing psychological theories of attention in learning, as reviewed below, can be interpreted as modifying the kernel by rescaling stimulus representations either in the input space (Nosofsky, 1986; Sutherland & Mackintosh, 1971) or in the feature space (Kruschke, 2001; Mackintosh, 1975). Kernel learning (Lanckriet *et al.*, 2004) investigates the relationship between various forms of representation learning, in the context of convex optimization of mixture kernels (Argyriou, Micchelli, & Pontil, 2005; Micchelli & Pontil, 2005a) or multiple tasks (Evgeniou, Micchelli, & Pontil, 2005). Attention acting on the feature space may induce a convex family of kernels, but attention acting on the input space generally does not. Therefore, the goal of optimizing the kernel through attention to input space (as in exemplar models of categorization) is of particular interest. There is a need for separate mathematical analyses of attention on exemplars versus attention on feature dimension.

8.1.4 Plan of Chapter

The plan of the rest of this chapter is as follows. In Section 8.2, we review fundamental mathematical concepts and tools of reproducing kernel methods. Starting from vector spaces, we introduce the notions of linear functional, norm, duality mapping, inner product and its generalization semi-inner product (Section 8.2.1). When considering function spaces as special examples of vector spaces, the idea of the reproducing kernel is then introduced to link a function with its evaluation (Section 8.2.2). In Section 8.3, the mathematical problem of learning with finite samples is formulated. This is done by first revisiting Statistical Learning Theory and the need for regularization (Section 8.3.1), and the various norms used as regularizers (Section 8.3.2). Section 8.4 then proceeds by reviewing solution concepts afforded with the reproducing kernel methods. They include the celebrated

"representer theorem" and the widely used "kernel trick" (Section 8.4.1), sampling methods (Section 8.4.2), and maximal margin classifier, feature selection, and vector-valued functions as feature maps (Section 8.4.3). Section 8.5 deals with psychological underpinnings of kernel methods. We first provide a comprehensive review of various psychological models of human categorization (Section 8.5.1), and then attempt to cast these models in a unified framework using the language of kernel methods (Section 8.5.2), with challenges also discussed (Section 8.5.3). Finally, our chapter closes with a short summary (Section 8.6).

8.2 Mathematical Preliminaries

In this section, we review the mathematical background underlying the reproducing kernel method that is popular in contemporary machine learning. We first review a basic but important mathematical object, namely, "vector space." Then, we study functions defined on vector spaces, known as "functionals," and the duality mapping between the space of continuous linear functionals and the original vector space; such duality mapping is linked to inner product or semi-inner product defined on a Hilbert or a Banach (vector) space, respectively. Next, we focus on function spaces as a particular kind of vector space, and the evaluation functional as a particular kind of linear functional. This leads to the notion of "reproducing kernels," which plays a central role in formulating the regularization problem in machine learning.

8.2.1 Vector Spaces and Functionals

A vector space V over the field $\mathbb{C}$ of complex numbers is a set equipped with the following two operations that are closed for all $u, v \in V$, $c \in \mathbb{C}$:

1. *vector addition*: $(u, v) \to u + v \in V$;
2. *scalar multiplication*: $(c, u) \to cu \in V$.

The vector addition operation is required to be commutative $u + v = v + u$, associative $w + (u + v) = (w + u) + v$, and to possess an identity $0 \in V$, namely $0 + u = u + 0 = u$ for all $u \in V$, such that every vector $u \in V$ has an additive inverse $-u \in V$, namely $u + (-u) = 0$. Essentially, these requirements about the vector addition operation $+$ make $(V, +)$ an Abelian group. The scalar multiplication operation, on the other hand, "couples" the Abelian group structure with the scalar field to provide richer structures of V. First, the scalar multiplication is associative and distributive with respect to vector addition. That is, there hold for all $u, v \in V$ and $a, b \in \mathbb{C}$ that

$$a(bu) = (ab)u,\ (a + b)u = au + bu,\ a(u + v) = au + av.$$

Second, the unity $1 \in \mathbb{C}$ in the field satisfies $1u = u$ for all $u \in V$. Note that a number field (in this case $\mathbb{C}$) must be associated with a vector space V – one

cannot speak of a vector space without specifying the associated field. However, the field can be any general number field, including the field $\mathbb{R}$ of real numbers.

A familiar example of vector space is $\mathbb{C}^n$ (or $\mathbb{R}^n$), where every vector is represented by an array of n complex (or real) numbers, addition is componentwise, and scalar multiplication is field multiplication on each of these numbers separately.

8.2.1.1 Linear Functionals and Norms

Over the vector space V, one can consider mappings from V to $\mathbb{C}$, that is, taking any vector in V as input to yield a number $c \in \mathbb{C}$ as output. Each such mapping, denoted as T, is called a *functional*, the totality of which we denote as $\mathbb{C}^V$.

An important property of the space $\mathbb{C}^V$ of functionals (with vectors in V as prescribed inputs) is that $\mathbb{C}^V$ itself is a vector space. In particular, linear combinations $c_1T_1 + c_2T_2$ of two functionals T_1, T_2 are a functional (where $T_1, T_2 \in \mathbb{C}^V$ and c_1, c_2 are scalars). Among all functionals, there is a particular subset called *linear functionals*, i.e., the map T is linear with respect to the *input* vectors u, v in V: $T(u + v) = T(u) + T(v)$. The set of all linear functionals forms a vector space, a subspace of $\mathbb{C}^V$.

Given a vector space V over $\mathbb{C}$, an important functional is the *norm* on V, usually denoted as $\|\cdot\|_V$ or simply $\|\cdot\|$. A norm is a non-linear functional which maps an element of V to a non-negative real value $\mathbb{R}_+$ such that

(i) *positivity*: $\|u\|_V \geq 0$ for all $u \in V$ and $\|u\|_V = 0$ if and only if $u = 0$,
(ii) *homogeneity*: $\|cu\|_V = |c|\|u\|_V$ for all $u \in V$ and $c \in \mathbb{C}$,
(iii) *triangle inequality*: $\|u + v\|_V \leq \|u\|_V + \|v\|_V$ for all $u, v \in V$.

A vector space endowed with a norm is called a *normed vector space*.

Any norm $\|\cdot\|_V$ on a vector space V induces a metric $d(\cdot, \cdot)$ on V by

$$d(u, v) = \|u - v\|_V, \quad u, v \in V.$$

It can be verified, by the three defining properties of a norm, that $d(u, v)$ is non-negative, symmetric, and satisfies the triangle inequality

$$d(u, v) \leq d(u, w) + d(w, v).$$

Thus, $d(u, v)$ qualifies as a metric and thus can induce a topology on V. Under this norm-induced topology, the convergence of a sequence to some point is equivalent to their metric distance vanishing in the limit.

For a normed vector space V, we can speak of continuity of any linear functional T mapping V to $\mathbb{C}$: continuity is with respect to the norm-induced topology on V. The set of all continuous linear functionals on V is denoted as V^*; it is a vector subspace of $\mathbb{C}^V$, and is called the "dual space" or simply the "dual" of V. It can be equipped with the so-called "dual norm" (induced from the norm $\|\cdot\|_V$ on V):

$$\|T\|_{V^*} := \sup_{u \in V, u \neq 0} \frac{|T(u)|}{\|u\|_V}.$$

(One can prove that $\|T\|_{V^*}$ as defined indeed satisfies the three axioms of a norm.)

When $\|T\|_{V^*} \leq C$ where C is some positive real number, then we say that T is a bounded functional. When $\|T\|_{V^*} < +\infty$, then T is continuous. Equivalently, $\lim_{n\to\infty} T(u_n) = 0$ whenever u_n converges to $0 \in V$. The Hahn–Banach theorem, a fundamental result in functional analysis, states that every continuous linear functional on a subspace of the input vector space V can be extended in a norm-preserving way to the whole space V. Hence, for each $u \in V$ there exists a continuous linear functional $T \in \mathbb{C}^V$ such that

1. $|T(u)| = \|T\|_{V^*}\|u\|_V$;
2. $\|T\|_{V^*} = \|u\|_V$.

We call any linear functional satisfying (i) a "dual element" or simply "dual" *of* u (or *with respect to* u) and usually denote it as u^*. When u maps to u^*, cu maps to $\bar{c}u^*$ for any $c \in \mathbb{C}$. This is to say, vectors in the same direction in V map to identical direction in the dual space V^*. When (ii) is further satisfied, u^* is called the "canonical dual" of u. In general, one can require instead of (ii): $\|T\|_{V^*} = \gamma(\|u\|_V)$ for a strictly monotone function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ that satisfies $\gamma(0) = 0$; $\lim_{t\to\infty} \gamma(t) = \infty$; the function γ is called a "gauge function".

The canonical dual $u^* \in V^*$ of $u \in V$ is unique when the norm $\|\cdot\|_V$ is uniformly convex. When the canonical dual of each $u \in V$ is chosen in the dual space V^*, we call the mapping $u \to u^*$ the *duality mapping*. This mapping is norm-preserving but is generally non-linear. The pairing of two vector spaces V and V^* is to specify a particular $T = u^*$ such that $T(u) = u^*(u) = \gamma(\|u\|_V)\|u\|_V$. The duality mapping may be linear when the gauge function $\gamma(t) = ct$ is linear and when the norm $\|\cdot\|_V$ is a special kind – this is the case of Hilbert space to be discussed next.

8.2.1.2 Inner Products and Hilbert Spaces

Hilbert spaces are a special case of normed vector spaces that are well-understood and widely used in physics and engineering applications. A Hilbert space is characterized by the existence of an inner product operation on a vector space. The inner product can uniquely induce a norm, a Hilbert 2-norm; conversely, a norm satisfying a certain equality can uniquely induce an inner product. Equivalently, a Hilbert space is a vector space that is self-dual, with linear duality mapping. The relationship between the existence of an inner product and self-dual property is reviewed next.

An inner product on a vector space V, denoted by $\langle\cdot,\cdot\rangle_V$ or simply $\langle\cdot,\cdot\rangle$, is by definition a function mapping $V \times V$ to $\mathbb{C}$ such that for all $u, v, w \in V$ and $c \in \mathbb{C}$

(i) *positivity*: $\langle u, u\rangle \geq 0$ for all $u \in V$ and $\langle u, u\rangle = 0$ if and only if $u = 0$;
(ii) *linearity about the first variable*: $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$ and $\langle cu, v\rangle = c\langle u, v\rangle$;
(iii) *conjugate symmetry*: $\langle u, v\rangle = \overline{\langle v, u\rangle}$.

It can be shown that (ii) and (iii) lead to linearity about the second variable in $\langle u, v\rangle$.

An inner product $\langle \cdot, \cdot \rangle$ on a vector space V induces a norm on V by

$$\|u\| := \langle u, u \rangle^{1/2}, \ \ u \in V.$$

(One can prove that the induced quantity is indeed a norm, that is, it satisfies the three axioms for norms listed in Section (8.2.1.1.) Furthermore, this norm *as induced from the inner product* satisfies the "Parallelogram Law" (also known as "polarization identity")

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2), \ \ u, v \in V.$$

Conversely, a norm on V induces an inner product:

$$\langle u, v \rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2)$$

when the norm $\| \cdot \|$ satisfies the Parallelogram Law. A vector space equipped with an inner product is called an *inner product space*. An inner product space whose induced norm is complete is called a Hilbert space.

On an inner product space V over $\mathbb{R}$, the angle between two non-zero vectors u, v can be measured by

$$\arccos \frac{\langle u, v \rangle}{\|u\| \|v\|}.$$

Thus, two vectors $u, v \in V$ are said to be orthogonal to each other if $\langle u, v \rangle = 0$. This leads to the notion of "orthogonal projection" of a vector u onto a closed subspace V_0 – it is the vector $v_0 \in V_0$ such that their difference $u - v_0$ is orthogonal to each vector in V_0:

$$\langle u - v_0, v \rangle = 0, \ \forall v \in V_0.$$

The orthogonal projection v_0 of u coincides with the best approximation of u in V_0, that is,

$$\|u - v_0\| = \inf_{v \in V_0} \|u - v\|.$$

Note that, fixing one slot of $\langle \cdot, \cdot \rangle$ turns the inner product into a linear functional $T(\cdot) = \langle \cdot, u_0 \rangle$ upon fixing u_0 – a linear functional can be viewed as an u_0-indexed inner product for some vector $u_0 \in V$. Furthermore, the dependency on u_0 is linear, due to linearity with respect to the second variable in an inner product – there is a linear (duality) map between $T \in V^*$ and $u_0 \in V$. It turns out that every continuous linear functional arises this way, as guaranteed by the fundamental theorem in functional analysis, namely the Riesz representation theorem:

Theorem 8.1 (Riesz representation theorem for Hilbert spaces) *Every continuous linear functional v^* on a Hilbert space V is representable as a v-indexed inner product $v^*(\cdot) = \langle \cdot, v \rangle$. In other words, every element in the dual space V^* of linear functionals on V can be represented by an element of the original vector space V.*

To summarize, a Hilbert space can be equivalently defined as a normed vector space

(i) equipped with an inner product; or
(ii) with norm-satisfying Parallelogram Law; or
(iii) with duality mapping being linear.

8.2.1.3 Semi-inner Products and Banach Spaces

Moving beyond Hilbert spaces means that we have to sacrifice "inner product" operation and all the nice (linear) properties it brings. This is the setting of Banach spaces, which deal exclusively with "norm" (which is non-linear) and norm-induced metric. More precisely, Banach spaces are normed vector spaces in which the norm-induced metric is "complete" (we do not define "complete" here, since it deals with the issue of denseness of subspaces). It suffices to think of a Banach space to be well-behaved both as a vector space and as a metric space.

A Hilbert space is a Banach space with an inner product. So the only new operation on a vector space V that is introduced into a Hilbert space is an inner product. Hence, in studying general Banach spaces, we need to carefully delineate which properties of a Hilbert space are due to the existence of an inner product, and which are generic properties of a Banach space. An important bridge is provided by the less-popular notion of "semi-inner product" on a normed vector space.

Semi-inner products were introduced into mathematics (Giles, 1967; Lumer, 1961) for the purpose of extending the geometric structure of inner product spaces to general vector spaces. A semi-inner product on a vector space V, hereby denoted by $[\cdot,\cdot]$, is a mapping from $V \times V$ to $\mathbb{C}$ such that for all $u, v, w \in V$ and $c \in \mathbb{C}$

(i) *positivity*: $\langle u, u\rangle \geq 0$ for all $u \in V$ and $\langle u, u\rangle = 0$ if and only if $u = 0$;
(ii) *linearity about the first variable*: $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$ and $\langle cu, v\rangle = c\langle u, v\rangle$;
(iii) *Cauchy-Schwarz inequality*: $|[u, v]|^2 \leq [u, u][v, v]$.

Compared with the definition of an inner product $\langle\cdot,\cdot\rangle$, the first two stipulations are identical, and only (iii) differs: the only property of inner products that a semi-inner product need not possess is conjugate symmetry, that is, $[u, v] \neq \overline{[v, u]}$, which is equivalent to non-linearity with respect to the second argument: $[w, u + v] \neq [w, u] + [w, v]$. Quite pleasingly, $[u, u]^{1/2} \equiv \|u\|$ defines a norm on V. Furthermore, as Giles (1967) showed, $[u, cv] = \bar{c}[u, v]$, that is, the semi-inner product is (conjugate) homogeneous with respect to its second argument. As an example, we give here semi-inner product on the space of all sequences $\ell^1(\mathbb{N})$ under ℓ^1-norm:

$$[u, v] := \|v\|_{\ell^1} \sum_{j\in\mathbb{N}} u_j \operatorname{sgn}(v_j), \quad u, v \in \ell^1(\mathbb{N})$$

where $\operatorname{sgn}(t)$ denotes the sign of $t \in \mathbb{R}$.

A semi-inner product on a vector space is a surrogate to the inner product. It plays similar roles as an inner product by providing a representation for continuous linear functionals (elements of the dual space) by elements of the original vector space (Riesz representation theorem). James (1964) showed an analogue to (in fact

an extension of) the classical Riesz representation theorem: every element of the dual space V^* can be represented by a unique element in V. Formally,

Theorem 8.2 (Riesz representation theorem for Banach space) *Assume that V is a reflexive Banach space. Every continuous linear functional v^* on V is representable as a v-indexed semi-inner product:*

$$v^*(u) = [u, v] \text{ for all} u \in V.$$

(Note that linearity with respect to the first argument of $[u, v]$ ensures that v^* is a linear functional.)

Semi-inner products also provide a means for measurement of angle (in particular defines a kind of orthogonality) between two vectors. In the absence of an inner product, one may still have a notion of "pseudo-orthogonality" (James, 1964), extending that of orthogonality:

Definition 8.3 A vector u is said to be *pseudo-orthogonal* with respect to v in a vector space V if $\|u + cv\| \geq \|u\|$ for all $c \in \mathbb{C}$.

Giles (1967) showed that u is pseudo-orthogonal to v if and only if $[v, u] = 0$. Pseudo-orthogonality is not a symmetric relation. The best approximation property, nevertheless, is still captured: v_0 is the best approximation of u to a closed subspace V_0 in V if and only if their difference $u - v_0$ is pseudo-orthogonal to V; the computation of v_0 involves a system of non-linear equations.

8.2.1.4 Duality Mapping and Generalized Semi-inner Products

The essence of semi-inner products is reflected in the duality mapping. Recall that starting from a vector space V, we defined its "dual" V^*, which is a vector space. One can continue this process and define the "double dual" V^{**} as the dual of V^*, two kinds of "triple duals," respectively, $(V^{**})^*$ as the dual of V^{**} and $(V^*)^{**}$ as the double dual of V^*, etc. It turns out that first, $(V^{**})^* = (V^*)^{**}$ so the notation V^{***} can be used unambiguously, and second, $V^{***} = V^*$. Here equality is in the sense of a "linear isomorphism." So what may appear at first as an endless process yields only two new vector spaces V^* and V^{**}, in addition to the vector space V one starts with. In general, $V \neq V^{**}$; it can be shown that V^{**} is larger than V and includes the latter as a subspace. When $V^{**} = V$, then the vector space V is called *reflexive*. When V^* and V are linearly isomorphic $V^* = V$, the vector space is called *self-dual*; this happens when the duality mapping is linear – in this case, the vector space is equipped with an inner product.

The standard semi-inner product introduced by Lumer (1961) in essence provides a "canonical" duality map between V and V^*, i.e., canonical in the sense of satisfying $\|v^*\|_{V^*} = \|v\|_V$. Nath (1971) relaxed the Cauchy–Schwarz inequality to Hölder inequality and obtained a definition of semi-inner products of order p:

(iii) Hölder inequality: $|[u, v]| \leq ([u, u])^{1/p}([v, v])^{1/q}, \quad u, v \in V,$

for $p, q \in (1, \infty)$ and satisfying $p^{-1} + q^{-1} = 1$. Nath's definition amounts to relaxing the duality mapping to $\|v^*\|_{V^*} = \|v\|_V^{p-1}$.

Zhang and Zhang (2010) investigated the most general form of a semi-inner product that might be compatible with a norm. Here, compatibility means that $[v, v]$ is indeed monotonically related to the norm $\|v\|_V$ endowed on V. It turns out that condition (iii) can be extended to the following inequality:

(iii) (generalization of Hölder inequality): $|[u, v]| \leq \varphi([u, u])\psi([v, v]), \ \ u, v \in V,$

where φ and ψ satisfy $\varphi(t)\psi(t) = t$, and some other conditions, including monotonicity. This corresponds to a generalized duality mapping requiring $\|v^*\|_{V^*} = \varphi^{-1}(\|v\|_V)/\|v\|_V$.

8.2.1.5 Bases in Hilbert and Banach Spaces

In a vector space V, one can find a set $E = \{u_j : j \in \mathbb{I}\}$ of linearly independent vectors to serve as its basis, so that each vector $v \in V$ can be expressed by as a sequence of numbers or "coordinates" using the basis (here we assume countable basis, with $\mathbb{I} = \{1, 2, \cdots, m\}$ an index set allowing $m = \infty$). Linear independence is in the sense that (i) no element (vector) in E can be written as a linear combination of other elements in it; and (ii) all elements of V can be written as a unique linear combination of elements of E.

When V is a Hilbert space with an inner product $\langle \cdot, \cdot \rangle$, it turns out that one can always construct an orthonormal basis such that: $\langle u_i, u_j \rangle = \delta_{ij}$, the Kronercker delta. In this case each vector in V has a simple decomposition:

$$u = \sum_{j \in \mathbb{I}} \langle u, u_j \rangle u_j.$$

In other words, the coordinates are simply given by projection of the vector u onto each of the basis vectors v_j's.

For a general vector space (not being equipped with an inner product), due to the lack of definition of orthogonality, we cannot hope for the existence of such an orthonormal basis. Instead, when V is reflexive, we can use the semi-inner product to construct *a pair* of pseudo-orthogonal bases $\{u_j : j \in \mathbb{I}\}$ and $\{v_j : j \in \mathbb{I}\}$:

$$[u_j, v_k] = \delta_{jk},$$

where pseudo-orthogonality between two vectors is in the sense of James discussed earlier. In analogy to the Hilbert space situation with an inner product, any element $u \in V$ of a Banach space with a semi-inner product can be decomposed as

$$u = \sum_{j \in \mathbb{I}} [u, v_j] u_j.$$

Note that the projection ($c_j = [u, v_j]$) and reconstruction ($u = \sum_j c_j u_j$) use two different sets of vectors (as opposed to one set, in the case of Hilbert base).

8.2.1.6 Frames in Hilbert and Banach Spaces

A basis of a vector space V is a minimum set that can be used to represent or reconstruct any element of V. The vectors in a basis "span" the entire space V. This is the irredundant representation of vectors, which is unique. Sometimes, it is convenient to consider over-redundant (and hence non-unique) representations, by allowing more vectors in the set than an independent set. Such representation is called a "frame," which brings more flexibility in representing and approximating vectors.

Formally, a collection of vectors $\{v_j \in V : j \in \mathbb{I}\}$ is called a frame for a Hilbert space V if there exist two constants α, β (with $0 < \alpha \leq \beta$) such that

$$\alpha \|u\|^2 \leq \sum_{j \in \mathbb{I}} |\langle u, v_j \rangle|^2 \leq \beta \|u\|^2$$

for all $u \in V$. When the frame $\{v_j : j \in \mathbb{I}\}$ satisfies linear independence, i.e., none of its members can be expressed as a linear combination of other members

$$v_j \notin \overline{\text{span}}\{v_k : k \in \mathbb{I};\ k \neq j\},$$

the frame is called a Riesz basis for a Hilbert vector space V.

Given a frame in a Hilbert space, we can construct an "analysis operator" $\mathcal{A} : V \to \ell^2(\mathbb{I})$ defined by

$$\mathcal{A}u := \{\langle u, v_j \rangle : j \in \mathbb{I}\}.$$

and a "synthesis operator" from $\ell^2(\mathbb{I})$ to V, as the dual operator $\mathcal{A}^\dagger$ of $\mathcal{A}$, given by

$$\mathcal{A}^\dagger c := \sum_{j \in \mathbb{I}} c_j v_j.$$

By the definition of frames, the operator $\mathcal{S} := \mathcal{A}^\dagger \mathcal{A}$ satisfies, $\forall u \in V$,

$$\alpha \|u\|^2 \leq \langle \mathcal{S}u, u \rangle \leq \beta \|u\|^2.$$

Thus, $\mathcal{S}$ is positive and has a bounded inverse. It is used to provide a sampling formula

$$u = \mathcal{S}^{-1}\mathcal{S}u = \sum_{j \in \mathbb{I}} \langle u, v_j \rangle (\mathcal{S}^{-1} v_j), \quad u \in V. \tag{8.1}$$

In a Hilbert space V, a dual frame of $\{v_j : j \in \mathbb{I}\}$ is any frame $\{u_j : j \in \mathbb{I}\}$ for V such that for all $u \in \mathcal{H}$

$$u = \sum_{j \in \mathbb{I}} \langle u, v_j \rangle u_j.$$

By Equation (8.1), $\{\mathcal{S}^{-1} v_j : j \in \mathbb{I}\}$ is a dual frame of $\{v_j : j \in \mathbb{I}\}$. It is called the canonical dual frame of $\{v_j : j \in \mathbb{I}\}$, where "canonical dual" is in the L^2 sense (Li, 1995). In the case of Banach spaces, frames and Riesz base can still be defined analogously despite of a lack of the inner product operator, which is instrumental for defining a Hilbert space frame and Riesz basis. Zhang and Zhang (2011) provided a definition using the semi-inner product, improving an earlier characterization by Casazza, Christensen, and Stoeva (2005).

Definition 8.4 Let X_d be a chosen discrete sequence space. We call $\{v_j : j \in \mathbb{I}\} \subseteq V$ a frame for V modeled on X_d if $\{[u, v_j] : j \in \mathbb{I}\} \in X_d$ for all $u \in V$ and there exist two constants α, β (with $0 < \alpha \leq \beta$) such that

$$\alpha\|u\|_V \leq \|\{[u, v_j] : j \in \mathbb{I}\}\|_{X_d} \leq \beta\|u\|_V \quad \text{for all } u \in V.$$

We call $\{v_j : j \in \mathbb{I}\} \subseteq V$ a Riesz basis for V modeled on X_d if $\text{span}\{v_j\} = V$, $\sum_j c_j v_j$ converges in X_d for all $c \in X_d$, and there exist two constants α, β (with $0 < \alpha \leq \beta$) such that

$$\alpha\|c\|_{X_d} \leq \|\sum_{j\in\mathbb{I}} c_j v_j\|_V \leq \beta\|c\|_{X_d} \quad \text{for all } c \in X_d.$$

Analogous expressions for analysis operator and synthesis operator were established for V and V^*, generalizing those for Hilbert spaces (Zhang & Zhang, 2011).

8.2.2 Function Spaces and Reproducing Kernels

Having introduced vector spaces and studied their properties, we now focus on a special kind of vector space, called "function spaces," in which elements are functions taking inputs from an arbitrary set. More specifically, we study the set $\mathbb{C}^{\mathcal{X}}$ of all complex-valued functions on a prescribed set $\mathcal{X}$. That the set of all functions $f : \mathcal{X} \to \mathbb{C}$ forms a vector space in its own right can be seen by verifying all the requirements of a vector space, where vector addition (i) and scalar multiplication (ii) are, respectively, defined by

$$(f + g)(x) = f(x) + g(x), \quad x \in \mathcal{X}$$

and

$$(cf)(x) = c(f(x)), \quad x \in \mathcal{X}$$

for $f, g \in \mathbb{C}^{\mathcal{X}}, c \in \mathbb{C}$. Below, these functions are also referred to as $\mathcal{X}$-functions, to remind the reader of the fact that the two operations (i) and (ii) are defined "point-wise" on $\mathcal{X}$ in forming the vector space $\mathcal{B}$ of functions,

$$\mathcal{B} = \{f : f \text{ is a } \mathcal{X}\text{-function}\},$$

regardless of the structure of the prescribed input space $\mathcal{X}$.

Recall (from Section 8.2.1.1) that a functional $T : V \to \mathbb{C}$ takes its inputs from a vector space V (whereas in general, a function $f : \mathcal{X} \to \mathbb{C}$ takes its inputs from an arbitrary space $\mathcal{X}$). In the case when the underlying vector space (serving as the input space for functionals) is a space of functions, then a functional maps any function to a number; a functional can be thought of as a "function of function" (as reflected by the suffix "-al"). For example, with respect to the function space $\mathcal{B}$ (on some prescribed set $\mathcal{X}$, without any structural assumptions about $\mathcal{X}$), a functional (whether linear or non-linear) on $\mathcal{B}$ takes any $\mathcal{X}$-function $f \in \mathcal{B}$ and produces a number.

norm $\|\cdot\|_{\mathcal{B}}$ of the function space $\mathcal{B}$ is an example of non-linear functionals, following the axioms that define the norm $\|\cdot\|_V$ of any vector space V. An important example of linear functionals is the "point evaluation" of a function. A point evaluation functional $T_{\text{ev}(\cdot)} : \mathcal{B} \to \mathbb{C}$ takes as input a $\mathcal{X}$-function $f : \mathcal{X} \to \mathbb{C}$ (an element of the function space $\mathcal{B}$) and gives as output a number (an element of $\mathbb{C}$) by restricting the $\mathcal{X}$-function at an evaluated point x_0:

$$T_{\text{ev}(x_0)}(f) = f(x)|_{x=x_0} = f(x_0)$$

where $x_0 \in \mathcal{X}$ is an element of the prescribed input space $\mathcal{X}$. The point evaluation functional is, of course, indexed by the given point x_0. The notion of "continuity" for evaluation functionals relies on the norm-induced topology on $\mathcal{B}$ as a vector space.

Apart from evaluation functionals, function spaces may also have "reproducing kernels." For a Banach space $\mathcal{B}$ (including Hilbert space $\mathcal{H}$) of functions $\mathcal{X} \to \mathbb{C}$, a reproducing kernel K ("kernels" for short) is a bivariate function: $\mathcal{X} \times \mathcal{X} \to \mathbb{C}$ that satisfies additional conditions. A consequence of Riesz representation theorem (Theorem 8.1) (and its Banach space extension, Theorem 8.2) is that point evaluation can be represented by an inner product (and its Banach space extension, semi-inner product) through the use of reproducing kernels. This is the mathematical foundation of theories of sampling and approximation, and the starting point for developing kernel methods for regularzied learning, whose goal is to generalize point evaluation from given sample inputs to all points (in a prescribed input space $\mathcal{X}$).

8.2.2.1 Reproducing Kernel Hilbert Spaces (RKHS)

In a Hilbert space $\mathcal{H}$ of functions, reproducing kernels are such that $K(x_0, \cdot) \in \mathcal{H}, K(\cdot, x_0) \in \mathcal{H}$ for all $x_0 \in \mathcal{X}$. Reproducing kernels offer a convenient device for "sampling" the function at input points. The reproducing property of the kernel K links any function $f : \mathcal{X} \to \mathbb{C}$ to its point evaluation:

$$f(x_0) = \langle f(\cdot), K(\cdot, x_0)\rangle = \langle K(x_0, \cdot), f(\cdot)\rangle \tag{8.2}$$

with $\langle\cdot,\cdot\rangle$ denoting the inner product of the RKHS satisfying $K(x, x') = K(x', x)$ and

$$K(x, x') = \langle K(x, \cdot), K(\cdot, x')\rangle.$$

For this reason, the reproducing kernel plays the role of an "evaluation functional" – given a function and a point $x_0 \in \mathcal{X}$ of the sample space, K is used to return a number which is the function's value $f(x_0)$ at that point. A Reproducing Kernel Hilbert Space (RKHS) is defined as a function space in which point evaluations are continuous. Every RKHS is completely characterized by a kernel function K; the space is normed by $\|\cdot\|_{\mathcal{H}_K}$, a 2-norm as induced by its inner product $\langle\cdot,\cdot\rangle$:

$$K(x, x) = \langle K(x, \cdot), K(\cdot, x)\rangle = \|K(x, \cdot)\|^2_{\mathcal{H}_K}.$$

(Note the 2-norm is on the Hilbert space of functions $\mathcal{X} \to \mathcal{C}$, and not on the input space $\mathcal{X}$.) Well-known examples of kernel functions are polynomials, radial basis functions, etc. It is long known that a bivariate function K can become a reproducing kernel if and only if the corresponding kernel matrix $K(x_i, x_j), i = 1, \cdots, n$ is positive semi-definite for any arbitrary set of sample points in $\mathcal{X}$ and for any set-size n (Aronszajn, 1950), or that there exists a mapping Φ from $\mathcal{X}$ to some inner product space $\mathcal{W}$ such that

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{W}}, \;\; x, x' \in \mathcal{X}.$$

8.2.2.2 Reproducing Kernel Banach Spaces (RKBS)

We define an RKBS as a reflexive Banach space $\mathcal{B}$ of functions on $\mathcal{X}$ such that the dual space $\mathcal{B}^*$ is isometrically isomorphic to (and hence can be identified with) a Banach space of functions on $\mathcal{X}$ and point evaluations are continuous linear functionals on both $\mathcal{B}$ and $\mathcal{B}^*$ (as a space of functions). A bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a reproducing kernel in a RKBS if and only if there exist mappings Φ from $\mathcal{X}$ to some reflexive Banach space $\mathcal{W}$ and $\Phi^* : \mathcal{X} \to \mathcal{W}^*$ such that

$$K(x, x') = (\Phi(x), \Phi^*(x'))_{\mathcal{W}}, \;\; x, x' \in \mathcal{X}.$$

where $(\cdot, \cdot)_{\mathcal{W}}$ denotes the "dual pairing" of $\mathcal{W}$ with $\mathcal{W}^*$, which is a bilinear form. In this case, $K(x, \cdot) \in \mathcal{B}$, $K(\cdot, x) \in \mathcal{B}^*$ for all $x \in \mathcal{X}$, and K has the following reproducing properties:

$$f(x) = (f, K(\cdot, x))_{\mathcal{B}}, \;\; f^*(x) = (K(x, \cdot), f^*)_{\mathcal{B}}, \; f \in \mathcal{B}, \; f^* \in \mathcal{B}^*.$$

Here $(\cdot, \cdot)_{\mathcal{B}}$ is a bilinear form ("dual pairing") on $\mathcal{B} \times \mathcal{B}^*$:

$$(f, g^*) := g^*(f), \;\; f \in \mathcal{B}, \; g^* \in \mathcal{B}^*.$$

As a special case of RKBS, we have the reproducing kernel Hilbert space (RKHS), where $\mathcal{B}$ is simply a Hilbert space $\mathcal{H}$ of functions. A Hilbert space is a Banach space with the addition of *any* of the following conditions:

(i) the semi-inner product $[\cdot, \cdot]$ on $\mathcal{B}$ is conjugate symmetric: $[f, g] = \overline{[g, f]}$ (and hence is an inner product);
(ii) the dual space $\mathcal{B}^*$ of $\mathcal{B}$ is isometrically isomorphic to $\mathcal{B}$;
(iii) the duality mapping from $\mathcal{B} \ni f \mapsto f^* \in \mathcal{B}^*$ is linear.

8.2.2.3 Operator-valued Reproducing Kernels

In the above discussions, the Banach space $\mathcal{B}$ of functions f is scalar-valued, $f : \mathcal{X} \to \mathbb{C}$. In machine learning applications (especially kernel learning), there is a need to consider vector-valued functions $f : \mathcal{X} \to \mathbb{C}^m$, where m is the dimension of the output space $\Lambda = \mathbb{C}^m$.

The first step is to extend the Reproducing Kernel Hilbert Space (RKHS) to vector-valued RKHS, as formally defined below (Micchelli & Pontil, 2005a,b).

Definition 8.5 A Λ-valued RKHS is a Hilbert space $\mathcal{H}$ of functions from the input space $\mathcal{X}$ to the output vector space Λ such that for each $x \in \mathcal{X}$ and for each $f \in \mathcal{H}$, the point evaluation $f \mapsto f(x)$ is a continuous linear operator from $\mathcal{H}$ to Λ.

In this definition, $\mathcal{H}$ is the space of vector-valued functions with an inner product operation $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ taking any two such functions into a number $\mathbb{C}$. A kernel $\mathcal{K}$ in a Λ-valued RKHS is a bivariate, operator-valued functions of $\mathcal{X}$ such that, when evaluated in each variable, the operator-valued functions $\mathcal{K}(\cdot, x)$ and $\mathcal{K}(x, \cdot)$ satisfy $\forall u \in \Lambda$

$$\langle f(x), u \rangle_{\Lambda} = \langle f(\cdot), \mathcal{K}(\cdot, x)u \rangle_{\mathcal{H}}, \qquad \langle \mathcal{K}(x, \cdot)u, f(\cdot) \rangle_{\mathcal{H}} = \langle u, f(x) \rangle_{\Lambda}, \tag{8.3}$$

and hence they are equal. Here, $\mathcal{K}(x, x')$ is a linear operator (an $m \times m$ matrix) from Λ to Λ; when we speak of $\mathcal{K}(\cdot, \cdot)$ as a bivariate function, we mean that it takes input from $\mathcal{X} \times \mathcal{X}$ and outputs such an operator over vector space Λ, that is, an $m \times m$ matrix. $\mathcal{K}$ serves as a point evaluation map for the vector-valued function f. The kernel $\mathcal{K}$ satisfies the property that $\forall u, v \in \Lambda$

$$\langle u, \mathcal{K}(x, x')v \rangle_{\Lambda} = \langle \mathcal{K}(x, \cdot)u, \mathcal{K}(\cdot, x')v \rangle_{\mathcal{H}} = \langle \mathcal{K}(x', x)u, v \rangle_{\Lambda} \tag{8.4}$$

and that $\mathcal{K}(x, x') = (\mathcal{K}(x', x))^{\dagger}$, where $\dagger$ stands for adjoint transformation of an operator (i.e., conjugate transpose of a matrix) in the sense that

$$\langle \mathcal{K}u, v \rangle_{\Lambda} = \langle u, \mathcal{K}^{\dagger} v \rangle_{\Lambda}. \tag{8.5}$$

Clearly, $\mathcal{K}(x, x)$ is self-adjoint and hence non-negative, satisfying the following:

$$|\langle \mathcal{K}(x, x')u, v \rangle_{\Lambda}| \leq \langle \mathcal{K}(x, x)u, u \rangle_{\Lambda}^{1/2} \langle \mathcal{K}(x', x')v, v \rangle_{\Lambda}^{1/2} = \|\mathcal{K}(x, \cdot)u\|_{\mathcal{H}} \|\mathcal{K}(x', \cdot)v\|_{\mathcal{H}}. \tag{8.6}$$

Micchelli and Pontil (2005a,b) showed that an operator-valued bivariate function $\mathcal{K}$ is a reproducing kernel if and only if for any sequence ($i \in \mathbb{I}$) of points $x_i \in \mathcal{X}$ and any sequence of bounded vectors $u_i \in \Lambda$,

$$\sum_{i,j} \langle \mathcal{K}(x_i, x_j)u_j, u_i \rangle_{\Lambda} \geq 0.$$

This is analogous to Aronszajn's (1950) characterization of reproducing kernels for scalar-valued Hilbert spaces.

The above result has been extended to vector-valued Reproducing Kernel Banach Spaces in Zhang and Zhang (2013), with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ in Equations (8.3), (8.4) and (8.6) replaced by semi-inner product $[\cdot, \cdot]_{\mathcal{B}}$. In particular, instead of Equation (8.5), the *generalized adjoint* $\mathcal{K}^{\dagger}$ of an operator $\mathcal{K}$ is defined by

$$[\mathcal{K}u, v]_{\Lambda} = [u, \mathcal{K}^{\dagger} v]_{\Lambda}. \tag{8.7}$$

8.2.2.4 Measurable Functions as Vector Spaces: $L^p(\mathcal{X})$, $M(\mathcal{X})$, $C_0(\mathcal{X})$, etc.
When the underlying set $\mathcal{X}$ is endowed with a measure, the space of measurable functions on $\mathcal{X}$ has further properties. For example, the space $L^p(\mathcal{X}, \Omega, \mu)$ ($1 \leq p < +\infty$) of μ-measurable functions on $\mathcal{X}$, when endowed with the norm

$$\|f\|_{L^p} := \left(\int_\Omega |f(x)|^p d\mu\right)^{1/p} < +\infty,$$

is a normed vector space. When $p = +\infty$, $L^\infty(\mathcal{X}, \Omega, \mu)$ can be endowed with the following norm

$\|f\|_{L^\infty} = \inf\{\alpha > 0 : \{x : |f(x)| > \alpha\}$ is a set with measure zero with respect to $\mu\}$.

Generalizations of the above L^p spaces include the Orlicz spaces (Rao & Ren, 1991) and L^p spaces over other measure spaces.

The dual space of $L^p(\mathcal{X}, \Omega, \mu)$ when $1 < p < +\infty$ can be identified with $L^q(\mathcal{X}, \Omega, \mu)$, where q is the conjugate number of p satisfying $p^{-1} + q^{-1} = 1$. More accurately, by identification, we mean that there exists a mapping $L^q \ni g \mapsto \int_\Omega fg d\mu$, indexed by g, that defines an isometric isomorphism of $L^q(\mathcal{X}, \Omega, \mu)$ onto $(L^p(\mathcal{X}, \Omega, \mu))^*$. The duality mapping from an element f in $L^p(\mathcal{X}, \Omega, \mu)$ ($1 < p < +\infty$) to an element f^* of $L^q(\mathcal{X}, \Omega, \mu)$ is given by

$$f \to f^* = \frac{\bar{f}|f|^{p-2}}{\|f\|_{L^p}^{p-2}}.$$

semi-inner products on $L^p(\mathcal{X}, \Omega, \mu)$, $(1 \leq p < \infty)$ take the form:

$$[g, f] := \frac{\int_\Omega g\bar{f}|f|^{p-2} d\mu}{\|f\|_{L^p}^{p-2}}.$$

The space $L^2(\mathcal{X}, \Omega, \mu)$, a special case of $L^p(\mathcal{X}, \Omega, \mu)$, is a Hilbert space.

Complications arise when $p = 1$ and $q = \infty$. The dual space of $L^1(\mathcal{X}, \Omega, \mu)$ is $L^\infty(\mathcal{X}, \Omega, \mu)$ when μ is σ-finite, so the latter maps onto $(L^1(\mathcal{X}, \Omega, \mu))^*$ isometric isomorphically through the map: $L^\infty \ni g \mapsto \int_\Omega fg d\mu$. However, neither L^1 nor L^∞ is reflexive as long as they are infinite-dimensional. As a matter of fact, $(L^\infty(\mathcal{X}, \Omega, \mu))^* = G(\mathcal{X}, \mu)$, where $G(\mathcal{X}, \mu)$ is the set of finitely additive measures on $\mathcal{X}$ that is absolutely continuous w.r.t. μ and have finite total variation. In general, the space $G(\mathcal{X}, \mu)$ is strictly larger than $L^1(\mathcal{X}, \Omega, \mu)$.

An important subspace of $L^\infty(\mathcal{X}, \Omega, \mu)$ is the space $C_0(\mathcal{X})$ of continuous functions f on $\mathcal{X}$ such that for all $\varepsilon > 0$, $\{x \in \mathcal{X} : |f(x)| \geq \varepsilon\}$ is compact, endowed with the sup norm

$$\|f\|_{C_0} = \max_{x \in \mathcal{X}} |f(x)|.$$

The dual of $C_0(\mathcal{X})$ is $M(\mathcal{X})$, the set of regular Borel measures on $\mathcal{X}$ with total variation norm (note the Borel measure can be real or complex-valued). For $\mu \in M(\mathcal{X})$, $\|\mu\|$ is defined as $|\mu|(\mathcal{X}) = \mu_+(\mathcal{X}) + \mu_-(\mathcal{X})$. The mapping $M(\mathcal{X}) \ni \mu \mapsto \int_\mathcal{X} f d\mu$ is an isometric isomorphism of $M(\mathcal{X})$ onto $C_0(\mathcal{X})^*$. The dual of $M(\mathcal{X})$ (as a normed vector space) is $L^\infty(M(\mathcal{X}))$, see Conway (1990, p. 76).

When $\mathcal{X}$ is discrete, then $\ell^1 = L^1(\mathbb{N}, 2^{\mathbb{N}}, \mu) = M(\mathbb{N})$, whose dual space is $c_0 = C_0(\mathbb{N})$, with $c_0 \subset l^\infty = L^\infty(\mathbb{N}, 2^{\mathbb{N}}, \mu)$ and μ having measure 1 at each point in $\mathbb{N}$.

From measure theory, the specification of measurable functions depends on the measure itself. If $(\mathcal{X}, \Omega, \nu)$ is a σ-finite measure space, and μ is a complex-valued measure on $(\mathcal{X}, \Omega)$ that is absolutely continuous w.r.t. ν, then there is a unique complex-valued function f in $L^1(\mathcal{X}, \Omega, \nu)$ such that $\mu(\Delta) = \int_\Delta f d\nu$ for every $\Delta \in \Omega$. The quantity $f = d\mu/d\nu$ is called the Radon–Nikodym derivative. We have the following facts: (i) for any $g \in L^1(\mathcal{X}, \Omega, \mu)$ then $gf \in L^1(\mathcal{X}, \Omega, \nu)$ and $\int g d\mu = \int g f d\nu$; (ii) for $\Delta \in \Omega, |\mu|(\Delta) = \int_\Delta |f| d\nu$. This is to say, $|d\mu/d\nu| = d|\mu|/d\nu$; (iii) for $\Delta \in \Omega$, there exists an f such that $|f| = 1$ a.e. w.r.t. $|\mu|$, and $\mu(\Delta) = \int_\Delta f d|\mu|$.

8.2.2.5 Integral Operators and Spectral Methods

For a Hilbert space of measurable functions, a key operator that had been studied (see, e.g., Smale & Zhou, 2007) is the linear integral operator ("Mercer map") $T_\mu : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$:

$$(T_\mu f)(\cdot) \equiv \int_{\mathcal{X}} K(\cdot, y) f(y) d\mu(y) \tag{8.8}$$

associated with a positive semi-definite kernel K. When $\mathcal{X}$ is compact, the so-called Mercer's Theorem guarantees that the spectrum of T is discrete with the existence of an orthonormal basis. In fact, the above map (8.8) can be shown to be injective, with the range of the operator T_μ exactly being the RKHS with kernel K. This is the mathematical foundation underlying the heat equation/graph Laplacian approach to manifold learning (Coifman & Maggioni, 2006). The class of so-called "spectral clustering" methods in machine learning uses graph Laplacian as the kernel (Ng, Jordan, & Weiss, 2002; Weiss, 1999).

To introduce the details of the Mercer theorem, let us assume that $\mathcal{X}$ is a compact metric space, μ is a finite positive Borel measure on $\mathcal{X}$, and K is a continuous reproducing kernel on $\mathcal{X}$. In this case, the integral operator T_μ is bounded, compact, self-adjoint, and positive. By the spectral theorem for compact operators on a Hilbert space, there exists an orthonormal basis $\{u_j : j \in \mathbb{I}\}$ for $L^2(\mathcal{X}, \mu)$ and a sequence $\{\lambda_j : j \in \mathbb{I}\}$ of non-increasing non-negative constants that tend to zero as j goes to infinity such that

$$T_\mu u_j = \lambda_j u_j.$$

The Mercer theorem states that if μ is non-degenerate, that is, every non-empty open subset in $\mathcal{X}$ has positive measure under μ, then

$$K(x, y) = \sum_{j \in \mathbb{N}} \lambda_j u_j(x) \overline{u_j(y)}, \quad x, y \in \mathcal{X},$$

where the series converges absolutely and uniformly for $x, y \in \mathcal{X}$.

The structure of the RKHS $\mathcal{H}_K$ with kernel K becomes clearer in view of the above expansion. For simplicity, let us assume that all the eigenvalues λ_j of T_μ are

positive. The space $\mathcal{H}_K$ then consists of all the functions of the form $\sum_{j\in\mathbb{N}} a_j u_j$ where

$$\sum_{j\in\mathbb{N}} \frac{|a_j|^2}{\lambda_j} < +\infty.$$

The inner product in $\mathcal{H}_K$ takes the form

$$\left\langle \sum_{j\in\mathbb{N}} a_j u_j, \sum_{j\in\mathbb{N}} b_j u_j \right\rangle = \sum_{j\in\mathbb{N}} \frac{a_j \overline{b_j}}{\lambda_j}.$$

Considering the effect of changing measures in the Mercer theorem, we let ν be another finite positive Borel measure on $\mathcal{X}$ that is absolutely continuous with respect to μ. Set $g = \frac{d\nu}{d\mu}$. Also denote by T_μ the integral operator associated with μ and the kernel K, and by T_ν that associated with ν and kernel

$$\tilde{K}(x,y) = \frac{K(x,y)}{\sqrt{g(x)g(y)}}, \quad x, y \in \mathcal{X}.$$

One sees that

$$T_\nu\left(\frac{u_j}{\sqrt{g}}\right) = \frac{1}{\sqrt{g}} T_\mu(u_j) = \lambda_j \frac{u_j}{\sqrt{g}} = \lambda_j \left(\frac{u_j}{\sqrt{g}}\right).$$

In other words, T_ν and T_μ have the same eigenvalues, and the eigenfunctions differ by $\sqrt{g}$, the square-root of the Radon–Nikodym derivative $d\nu/d\mu$.

8.3 Learning as Regularization: Problem Formulation

In this section, we rigorously formulate the problem of learning from finite samples using mathematical tools. We will review Statistical Learning Theory, and highlight the importance of "generalization" over mere (retroactive) goodness-of-fit as the goal of learning. We emphasize decomposing errors of generalization into various sources, and investigate different regularizers used in kernel methods to achieve smoothness or sparsity objectives.

8.3.1 Statistical Learning Theory

Learning to classify inputs through finite samples is an intrinsically ill-posed problem (that is, underconstrained), and *regularization* is a key towards recovery of stable and generalizable solutions. In fact, statistical machine learning is now laid on the solid foundation of the *regularized loss minimization* framework (Cucker & Smale, 2002; Poggio & Smale, 2003). In its application to classification, the first step is to define an *empirical error* term which measures how well the model fits the data. The goal of learning is, however, not to just fit well existing data, but to generalize to new inputs. Learning theory shows that, to achieve good generalization, the empirical error should be minimized subject to (the constraint of) a data-dependent upper bound on the complexity of the function space.

8.3.1.1 Assumptions on Input and Output Spaces

The prescribed input space $\mathcal{X}$, such as the set of human faces, consists of a set of examples, with minimal structural assumptions. Typically, $\mathcal{X}$ is assumed to be measurable (i.e., admits a Borel measure), so one can quantify the relative frequency that any particular example occurs. It is not necessary to assume $\mathcal{X}$ to explicitly have a vector space structure or a metric space structure. In the reproducing kernel methods, input structure is implicitly encoded in the kernel function $K(\cdot,\cdot)$ defined on $\mathcal{X}\times\mathcal{X}$: the kernel function will have an associated vector space (i.e., feature space) and associated metric (via norm on the feature space). On the other hand, the output space $\mathcal{Y}$ can be a number field such as $\mathbb{R}$ (resulting in scalar-valued functions to be learnt), or a vector space $\mathbb{R}^n$ (resulting in vector-valued functions to be learnt). The set of known pairings, i.e., samples, is denoted as $\mathcal{Z} = \{(x_j, y_j) : j = 1, \cdots, n\}$. The joint probability distribution on $(\mathcal{X},\mathcal{Y})$ is assumed to exist, and expectation with respect to it is denoted as $\mathrm{E}_{x,y}$.

8.3.1.2 Learning as Optimization Problem

Classification algorithms based on regularization learning solve the following optimization problem

$$\min_{f\in\mathcal{B}} \{L_{\mathcal{Z}}(f) + \lambda R(f)\} = \inf_{f\in\mathcal{B}} \frac{1}{n}\sum_{j=1}^{n} L(y_j, f(x_j)) + \lambda R(f) \tag{8.9}$$

where

- f denotes a learned input-output mapping, called "model";
- $\mathcal{B}$ is the space of possible models, i.e., the hypothesis space $\mathcal{B} \ni f$;
- $\mathcal{Z}$ denotes the training set or sample data, consisting of known input–output pairs;
- $L_{\mathcal{Z}}(f)$ is the empirical error of the model f on the training set $\mathcal{Z}$;
- R is a regularizer on the hypothesis space "penalizing" more complex models and preferring simpler ones (Occam's razor);
- λ is a positive regularization parameter that balances the need for data-fitting and for complexity-reduction, operationalizing "optimal" generalization.

In the celebrated RKHS methods (Schölkopf & Smola, 2001; Shawe-Taylor & Cristianini, 2004), the space of models being considered consists of the Hilbert function space $\mathcal{H}$ with a penalty of the form $R(f) = \phi(\|f\|_{\mathcal{H}})$, with the norm measuring the complexity or smoothness of the function, and $\phi(\cdot)$ a non-negative function. The sample set $\mathcal{Z}$ consists of n pairs of points in input space $\mathcal{X}$ and output space $\mathcal{Y} \equiv \mathbb{R}$, with empirical loss (measuring a classifier's performance on known samples) given by $L_{\mathcal{Z}}(f) = \frac{1}{n}\sum_{j=1}^{n} L(f(x_j), y_j)$, $L(\cdot,\cdot) \geq 0$ being a proper loss function that achieves 0 if its two arguments take equal value.

8.3.1.3 Risk of Generalization and Error Decomposition

From a statistical learning perspective, the generalization performance of any proposed learning scheme based on finite samples is evaluated by measuring a

quantity called "excess risk." Formally, denote $f_{\mathcal{Z},\lambda}$ as the optimal solution to Equation (8.9), we are interested in bounds of the excess risk $\Delta(f_{\mathcal{Z},\lambda})$ defined as

$$\Delta(f_{\mathcal{Z},\lambda}) = \mathrm{E}_{x,y}[L(y,f_{\mathcal{Z},\lambda}(x))] - \min_{f\in\mathcal{B}} \mathrm{E}_{x,y}[L(y,f(x))]$$

where $\mathcal{B}$ is constructed based on the distribution $\mu(x)$ over the input space $\mathcal{X}$.

The excess risk $\Delta(f_{\mathcal{Z},\lambda})$ can be decomposed into three components, reflecting *sample error*, *approximation error*, and *hypothesis error* as different sources for the risk about modeling. To this end, we define the generalization error $\mathrm{Er}(f)$ and empirical error $\mathrm{Er}_{\mathcal{Z}}(f)$ as

$$\mathrm{Er}(f) = \mathrm{E}_{x,y}[L(y,f(x))],\ \mathrm{Er}_{\mathcal{Z}}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i))$$

and

$$f_* = \arg\min_{f\in\mathcal{B}} \mathrm{Er}(f),\ f_\lambda = \arg\min_{f\in\mathcal{B}} \left(\mathrm{Er}(f) + \lambda\|f\|_{\mathcal{B}}\right).$$

According to Cucker and Smale (2002), the generalization error of $f_{\mathcal{Z},\lambda}$ is decomposed as

$$\Delta(f_{\mathcal{Z},\lambda}) \le -\lambda\|f_{\mathcal{Z},\lambda}\|_{\mathcal{B}^{(n)}} + S(\mathcal{Z},\lambda) + H(\mathcal{Z},\lambda) + D(\lambda),$$

where sample error $S(\mathcal{Z},\lambda)$, hypothesis error $H(\mathcal{Z},\lambda)$, and approximation error $D(\lambda)$ are given by

$$\begin{aligned} S(\mathcal{Z},\lambda) &= \mathrm{Er}(f_{\mathcal{Z},\lambda}) - \mathrm{Er}_{\mathcal{Z}}(f_{\mathcal{Z},\lambda}) + \mathrm{Er}_{\mathcal{Z}}(f_\lambda) - \mathrm{Er}(f_\lambda),\\ H(\mathcal{Z},\lambda) &= \left\{\mathrm{Er}_{\mathcal{Z}}(f_{\mathcal{Z},\lambda}) + \lambda\|f_{\mathcal{Z},\lambda}\|_{\mathcal{B}^{(n)}}\right\} - \left\{\mathrm{Er}_{\mathcal{Z}}(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{B}}\right\},\\ D(\lambda) &= \mathrm{Er}(f_\lambda) - \mathrm{Er}(f_*) + \lambda\|f_\lambda\|_{\mathcal{B}}. \end{aligned}$$

The sample error $S(\mathcal{Z},\lambda)$ can be well bounded using the standard concentration inequality technique, and the approximation error $D(\lambda)$ is often assumed to be $O(\lambda^q)$ where $q \in (0,1]$. The main challenge is to bound the hypothesis error $H(\mathcal{Z},\lambda)$, which in the conventional analysis of generalization error is always upper bounded by zero due to the fact $f_{\mathcal{Z},\lambda}$ optimizes the objective function in Equation (8.9). Here, $\mathcal{B}^{(n)}, n \in \mathbb{N}$ is a sequence of Banach spaces used to model input–output mapping functions as sample size n increases. Suppressing H amounts to finding the series of Banach spaces such that $\lim_{n\to\infty}\mathcal{B}^{(n)} = \mathcal{B}$ where convergence is with respect to the probability distribution under which $\mathrm{E}_{x,y}$ is performed.

8.3.2 Regularization by Different Norms

Effective regularization is key to solving the optimal classification problem. Various norms are employed as regularizers for achieving different goals.

8.3.2.1 Norms to Enforce Smoothness

The RKHS theory provides that specifying a kernel function K amounts to specifying some feature mapping $\Phi(\cdot) : x \mapsto \Phi(x)$, where $\Phi(x)$ is an internal representation of input x, i.e., $\Phi(x)$ is a vectorial representation of x's "features." The kernel function itself measures the similarity between any two inputs x_i and x_j; the measure is necessarily symmetric: $K(x_i, x_j) = K(x_j, x_i)$. Such similarity measure has the natural interpretation of projections of two vectors $\Phi(x_i)$ and $\Phi(x_j)$ onto one another in the (implicitly referenced) feature space $\mathcal{W}$, which is a Hilbert space itself.

There are many advantages of the RKHS approach owing to the existence of an inner product in a Hilbert space. In particular, computation of the similarity between inputs need not rely on the availability of the explicit form of Φ; rather, similarity is typically given directly in some analytic form of a reproducing kernel K (e.g., a Radial Basis Function). However, this L^2-based approach has its limitations, most notably the assumed symmetry of the similarity structure (as modeled by the symmetric inner product), in violation of empirical evidence from human cognition (see below).

Besides regularization with RKHS norms, there have been other choices of norms that are Fréchet differentiable. Regularization about these norms typically results in a smooth minimization problem and thus possesses efficient numerical algorithms. Learning with such norms includes the ℓ^p-coefficient regularization with $p > 1$ (Tong, Chen, & Yang, 2010), and the smooth RKBS approach by Zhang, Xu, & Zhang (2009).

8.3.2.2 Norms to Enforce Sparsity

On par with the massive amount of work on RKHS methods, which is essentially L^2-based, there is an equal body of work on stimulus/signal representation based on ℓ^1-regularization and sparse approximation. Thanks to the development of Lasso in statistics (Tibshirani, 1996), compressive sensing (Candès, Romberg, & Tao, 2006), and cubic ℓ^1 spline method for minimum norm interpolation (Lavery, 2000), it is now recognized that it is the ℓ^1-norm that enforces sparsity solution and optimal feature selection. Unlike RKHS, where the existence of a feature map is guaranteed but does not need to be computed, properly selecting a feature basis is crucial for the success of ℓ^1 regularization.

It should be noted that ℓ^1-regularization is seen as a convex surrogate to ℓ^0-regularization, which is truly the sparsity-enforcing norm (since ℓ^0-norm counts the number of non-zero components of a vector). Because of combinatoric explosion, ℓ^0-regularization is computationally unfeasible, and therefore replaced by ℓ^1-regularization (Candès & Tao, 2004).

In pursuit of further sparsity, ℓ^q norms with $0 < q < 1$ have been considered (see, e.g., Foucart & Lai 2009). Such semi-norms are even closer to the ℓ^0-norm compared with the ℓ^1-norm. In particular, $\ell^{1/2}$-regularization (Xu *et al.*, 2010) has received considerable attention for its advantages in developing efficient numerical methods.

8.3.2.3 Different Norms on Kernelized Functions

Given arbitrary n sample points in the input space $\mathcal{X}$ and a bivariate function K on $\mathcal{X} \times \mathcal{X}$, consider the Banach space $\mathcal{B}$ of functions on $\mathcal{X}$ which contains the following linear combinations of its elements

$$\sum_{j=1}^{n} c_j K(x_j, \cdot) \tag{8.10}$$

for arbitrary $c = \{c_j \in \mathbb{C} : j = 1, \cdots, n\}$. We can endow the space $\mathcal{B}$ with the following ℓ^1-norm:

$$\left\| \sum_{j=1}^{n} c_j K(x_j, \cdot) \right\|_{\mathcal{B}} = \sum_{j=1}^{n} |c_j| = \|c\|_{\ell 1}. \tag{8.11}$$

or an L_2 (RKHS) norm

$$\left\| \sum_{j=1}^{n} c_j K(x_j, \cdot) \right\|_{\mathcal{B}} = \left(\sum_{i,j=1}^{n} c_i K(x_i, x_j) c_j \right)^{1/2} = \|c\|_K. \tag{8.12}$$

Such flexibility highlights the fact that reproducing kernels are useful tools to construct the function space where functions (as vectors) may be endowed with various norms determined by practical applications.

8.4 Learning with Reproducing Kernels: Solution Concepts

This section provides an in-depth explanation of the celebrated "kernel methods." The suite of methods include representer theorem, feature map, "kernel trick" and kernelization, maximal margin classifier, etc. These computational tools are grounded in solid mathematical techniques and derivations, and support the conceptual core for mathematical modeling of human categorization.

8.4.1 Power of Reproducing Kernels

In this section, we review the usefulness of reproducing kernels of a function space in regularized learning. Among those are (i) representer theorem, which allows the optimal solution of the regularized learning problem to be representable through the kernel function; (ii) feature map, which allows an input stimulus to have feature representation in a vector space; and (iii) kernel trick, which allows the computation to bypass exact feature representation.

8.4.1.1 Representer Theorem

The now-popular RKHS theory guarantees that the optimal classifier f_{opt} is a weighted sum of similarities (as measured through a symmetric, positive semi-definite kernel function K) to the sampled points:

$$f_{opt}(\cdot) = \sum_{j=1}^{n} c_j K(x_j, \cdot). \tag{8.13}$$

The above statement is known as the "representer theorem."

In the case of RKBS in general, denote $G(\cdot, x) = K(x, \cdot)^*$. Following the discussions of Section 3.1.2, we formulate the following regularized risk minimization problem (under a general regularization function ϕ)

$$\inf_{f \in \mathcal{B}} \frac{1}{n} \sum_{j=1}^{n} L(f(x_j), y_j) + \lambda \phi(\|f\|_{\mathcal{B}}), \tag{8.14}$$

where $\mathcal{B}$ is a Banach space of functions $f : \mathcal{X} \to \mathbb{C}$. Its solution f_{opt} is characterized by the following "representer theorem":

Theorem 8.6 (Zhang & Zhang, 2012) *Let L be continuous and proper (i.e., $L(s,t) \geq 0$, achieving 0 if and only if $t = s$) and let ϕ be continuous and non-decreasing with $\lim_{t\to\infty} \phi(t) = +\infty$. Then (8.14) has a minimizer f_{opt} in $\mathcal{B}$, which has the form*

$$f^*_{opt}(\cdot) = \sum_{j=1}^{n} c_j \, K(x_j, \cdot)^* = \sum_{j=1}^{n} c_j \, G(\cdot, x_j) \tag{8.15}$$

for some constants $c_j \in \mathbb{C}, j = 1, \cdots, n$. Here f^*_{opt} represents the dual of f_{opt}.

Note that, unlike the RKHS case, the kernels are generally not positive definite once the number of sampling points exceeds two. The essence of a representer theorem is to represent the dual function of the minimizer as a linear combination of the point evaluation functionals at x_j, $1 \leq j \leq n$. The fact that the representer theorem is really a statement about an element of the dual space has not been previously appreciated due to the reason that in an RKHS, $f^* = f$.

8.4.1.2 Feature Spaces and Feature Maps

If stimuli are represented as individual points in the sample space $\mathcal{X}$, similarity between inputs in $\mathcal{X}$ can be measured by the inner product of their respective features in some feature space, which is a vector space denoted as $\mathcal{W}$. In the standard RKHS theory, the feature space $\mathcal{W}$ is further assumed to be endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$. The feature map $\Phi(\cdot) : \mathcal{X} \to \mathcal{W}$ is captured by the reproducing kernel K:

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{W}}, \tag{8.16}$$

that is necessarily symmetric: $K(x_i, x_j) = K(x_j, x_i)$. The important advantage for the RKHS framework is its *implicit* reference to a feature representation (i.e., mapping Φ) – note that inputs in $\mathcal{X}$ for classification typically do not have a Hilbert space or even vector space structure for that matter; however, one is able to assume the existence of a (usually much higher dimensional) feature space (which is a Hilbert space) that the input space can be mapped into.

These same considerations can be extended to Banach spaces in general. We have proven:

Theorem 8.7 (Zhang, Xu, & Zhang, 2009) *Let $\mathcal{B}$ be a uniform Banach space of functions on an input space $\mathcal{X}$ where point evaluations are continuous linear functionals. Then*

1. *There exists a unique function (called "semi-inner product reproducing kernel") $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $G(x, \cdot) \in \mathcal{B}$ for all $x \in \mathcal{X}$ and*
$$f(x) = [f, G(x, \cdot)]_{\mathcal{B}} \text{ for all } f \in \mathcal{B} \text{ and } x \in \mathcal{X}. \tag{8.17}$$
2. *A function $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the reproducing kernel of $\mathcal{B}$ if and only if there exists some mapping Φ from $\mathcal{X}$ to a uniform Banach space $\mathcal{W}$ such that*
$$G(x, y) = [\Phi(x), \Phi(y)]_{\mathcal{W}} \text{ for all } x, y \in \mathcal{X}. \tag{8.18}$$

Here, "uniform" means that the Banach space of functions under consideration is uniform convex and uniform Fréchet differentiable, conditions that guarantee the existence and uniqueness of a semi-inner product. Since a semi-inner product is generally non-symmetric unless the Banach space becomes a Hilbert space, similarity is non-symmetric in general – there may be x and y such that

$$G(x, y) = [\Phi(x), \Phi(y)]_{\mathcal{W}} \neq [\Phi(y), \Phi(x)]_{\mathcal{W}} = G(y, x).$$

In the case of vector-valued RKHS $\mathcal{H}$ or vector-valued RHBS $\mathcal{B}$ where the function under question is vector-valued (e.g., taking values in $\Lambda = \mathbb{C}^m$), the characterization of kernel $\mathcal{K}$ is through the existence of an operator-valued mapping $\mathcal{W}(\cdot)$ from $\mathcal{X}$ to the space of operators from the feature space $\mathcal{W}$ to Λ (that is, the space of $m \times \dim \mathcal{W}$ matrices):

$$\mathcal{K}(x, x') = \mathcal{W}(x')\mathcal{W}^{\dagger}(x), \quad x, x' \in \mathcal{X}. \tag{8.19}$$

Note that $\mathcal{W}$ is an operator-valued feature map, that is, an $m \times \dim \mathcal{W}$ matrix for each input value x. So this is the analogue to "feature map" Φ in the scalar-valued function learning, except now the "features" are operators (matrices) instead of "vectors." When $m = 1$, $\mathcal{W}$ is simply the map Φ; viewed in this way, $\mathcal{W}$ is simply a multitude of feature maps. The above representation formula says that the kernel operator $\mathcal{K}(x, x')$ is simply the multiplication of two matrices $\mathcal{W}(x')$ and $\mathcal{W}^{\dagger}(x)$. Micchelli and Pontil (2005a, b) derived this result for the vector-valued RKHS case, and Zhang and Zhang (2013) extended Equation (8.19) to the case of vector-valued RKBS. The only difference between the two cases is that † is a generalized adjoint defined in Equation (8.7) through the use of semi-inner product on $\mathcal{B}$.

8.4.1.3 Kernel Trick

Measurement of similarity of two input points, x_i and x_j, via their feature representations, $\Phi(x_i)$ and $\Phi(x_j)$, is through their inner product, or semi-inner product in general. The "kernel trick" refers to the fact that one does not need explicit

knowledge of feature representation, that is, the function form of $\Phi(\cdot)$, to compute the similarity. Rather, knowledge of the given kernel function $K(\cdot,\cdot)$ suffices. Hence, existence of the feature representation is guaranteed but not invoked in computation. It is well-known in the RKHS theory that a bivariate function makes a reproducing kernel if and only if it is representable as a similarity measurement by Equation (8.16), see Aronszajn (1950).

It should be mentioned that the kernel trick allows us to "kernelize" an input representation, so that the input space needs not be a Hilbert space of vectors, but rather an arbitrary space with a kernel function. This leads to popular algorithms such as kernel principle component analysis (kPCA), kernel density estimation, kernel regression, etc. In those algorithms, what matters is the result of projection of feature vectors $\Phi(x_i)$ and $\Phi(x_j)$ onto one another, so one can use the evaluation of the kernel function $K(x_i, x_j)$ instead, without the need to specify the function form Φ of the non-linear mapping from the input space $\mathcal{X}$ to the feature space $\mathcal{W}$.

8.4.2 Sampling via Reproducing Kernels

Kernels provide a convenient tool for sampling – they provide a means of dealing with input samples by appropriate interpolation.

8.4.2.1 Shannon Sampling

Recall that in RKHS, the sinc kernel provides a way for sampling and exact reconstruction as long as the input space and its sampling each satisfies some conditions (in terms of band-limitedness and Nyquist rate, respectively). Kernels in RKBS have similar utilities. In fact, the following Shannon sampling theorem type reconstruction formula has been established:

Theorem 8.8 (Zhang & Zhang, 2011) *Let $\mathcal{B}$ be an RKBS on $\mathcal{X}$ with the semi-inner product reproducing kernel G and let $\{x_j : j \in \mathbb{I}\}$ be a sequence of sampling points in $\mathcal{X}$. If $G(x_j,\cdot)$ and $G(x_j,\cdot)^*$ form an X_d-Riesz basis and X_d^*-Riesz basis for $\mathcal{B}$ and $\mathcal{B}^*$, respectively, then the standard reconstruction operator $S : \mathcal{B} \to \mathcal{B}$ defined by*

$$(Sf)(x) := \sum_{j\in\mathbb{I}} f(x_j)G(x_j, x)$$

is bijective and bounded. Furthermore,

$$f(x) = \sum_{j\in\mathbb{I}} f(x_j)(S^{-1}G(x_j,\cdot))(x), \quad \forall f \in \mathcal{B}, x \in \mathcal{X}.$$

8.4.2.2 Sample-based Hypothesis Space with ℓ^1 Norm

The key challenge in bounding the excess risk $\Delta(f_{\mathcal{Z},\lambda})$ arises from the fact that the solution space $\mathcal{B}^{(n)}$ used for regularized empirical risk minimization depends on training sample inputs $\{x_i\}_{i=1}^n$, and different sets of training examples will result in different subspaces $\mathcal{B}^{(n)}$ of the RKBS $\mathcal{B}$. This problem has been studied in (Shi,

Feng, & Zhou, 2011; Wu & Zhou, 2008) and is referred to as *sample dependent hypothesis space*.

Clearly, the infinite-dimensional L^1-spaces are not reflexive. However, we are not seeking to represent all the continuous linear functionals by kernel functions, but only continuous point evaluation functionals.

Suppose that $\mathcal{X}$ is a metric space. To construct a reproducing kernel, we require K to be such that for each $x \in \mathcal{X}$, $K(\cdot, x)$ belongs to $C_0(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ that vanish at infinity, and satisfy the denseness property

$$\overline{\text{span}}\{K(\cdot, x) : x \in \mathcal{X}\} = C_0(\mathcal{X}). \tag{8.20}$$

We next consider the embedding of signed Borel measures on $\mathcal{X}$ into $\mathcal{B}$ as induced by K. Specifically, we construct our space as

$$\mathcal{B} := \left\{ f_\mu := \int_X K(t, \cdot) d\mu(t) : \ \mu \text{ is a Borel measure on } \mathcal{X} \right\}$$

and define the norm of f_μ in $\mathcal{B}$ as the total variation of the corresponding measure μ – the set of functions given by Equation (8.10) lies in $\mathcal{B}$ and corresponds to μ taking discrete measures $\sum_{j=1}^n c_j \delta_{x_j}$ on a countable support in $\mathcal{X}$, whose total variation is exactly the ℓ^1-norm of the coefficient (8.11). Therefore, $\mathcal{B}$ indeed possesses the ℓ^1-norm as desired. Kernel embedding of probability measure is a hot research topic (Fukumizu, Lanckriet, & Sriperumbudur, 2011; Sriperumbudur *et al.*, 2010), although in Sriperumbudur *et al.* (2011), measures are embedded into $\mathcal{B}^*$, the dual of $\mathcal{B}$ (different from the approach discussed here).

This construction allows the introduction of the bilinear form on $\mathcal{B} \times C_0(\mathcal{X})$

$$(f_\mu, g) := \int_X g(t) d\mu(t), \ \ f_\mu \in \mathcal{B}, \ g \in C_0(\mathcal{X}).$$

Since the dual of $C_0(\mathcal{X})$ is precisely the set of Borel measures $M(\mathcal{X})$, K is the reproduce kernel of both $\mathcal{B}$ and $C_0(\mathcal{X})$ in the sense that

$$f(x) = (f, K(\cdot, x)), \quad g(x) = (K(x, \cdot), g) \ \text{ for all } f \in \mathcal{B}, \ g \in C_0(\mathcal{X}), \ x \in \mathcal{X}.$$

Song, Zhang, and Hickernell (2013) show that $\mathcal{B}$ satisfies the representer theorem if and only if

$$\|K[\mathbf{x}] K_{\mathbf{x}}(\cdot)\|_{\ell^1} \le 1 \text{ for all pairwise distinct } \mathbf{x} = \{x_j \in \mathcal{X} : j = 1, \cdots, n\}, \tag{8.21}$$

where $K[\mathbf{x}]$ is the matrix $[K(x_j, x_k) : 1 \le j, k \le n]$ and $K_{\mathbf{x}}(\cdot) := (K(\cdot, x_j) : 1 \le j \le n)^T$. It turns out that the Shepard (exponential) kernel satisfies the above relation, while the Gaussian kernel does not.

8.4.3 Kernel Method as a Unifying Feature and Similarity

Kernels provide unifying perspective on similarity-based generalization and feature-based generalization. We view similarity and feature representation as two

sides of the categorization "coin," which can be unified via an RKBS framework. Using the semi-inner product operator tool, RKBS provides a unified treatment of regularized learning to deal with similarity-based generalization, using Hilbert-space 2-norm as a regularizer, and to deal with sparsity-based feature representation and selection, using 1-norm for regularization. According to this viewpoint, a psychological stimulus will have *dual* representations: one in the input space (where processing is described by exemplars and their similarity), and the other in the feature space (where processing is described by feature-outcome associations). So the two main issues of regularized learning, kernel design and feature selection, are heavily intertwined, with the former addressed through kernel learning in multi-task setting and the latter addressed in representational learning in sparse encoding setting.

8.4.3.1 Dictionary Learning and Feature Selection

A key element in representational learning is to acquire a set of basis functions in infinite-dimensional function space. Neurophysiological evidence suggests that setting up such a proper representational basis is accomplished in brain areas for primary sensory representation, and through means of sparse encoding. For example, in Olshausen and Field (1996), a sparse representation c of an image y is generated under an internal representation $u = \{u_k \in \mathcal{W} : k \in \mathbb{I}\}$ (where $\mathbb{I}$ is an index set) via

$$\tilde{L}(u, x_j) := \min_{c \in \mathbb{R}^n} L\left(\sum_{k \in \mathbb{I}} c_k u_k, x_j\right) + \lambda \|c\|_{\ell^1}, \tag{8.22}$$

where the ℓ^1-norm $\|c\|_{\ell^1} := \sum_{k \in \mathbb{I}} |c_k|$ is employed as the regularizer. The internal representation, or "dictionary" u, on the other hand, is achieved through averaging over an ensemble of input images $\mathbf{x} = \{x^j \in \mathcal{X} : j = 1, \cdots, n\}$

$$\min_u \frac{1}{n} \sum_{j=1}^{n} \tilde{L}(u, x_j).$$

To guarantee that the dictionary $u = \{u_k : k \in \mathbb{I}\}$ is matched to the statistics of the input stimuli, the algorithm alternates between minimizing Equation (8.22) with respect to the coefficients (i.e., c_k's) given the current dictionary for an example data set, then taking a gradient descent step to update dictionary elements (i.e., u_k's, $k \in \mathbb{I}$):

$$u_l \leftarrow u_l + \epsilon \left\{ c_l \left(x_j - \sum_{k \in \mathbb{I}} u_k c_k \right) \right\} \tag{8.23}$$

where ϵ is the step-size and $\{\cdot\}$ indicates average over the current batch of data. Using a similar dual minimization scheme, LeCun's "convolution network" (Kavukcuoglu *et al.*, 2010), which has been very successfully applied to object recognition in vision, applied this unsupervised method for learning a multi-stage hierarchy of sparse features.

8.4.3.2 Maximal Margin Classifier and SVM

Once the set of basis functions of a feature space is established, classification is achieved by mapping input samples to points in the feature space – those from distinct categories are separated by a hyperplane serving as the "decision boundary." The distance from the decision boundary for each example constitutes the "margin," and it changes as one moves the placement and orientation of the separating hyperplane around. Maximal margin classifier is a learning algorithm that seeks to produce as large a margin as possible given input data, which would lower generalization error according to statistical learning theory. An example of such a classifier is the Support Vector Machine (SVM) which, through training, establishes a set of "support vectors" or training examples which contributes to the classifier.

SVM as an application of kernel methods in Hilbert space is well established. Der and Lee (2007) applied semi-inner products to the study of maximum margin classifiers in Banach spaces. Zhang, Xu, and Zhang (2009) investigated soft margin hyperplane classification

$$\inf_{w}\left\{\frac{1}{2}\|w\|_{\mathcal{W}}^2+\lambda\|\xi\|_{\ell^1}:w\in\mathcal{W},\ \xi:=(\xi_j\geq 0:1\leq j\leq n),\ b\in\mathbb{R}\right\} \tag{8.24}$$

subject to

$$y_j([\Phi(x_j),w]_{\mathcal{W}}+b)\geq 1-\xi_j,\ \ j=1,2,\ldots,n.$$

Here, Φ is a chosen feature map from $\mathcal{X}$ to the feature space $\mathcal{W}$, which is an RKBS with an semi-inner product $[\cdot,\cdot]_{\mathcal{W}}$, and λ is a fixed positive constant controlling the trade-off between margin maximization and training error minimization. The minimizer w_{opt} of Equation (8.24) was shown to belong to the closed convex cone spanned by $y_j\Phi(x_j), j=1,2,\ldots,n$, that is,

$$w_{opt}=\sum_{j=1}^{n}c_jy_j\Phi(x_j)$$

for some non-negative constant c_j's.

8.4.3.3 Kernel Learning and Vector-valued Maps

To provide a comprehensive framework for kernel learning and representational learning, the reproducing kernel Hilbert space methods can be extended to deal with learning of vector-valued functions (Micchelli & Pontil, 2005a,b), as motivated by the multi-task learning paradigm where the same input is being used for multiple objectives for inference. This vector-valued RKHS framework can be combined with the scalar-valued RKHS theory to provide a flexible framework for learning internal/feature representation (see also Micchelli & Pontil, 2007). Our work (Zhang & Zhang, 2013) has extended the vector-valued framework to the RKBS setting, which allows more explicit integration of feature and input spaces.

One such possible integration is illustrated below. We define a scalar-valued function $\tilde{f}$ on the extended input space $\tilde{\mathcal{X}} := \mathcal{X} \times \Lambda$:

$$\tilde{f}(x, u) := [f(x), u]_\Lambda$$

where $\tilde{f} \in \tilde{\mathcal{H}}$ constitutes a new function space (i.e., treating feature space as a portion of the input space). Then, applying the standard (scalar-valued) RKBS theory to $\tilde{\mathcal{H}}$, the scalar-valued semi-inner product kernel function $\tilde{G} : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}}$ for this space $\tilde{\mathcal{H}}$ can be characterized as (see corollary 3.3 in Zhang & Zhang, 2013)

$$\tilde{G}((x, u), (x', v)) = [\tilde{\Phi}(x, u), \tilde{\Phi}(x', v)]_\Lambda$$

where $\tilde{\Phi}$ is a vector-valued (as opposed to operator-valued) feature map. The representer theorem adopts the form

$$f^*_{opt} \in \overline{\text{span}} \left\{ (G(\cdot, x_j)u) : j = 1, \cdots, n,\ u \in \Lambda \right\}.$$

If we choose $\tilde{\Phi}(x, u) = \mathcal{W}(x)u$, then the optimal solution $\tilde{f}_{opt}$ satisfies $\tilde{f}_{opt}(x, u) = [f_{opt}(x), u]_\Lambda$ where f_{opt} is the solution to the vector-valued RKHS problem. The kernel

$$G(x, x') := \tilde{G}((x, f_{opt}(x)), (x', f_{opt}(x')))$$

is the optimal semi-inner product kernel for learning $\tilde{f}(x, f(x)) = \|f(x)\|_\Lambda$ with a feature map f. The vector-valued RKBS framework provides a unifying language and potential tool for addressing kernel learning and representational learning (Zhang & Zhang, 2013).

8.5 Psychological Underpinnings of Kernel Methods

8.5.1 Computational Models of Human Categorization

Category learning in cognitive psychology refers to how people acquire new concepts/categories. Typically, subjects are presented with input stimuli (a set of images, words, etc.) sequentially, classifying each and then receiving corrective feedback. A central question here concerns the mental representations people construct when acquiring new categories and storing old examples, and to perform categorical decisions. Variously mathematical/computational models have been investigated by cognitive psychologists, which fall into the following broad classes.

(i) *exemplar models*, which postulate that people judge the similarity of a test item to all remembered exemplars of each category. Medin and Schaffer's (1978) Context Model and Nosofsky's (1986) Generalized Context Model belong to this class.

(ii) *prototype models*, which postulate that people judge the similarity of a test item to a prototype of each category (Reed, 1972), where the prototype represents the average feature values of all category members (e.g., Smith & Minda, 1998).

(iii) *decision-bound models*, which postulate that people learn boundaries in stimulus space that separate categories. These were proposed by Ashby & his colleagues (Ashby & Gott, 1988; Ashby & Perrin, 1988; Maddox & Ashby, 1993).

(iv) *connectionist models*, which postulate that people learn associations between individual stimulus features and category labels with an error-driven mechanism (Gluck & Bower, 1988). In the ALCOVE model (Kruschke, 1992), associative learning is coupled with learning to selectively attend to relevant dimensions as individual exemplars are presented.

(v) *Bayesian models*, which postulate that people perform Bayesian inference based on combining prior assumptions about distributions of categories and features with feature representation of an object. Categories are represented as mixture distributions, with each component ("cluster") of this mixture given by a Gaussian distribution centered on some hypothetical stimulus. These so-called rational models of categorization was first proposed by Anderson (1991), followed up in Love, Medin, and Gureckis (2004), and popularized recently with elaborative statistical techniques by Griffiths and his associates (Austerweil & Griffiths, 2011, 2013; Sanborn, Griffiths, & Navarro, 2010).

We review each of these classes of categorization models below.

8.5.1.1 Exemplar Models

Exemplar models of categorization postulate that when a test stimulus is encountered, its similarities to the memory representation of every previously seen exemplar from each potentially relevant category are computed, and then the test stimulus is assigned to the category for which the sum of these similarities is greatest. The starting point for exemplar models is item-to-item similarity η_{ij} forming elements of a confusability matrix. The η's are non-negative numbers, typically obtained experimentally, with η_{ij} reflecting a human subject's judgment about similarity between item i and item j. With the assumption of $\eta_{ij} = \eta_{ji}$ with $\eta_{ii} = 1$, the similarity measure is made to be monotonically related to psychological distance through the multi-dimensional scaling technique (Shepard, 1957), and used in Luce's (1959) Choice Model.

Medin and Schaffer (1978), in their highly influential paper, started out by recognizing the following important characteristics of human categorization. Natural categories do not have well-defined rules or fixed boundaries separating them apart, and many natural concepts cannot be defined by a single set of critical features (Rosch, 1973). Rather, members vary in the degree to which they are judged to be good ("typical") examples of the category, and that items judged to be typical of a category possess features that are characteristic of the class but not necessary for defining the category – an example being that robin is a typical member of the category bird and has the characteristic feature that it flies, but not all birds fly, e.g., penguins. Based on these observations, Medin and Schaffer (1978) proposed (and tested) the so-called Context Theory of categorization,

in which similarity information between exemplars were used for classification judgments.

Nosofsky (1986), in his Generalized Context Model, formally linked the Context Model with Luce–Shepard conceptualization of identification/generalization to propose that the same stimulus similarity structure, revealed through the multidimensional feature representation, is used for stimulus identification and stimulus classification. Psychologically, this amounts to assuming that individuals store in their memories category exemplars which, during the categorization process, then become retrieved (not necessarily consciously) for comparison with the test stimulus to be classified. Formally, the probability of classifying a stimulus i in category J is written as

$$\text{Prob}(J|i) = \frac{b_J \sum_{j\in J} \eta_{ij}}{\sum_{K=1}^{m} (b_K \sum_{k\in K} \eta_{ik})}.$$

Here, upper-case symbols J, K refer to categories, and lower-case symbols i, j, k refer to individual exemplars in the corresponding category. The parameters b_J, b_K represent the biases in making categorization response in favor of J, K, etc.

With respect to memory retrieval process, Nosofsky and Palmeri (1997) further proposed a random-walk mechanism to describe speeded categorization: exemplars race among one another to be retrieved from memory, with rates determined by their similarity to a test item. The retrieved exemplars provide incremental information that enter into a random walk process for making classification decisions. In this way, the categorization process is linked with memory retrieval process explicitly. Earlier, Hintzman (1986) offered an exemplar model of episodic memory retrieval – the model, called MINERVA 2, was able to retrieve an abstracted prototype of the category when cued with the category name, and to retrieve category name when cued with an exemplar.

8.5.1.2 Prototype Models

Categorizing objects into psychological equivalence classes is historically thought of as prototype-based – humans supposedly average their experience with various exemplars to form a "prototype," or most typical member, of a category, compare new items to it, and render category membership judgment based on how similar the item is to this prototype. One common way of defining the prototype is as the centroid of all instances of the category in some psychological space.

Reed (1972) conducted experiments to determine how people make classifications when categories are defined by sets of exemplars and not by logical rules. Using schematic faces as stimuli and college students as subjects, it was concluded that the predominant strategy was to abstract a prototype representing each category and to compare the distance of novel patterns to each prototype, emphasizing those features which best discriminated the two categories.

In response to the emerging exemplar-based models of categorization (see Section 8.5.1.1), which were directly pitted against Reed's prototype model, Smith and Minda (1998) analyzed human performance when participants learned both

smaller, less differentiated categories and larger, more differentiated categories. They found that while the performance on former categories was better accounted for by an exemplar model, performance on the latter was better accounted for by a prototype model. Subjects' categorization strategies also differed at successive stages in learning these two types of categories: in the former case, the exemplar model dominated even early in learning, whereas in the latter case, the prototype model had a strong early advantage that gave way slowly. There seems to be a psychological transition, *from prototype-based to exemplar-based* processing during category learning. Smith and Minda (2000) further drew into question the particular experimental paradigm and category structure used to produce results that supported exemplar models.

8.5.1.3 Decision-bound Models

The decision-bound theory of categorization (mainly by Ashby and his associates) assumes that subjects partition the stimulus space into response regions in a multi-dimensional perceptual space, with these regions separated by boundaries called "decision bounds." When presented with a test stimulus, the subject determines which region the percept is in, and then emits the associated response. This class of models is closely related to the Signal Detection Theory (Green & Swets, 1966), the remarkably successful theory of human perceptual detection and discrimination.

Ashby and Gott (1988), as an application of General Recognition Theory (Ashby & Townsend, 1986), recognized that subjects often used deterministic decision rules despite considerable variations in exemplars across different attributes. Subjects' use of decision rules could be rather complex (e.g., quadratic rather than linear decision boundaries), and performed nearly optimally as prescribed by the Signal Detection Theory. Ashby and Perrin (1988) demonstrated that a perceptual stimulus can indeed be represented as a point in a multi-dimensional space, and that similarity among two stimuli is a function of the overlap of their perceptual distributions. A stochastic generalization of the static point representation of the percept to that with a multi-variate diffusion process was put forward in Ashby (2000), where the decision process was modeled as a variable (distance-to-bound) driving a univariate diffusion process with two absorbing barriers.

To compare exemplar-based models with the decision-bound model, Maddox and Ashby (1993) applied sophisticated model selection and model comparison measures, and reached the following conclusions: (i) when the exemplars from each category were normally distributed and the optimal decision bound was linear, the deterministic exemplar model and the decision-bound model provided roughly equivalent accounts of the data; (ii) when the optimal decision-bound was non-linear, the decision-bound model provided a more accurate account of the data than did exemplar models. When applied to categorization data collected by Nosofsky (1986) in which the category exemplars were not normally distributed, the decision-bound model provided excellent accounts of the data, in many cases significantly outperforming the exemplar models.

8.5.1.4 Connectionist Models

Neural networks provide a general modeling framework for achieving arbitrary input–output mapping, so naturally this framework has been used to model human categorization performance. Gluck and Bower (1988) built what is essentially the Perceptron architecture (Rosenblatt, 1957) to model categorization data. The adopted architecture had one layer of distinct input units (one per physical stimulus) feeding activation directly into one of the two output units corresponding to the output categories. The training phase involved setting the connection weights using the Rescola-Wagner (1972) rule minimizing the least mean squares (LMS) error. Because of the non-linear transfer function, this simple, two-layered network performed nonlinear discriminant analysis on input stimuli: optimal classification was achieved by predicting a criterion variable for the two categories from a set of independent variables. The network did not involve any intermediate hidden layer, so it did not need to invoke the celebrated back-propagation algorithm for error-driven learning of hidden nodes. To conclude, Gluck and Bower (1988) showed how a computationally simple architecture (namely, Perceptron) with error-driven learning rule accounted for empirical data of human categorization.

A major enhancement to the connectionist (Perception) model of category learning is Kruschke's (1992) ALCOVE (Attention Learning Covering Map) model, which incorporated both exemplar-based representation in exemplar models and error-driven learning of Perceptron. ALCOVE is a feed-forward connectionist network with three layers of nodes. Each input-layer node encodes a single psychological dimension; each hidden-layer node represents one exemplar, i.e., a particular conjunction of input dimensions in multi-dimensional stimulus space. It was assumed that the input spread by the hidden nodes in the model "cover" the entire multi-dimensional psychological space.

This three-layered connectionist network has two kinds of connection weights. The connection between the hidden layer and the output layer is essentially the two layer perceptron architecture: the output nodes are simply weighted average of hidden-layer activations, implementing the core idea of exemplar model:

$$a_k^{out} = \sum_j w_{kj} a_j^{hid}.$$

Response selection is via the Luce–Shepard choice rule using a_k as choice propensity.

The connection weights from input nodes (feature) to hidden nodes (feature conjunctions, which are viewed as exemplars), on the other hand, are modeled by Shepard's universal generalization kernel function, and these connections are gated by attentional weights (as separate input). So activation of hidden nodes can be written as

$$a_j^{hid} = \exp[-c(\sum_i \alpha_i |h_{ji} - a_i^{in}|^r)^{q/r}],$$

where a_j^{hid}, a_i^{in} are activation values of input node i (test stimulus) and of hidden node j (each j representing an exemplar), with α_i attention weights on input node

(dimension), $r = 1$ for city-block (ℓ^1) metric and $q = 1$ for exponential similarity. The hidden nodes therefore can be thought of as localized "receptive fields" in the multi-dimensional feature space. The dimensional weights α_i, which act multiplicatively on corresponding dimensions, serve to independently stretch and shrink each dimension of the input space so that across-categories stimuli are better separated and within-category stimuli are better concentrated.

Compared with standard feed-forward connectionist networks implementing parallel distributed processing (PDP), the key mechanism introduced by ALCOVE is attentional modulation of hidden nodes, which allows selective enhancement of relevant stimulus dimensions and sensitive tuning to correlated dimensions. Learning occurs as the change of connection strengths from the hidden-to-output layer, via error-driven learning, as in Gluck and Bower's implementing of Perceptron architecture. Error back-propagation is used to modify the attention weights α_i to exemplars, in addition to modifying the weights w_{ij}. Therefore, adjustment of attention weights is specific to each input dimension separately. In particular, it will allow the model to attend to relevant features and ignore irrelevant features. Looking at the time-course of the hidden layer representation will determine which dimensions of the given representation are most relevant to the task and how strongly to associate exemplars with categories.

ALCOVE has some nice computational properties. The use of exemplars as hidden nodes allows them to interact during learning, because error spreads in a non-linear fashion to connection strengths from individual feature dimensions. The interactive character in the learning is comparable to the Perceptron model, which accounts for base-rate neglect phenomena and certain forgetting effects. This interaction also implies that similar exemplars from the same category should enhance each other's learning – this further implies that prototypical exemplars should be learned faster than non-typical exemplars that lie at the category boundary, that the shape of category boundary, compared with clustering of exemplars, would have little direct influence on the difficulty of distinguishing categories.

8.5.1.5 Bayesian Models

Bayesian models of categorization address a slightly different problem, that of category learning or induction, that is, how category structures are created in the first place. While other computational models all assume *a priori* a fixed number of categories, Bayesian models focus on how clustering of objects occur from probabilistic representation of their features based on some prior distributions reflecting statistical regularities of the environment.

Anderson (1990, 1991) presented a rational model of human categorization which assumed that categorization reflected the derivation of optimal estimates of the probabilities of unseen features of objects. He performed a Bayesian analysis of what optimal estimations would be if categories formed a disjoint partitioning of the space and if features were independently displayed within a category, and proposed an incremental algorithm for calculating probabilistic classification. The prior probability a new object is classified into category k:

$$\text{Prob}(k) = \frac{cn_k}{(1-c) + c\sum_k n_k} \tag{8.25}$$

where n_k is the number of existing objects in category k, and c is a coupling constant which reflects a fixed probability that any two objects come from the same category. The prior probability that an object comes from an entirely new category not seen thus far is given by:

$$\text{Prob}(\text{new}) = \frac{1-c}{(1-c) + c\sum_k n_k}. \tag{8.26}$$

The likelihood function in Anderson's model, on the other hand, is expressed as the product of conditional probabilities through a Dirichlet process, assuming independence of features. For any feature dimension, let $c_i, i = 1, \cdots, J$ be the number of objects of a particular category showing i-indexed value (as one of the possible J values) on that feature dimension, and α_i is the prior probability of this value distribution for that category. The conditional probability that any object of category displays the i-indexed value is

$$\frac{\alpha_i + c_i}{\sum_j (\alpha_i + c_i)}.$$

Such a likelihood function is then combined with prior probability of class membership (8.25) and (8.26) to generate *a posterior* classification of a new object. Anderson's (1991) rational model was shown to demonstrate many desired properties, such as effects of central tendency of categories, extraction of basic-level categories, base-rate effects, probability matching and trial-by-trial learning. Anderson's model is an example of the Dirichlet Process Mixture Model in non-parametric Bayesian statistics.

Sanborn, Griffiths, and Navarro (2010), following Anderson (1991), further investigated efficient computation implementation of this rational scheme of probabilistic inference, using Markov chain Monte Carlo (MCMC) algorithms which approximate a probability distribution with a set of samples from that distribution. The advantage is twofold: the MCMC algorithms they investigated (both the Gibbs sampler version and the particle filters version) asymptotically approximate ideal Bayesian inference, with the quality of the approximation guaranteed. Second, they teased apart the underlying statistical model from the inference algorithm, therefore allowing these algorithms to be viewed as process models instead of just rational (normative) models. With respect to the empirical finding that people seem to shift from using a prototype representation early in training to using an exemplar representation late in training (Smith & Minda, 1998), Sanborn, Griffiths, and Navarro's (2010) model is able to explain such shift as a rational statistical inference. Overall speaking, Bayesian models cast category learning as a problem of density estimation: determining the probability distributions associated with different category labels from sequentially obtained data stream while assuming some kind of feature representation of individual objects.

To address how feature representation arise to begin with, Austerweil and Griffiths (2011, 2013) presented a computational framework for flexibly construct

feature representations for a set of observed objects. Their non-parametric Bayesian model learned a feature representation from the raw sensory data of a set of objects without specifying the number of features ahead of time. There was potentially an infinite number of features to represent the set of objects, but representations with higher number of features were penalized. Formally, the problem of feature learning reduces to a matrix factorization problem, in which the feature ownership matrix and feature-input matrix were simultaneously being sought.

Treating exemplar and prototype as two extremes, a number of models have explored possibilities of using clusters (each consisting of exemplars) to represent categories. Vanpaemel and Storms (2008) in the Varying Abstraction Model (VAM) formalized a set of interpolating models by partitioning instances of each category into clusters, where the number of clusters per category could range from 1 to the total number of exemplars belonging to that category. Love, Medin, and Gureckis (2004) in the SUSTAIN (Supervised and Unsupervised Stratified Adaptive Incremental Network) model treated the discovery of category substructure as affected not only by the structure of the world but also by the nature of the learning task and the learner's goals. Its neural network model recruited new nodes for its intermediate representations of categories whenever the current architecture failed to adequately capture the input–output pattern.

8.5.2 Unified Account of Psychological Categorization Models

The above-mentioned psychological models of categorization, while emphasizing different aspects of empirical phenomena of human categorization and category learning, reveal only partial pictures of the categorization "elephant." In this section we will review the interconnection of those psychological models of categorization, and investigate how the core ideas behind those models can be stated more precisely using terminologies from reproducing kernel methods and sparsity methods.

8.5.2.1 Relations Among Various Categorization Models

Ashby and Maddox (1993) examined several categorization models in terms of representations of stimulus and of category as well as processes for memory retrieval and for response selection. They showed some relationships between probabilistically weighted prototype model and the decision bound model, and how decision boundaries can also arise in exemplar-based models in general. That analysis, however, was restricted to categories with normally distributed attribute values.

Ashby and Alfonso-Reese (1995) showed equivalence of exemplar and of prototype models of categorization to different forms of density estimation. Identifying category similarity η_{iJ} with the probability of generating an item i from category J, and category bias B_J as prior probability of category J, led to the interpretation of exemplar model as kernel density estimation and prototype model as parametric

density estimation (of the prototype parameters), both as rational solutions to the problem of categorization.

8.5.2.2 Unified Account by Reproducing Kernels

Casting psychological models of categorization (exemplar, prototype, decision bound, connectionist, and to some extent Bayesian models) into the unifying language of kernel methods and regularized learning will further provide a coherent picture of these empirically driven modeling efforts.

Exemplar model of categorization relies on a similarity function, which is naturally conceptualized as the kernel function in machine learning (Jäkel, Schölkopf, & Wichmann, 2008a). Below, we fully elaborate the implications of such identification to human categorization models.

A test item, x, is classified according to a weighted sum of similarities to the training examples x_i: $f(x) = \sum_i c_i K(x_i, x)$. The exemplar weights, c_i, can be set to equal to the corresponding category labels, i.e. $c_i = y_i \in \{-1, 1\}$, as in Nosofsky (1986), or they can be learned to minimize classification error, as in Kruschke (1992). Reproducing kernel theory provides a key relation (8.16) between kernel similarity and feature representation under RKHS

$$\sum_i c_i K(x_i, x) = \left\langle \sum_i c_i \Phi(x_i), \Phi(x) \right\rangle. \tag{8.27}$$

Interpreting the term $\sum_i c_i \Phi(x_i) \equiv \alpha$ as a hypothetical stimulus ("prototype") in feature space, and $\langle \alpha, \Phi(x) \rangle$ as projections of the feature representation of stimulus x onto this prototype feature vector, then the exemplar model, which uses similarity in the input space ($\mathcal{X}$), becomes equivalent to a weighted prototype model represented by the prototype feature vector α in the feature space. Alternatively, we can view α as a vector of association weights, where each vector component represents the weighting of feature dimensions, and turn the exemplar model into a perceptron model, or to a linear decision-bound model. In particular, the maximal margin classifier, in which class boundary is sought after to separate the two classes of input stimuli as far apart as possible, bears resemblance to the decision-bound model of categorization.

For non-parametric Bayesian models of categorization, we can view this approach as similar to seeking sparsity in the vector c of exemplar weights, and hence bearing some resemblance to sparse SVM with hinge loss (where sparsity of support vectors results). However, the support vectors usually lie at the boundaries of the separating hyperplanes – they are hard-to-classify exemplars or "outliers." This is in sharp contrast to cluster models which seek to identify a small set of exemplars sufficient for representing the category structure.

Note that in the RKBS framework, the relationship generalizing Equation (8.27)

$$\sum_i c_i G(x_i, x) = \left[\sum_i c_i \Phi(x_i), \Phi(x) \right] = [\alpha, \Phi(x)]$$

still holds, due to linearity of semi-inner product with respect to the first argument. We can thus define $\alpha = \sum_i c_i \Phi(x_i)$, where exemplar weights c_i's are task-dependent and feature representation $\Phi(x_i)$ is task-independent. The α-vector defined above represents the feature dimension weights for the prototype. The interaction between exemplar weight c-vector and feature weight α-vector will be of core importance. Our analysis shows that the equivalence between the exemplar model and the prototype model is valid not only when perceptual "dimensions" are integral (Hilbert space), but also where feature representation is explicitly constructed (for ℓ^1-space, a Banach space).

8.5.2.3 Shepard Kernel and Multi-dimensional Input Scaling

Similarity plays a central role in psychological models of category learning, as it interacts with attention and generalization. In his seminal work on generalization, Shepard (1957, 1987) argued that generalization between stimuli is a direct function of their perceived similarity. There is a long tradition in multi-dimensional scaling (MDS) of modeling similarity as a decreasing function of distance in some metric stimulus space. For instance, the dominant assumption is that pair-wise similarity $K(x, x')$ between two stimuli x, x' is given by

$$K(x, x') = e^{-d_p(x,x')^p} \tag{8.28}$$

where d_p is the Minkowski (ℓ^p) metric over the input vector space $\mathcal{X}$

$$d_p(x, x') = (\sum_j |x_j - x'_j|^p)^{1/p} \tag{8.29}$$

with $p \geq 1$ and j ranges over the dimensions of the stimulus space, $\mathcal{X}$.

Shepard (1987) offered a normative justification for $p = 1$ (exponential kernel) based on Bayesian inference, under a model in which meaningful outcomes are associated with contiguous *consequential regions* of the stimulus space. Others have argued for $p = 2$ (Gaussian kernel), based on human performance in categorization tasks (Nosofsky, 1986). This view of psychological similarity maps naturally onto kernel methods, and indeed several psychological models founded on similarity, including models of categorization, can be formally recast as kernel methods (Jäkel, Schölkopf, & Wichmann, 2008a,b, 2009).

The standard non-metric MDS framework models stimulus similarity as a non-increasing function of the symmetric proximity data $\{K(x_i, x_j) \equiv K_{ij}, i, j = 1, \cdots, n\}$ on pairwise comparison in stimulus space $\mathcal{X}$. In traditional studies, proximity data (K_{ij} as a matrix) are always assumed to be symmetric or are symmetrized before being fed as inputs to MDS. This is because MDS seeks to embed the set of n input stimuli *metrically* into some feature space such that proximity is monotonically related to pairwise distance. Of course, when embedded into a Hilbert space, proximity is also monotone related to the inner product (and hence the kernel function). MDS has been tackled in contemporary machine learning as a kind of manifold learning, such as Local Linear Embedding (LLE) (Roweis &

Saul, 2000), Isomap (Tenenbaum, de Silva, & Langford, 2000), and Laplacian Eigenmaps (Belkin & Niyogi, 2003).

With respect to the Shepard kernel (8.28) with Minkowski metric (8.29), a consistent empirical finding is that separable psychological dimensions are best fit by $p = 1$ or sometimes $p < 1$, whereas integral psychological dimensions are fit by p closer to 2 (Garner, 1974; Shepard, 1964). Various proposals have been made for how learning transforms a stimulus space from having an integral representation ($p = 2$, Gaussian kernel) to a separable representation ($p = 1$, exponential kernel), where categories tend to be aligned with a consistent set of axes in the stimulus domain.

8.5.2.4 Incorporating Attention into Kernel Methods

Attention modulation of psychological similarity is a well-established principle in cognition. The pattern of pairwise similarities perceived among a set of stimuli depends on what aspects of the stimuli an observer considers currently relevant, which in turn depends on both the task and the set of stimuli present (Goldstone, 1994; Medin, Goldstone, & Gentner, 1993; Shepard, 1964; Tversky, 1977). Generalization is stronger between stimuli differing on task-irrelevant dimensions than between stimuli differing on task-relevant dimensions (Jones, Maddox, & Love, 2005; Nosofsky, 1986).

Attention might affect categorization by (i) influencing the sampling of input stimulus, through imposing weighting on individual exemplars on $\mathcal{X}$ (input filtering), or (ii) highlighting feature dimensions in the feature space, through modifying the shape of the kernel. Mathematically, (i) corresponds to controlling the exemplar weights c_i and hence imposing a measure on $\mathcal{X}$, while (ii) corresponds to parameterizing the family of kernels or equivalently the family of semi-inner products of a common basis or frame (e.g., the power exponent p in the semi-inner product kernel which, in the $p = 1$ case, is simply the projection operation).

Attentional effects on perceived similarity have been modeled by assuming learnable weights for each dimension (Nosofsky, 1986). In the language of kernels, it is given by the ℓ^p metric (8.28) where $p = 1$ or 2:

$$K(x, x'; \alpha) = \begin{cases} \exp\left(-\sum_j \alpha_j |x_j - x'_j|\right) & \text{Shepard (exponential) kernel} \\ \exp\left(-\sum_j \alpha_j (x_j - x'_j)^2\right) & \text{Gaussian kernel} \end{cases} \tag{8.30}$$

where j indexes the dimension of the input space $\mathcal{X}$, and α_j's are adjustable attention weights. By committing to the dimensions of the stimulus space and only leaving open their relative scaling, attention to the feature dimension serves to limit estimation error in kernel learning. In the popular ALCOVE model (Kruschke, 1992), attention weights are learned by gradient descent on a loss function.

Another conceptualizing of attention is that it controls the saliency of stimulus features, thereby influencing their learning rates (Kruschke, 2001; Mackintosh,

1975; Rescorla & Wagner, 1972). For instance, modifying the Perceptron learning rule (Rosenblatt, 1958), theories of cue associability have proposed that the change in association weights $w = [w_1, \cdots, w_m]^T$ follows gradient descent on loss $L_{\mathcal{Z}}$ over training data $\mathcal{Z}$

$$w_j \leftarrow w_j + \epsilon\, \alpha_j \frac{\partial}{\partial w_j} L_{\mathcal{Z}}(w)$$

where j is the feature dimension and α_j is the attention allocated to it. Associations from attended features are thus learned faster than associations from less-attended features. This role of attention can be shown to be equivalent to modifying the kernel by reweighting the dimensions of feature space (Matt Jones, personal communication).

Recently, stochastic techniques have been introduced to analyze ℓ^1-regularized loss minimization (Shalev-Shwartz & Tewari, 2011). The stochastic coordinate descent method updates the weight of a single feature at each iteration; this corresponds to selective attention to feature dimensions. The stochastic gradient descent (and its generalization, stochastic mirror descent, see Srebro, Sridharan, & Tewari, 2011) method, on the other hand, involves picking randomly one example from (a uniformly distributed) training set at each iteration, and updating the exemplar weight vector based on the chosen example; this corresponds to selective attention to exemplars. These techniques are expected to generate insights for addressing the issue of attentional control in category induction (attention to feature versus attention to exemplar).

8.5.3 Challenges for Kernel Methods

Despite the conceptual advantages provided by the reproducing kernel method, there are still challenges in finding an efficient, effective, and well-understood algorithm for categorization (in contrast to state-of-the-art reinforcement learning algorithms). The key ingredient lacking is a theory for deriving effective feature representation – there are certain representational mechanisms used for human categorization that have yet to be captured by any machine learning algorithm that approaches human inductive capacity.

8.5.3.1 Similarity vs. Dissimilarity

Psychological similarity was known to be incompatible with the metric axioms, in that it might be non-symmetric and violate triangle inequality (Tversky, 1977; Tversky & Gati, 1982). The general pattern observed regarding non-symmetry has been that x is rated as more similar to y than vice versa if y is more salient or has more features than does x. MDS models can account for at least some of these findings by incorporating stimulus- or response-bias parameters (Nosofsky, 1991), but non-symmetric similarity may also reflect stimulus representations themselves, as well as the process of comparison (Polk *et al.*, 2002). The RKBS methods avoid the symmetry requirement for its kernels, so in principle non-symmetry can be

accommodated more readily, via the semi-inner product operation. However, the metric of a Banach space is still symmetric, and satisfies the triangular inequality. So it appears that semi-inner product and metric/norm capture different aspects of the geometry, although from the former the latter is derivable. In machine learning, one may simultaneously consider both "similarity" and "dissimilarity" (as in von Luxburg, 2004, for instance); in the present context, they map to, respectively, norm-induced metric structure and semi-inner product structure.

Psychological similarity is notoriously labile, depending on both the stimulus set and the task (Tversky, 1977). Much of the flexibility of human cognition arguably lies in the fact that the features underlying similarity vary in an adaptive way from task to task (Medin, Goldstone, & Gentner, 1993). This view has led to claims that it is the choice of features, and not similarity itself, that determines behavior (Goodman, 1972). Nevertheless, similarity may play an important role in cognitive processing, as providing a compact (uni-dimensional) indication of how well two stimuli match across a large number of features (Goldstone, 1994). From this point of view, similarity can be viewed as the mind's implementation of the kernel trick, to perform learning and inference directly on the input space (i.e., perceptual representation), bypassing the often high-dimensional feature space. Flexibility of similarity then becomes flexibility in the choice of the kernel, K, or equivalently, of the feature map, Φ.

8.5.3.2 Flexible Feature Representation

There is also evidence that the dimensions of the stimulus space themselves can change, perhaps over longer time-scales.

To begin with, there is a long-recognized distinction between perceptually integral dimensions (2-norm) versus perceptually separable dimensions (1-norm) in the multi-dimensional psychological space where stimuli are represented (Garner, 1974). Moreover, perceptually integral dimensions ($p = 2$) can become separable ($p = 1$), either during the course of development or (perhaps temporarily) as a consequence of classification learning in adulthood. Goldstone and Steyvers (2001) investigated the process of dimension differentiation, whereby subjects learned a separable representation of previously integral dimensions. This change of representation entailed a qualitative shift of similarity, which affected the learnability of future categorization tasks. Jones and Goldstone (2013) showed that learning a categorization task with integral-dimension stimuli could induce an analytic coordinate frame (i.e., a set of orthogonal axes) for the stimulus space such that, when subjects were transferred to a new task, transfer performance was better if the new category boundary was aligned with this coordinate frame. When stimuli vary along integral dimensions and their distribution (within some task domain) had a non-spherical covariance structure (i.e., weighted l^2-norm), the dimension of greatest variation (i.e., the first principle component) was shown to become perceptually highlighted. Austerweil and Griffiths (2013) reported that new stimulus dimensions could emerge after the discovery of components useful in predicting some outcome variable. These examples illustrate how the perceptual

system (mediating similarity judgments) adapts to suit the needs of the cognitive system. How such mechanisms of adaptation lead to modifying kernels to improve learning has yet to be worked out.

8.6 Summary: Unifying Feature and Similarity by RKBS

Our chapter reviews the basic concepts of reproducing kernel Banach Space (RKBS) framework as potentially unifying similarity and feature representation, the two sides of the categorization "coin," both from the mathematical perspective and from the cognitive perspective. Mathematically, RKBS will provide a unified treatment of regularized learning under both the smoothness setting (L^2 or Hilbert space) and the sparsity setting (ℓ^1 space), through the semi-inner product tool. The connection and interaction between similarity-based generalization (under 2-norm) and the sparsity-based feature selection (under 1-norm) can now be addressed under a common framework. representer theorem continues to work so long as the regularizer is norm-based and the loss function is proper (i.e., it is non-negative, with its minimum zero achieved when $y_i = f(x_i), \forall i$). More importantly, the kernel function still depends on the norm used in the regularizer and not on the sample points. Because the Banach space approach includes the 2-norm (Hilbert space) and 1-norm as special cases, this approach has the potential in bringing fundamental breakthrough to learning, sampling, approximation, and feature representation after synthesizing kernel-based method with sparsity-based methods. Translating extant psychological models of categorization, e.g., the Exemplar Model, the Prototype Model, the Decision-Bound Model, the Connectionist Model, and the Cluster Model, in this universal framework of kernel methods allows a more comprehensive view of psychological processes underlying categorization and category learning. A psychological stimulus may have dual representations, one in the input space (where processing is described by exemplars and their similarities) and the other in the feature space (where processing is described by feature–outcome associations). The Shepard's Universal Law of Generalization plays the crucial role of a "kernel." The RKBS framework using the semi-inner product will bring together similarity-based cognitive mechanisms with feature-/attention-based cognitive mechanisms for categorization, and provide a unified account of objects with dual representations, respectively, in input space and in feature space. It may well be the case that the above-mentioned psychological models merely reflect different views about whether the classification algorithm (and attention modulation) is operating on the input space or the feature space, rather than contain computationally meaningful differences in the nature of category representations (as psychologists have debated over decades)! The real questions, then, become (i) the nature of the stimulus or feature representation, and (ii) the learning algorithms that find or approximate the optimal classifier (e.g., by updating exemplar and/or feature weights). A regularized learning framework thus provides the potential for a unification of all of these theoretical considerations,

facilitating comparison and enabling new, more principled psychological models that synthesize the insights offered by each.

Acknowledgement

The authors acknowledge research supports from ARO Grant No. W911NF121-0163 and AFOSR Grant No. FA9550-13-1-0025 to Jun Zhang and NSFC Grant No. 11101438 and No. 11222103 to Haizhang Zhang. We thank Matt Jones for inputs on human categorization literature and for critical reading of an earlier draft.

References

Anderson, J. R. (1990). *Cognitive psychology and its implications*. WH Freeman/Times Books/Henry Holt & Co.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Argyriou, A., Micchelli, C. A., & Pontil, M. (2005). Learning convex combinations of continuously parameterized basic kernels. Proceedings of the 18th Annual Conference on Learning Theory (COLT 2005). Bertinoro, Italy.

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

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, *44*, 310–329.

Ashby, F. G. & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G. & Gott, R. E. (1988). Decision rules in the perception and categorization of multi-dimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53.

Ashby, F. G. & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Ashby, F. G & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–150.

Ashby, F. G. & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, *63*, 173209.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*, 817–851.

Belkin, M. & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*, 1373–1396.

Candès, E. J. & Tao, T. (2004). Near-optimal signal recovery from random projections: universal encoding strategies. *IEEE Transactions on Information Theory*, *52*, 5406–5425.

Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*, 489–509.

Casazza, P., Christensen, O., & Stoeva, D. T. (2005). Frame expansions in separable Banach spaces. *Journal of Mathematical Analysis and Applications*, *307*, 710–723.

Coifman, R. R. & Maggioni, M. (2006). Diffusion wavelets. *Applied Computational Harmonics Analysis*, *21*, 53–94.

Cucker, F. & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society (New Series)*, *39*, 1–49.

Der, R. & Lee. D. (2007). Large-margin classification in Banach spaces. *AISTATS*, *2*, 91–98.

Evgeniou, T., Micchelli, C. A., & Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, *6*, 615–637.

Foucart, S. & Lai, M.-J. (2009). Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied Computational Harmonics Analysis*, *26*, 395–407.

Fukumizu, K., Lanckriet, G. R. G., & Sriperumbudur, B. (2011). Learning in Hilbert vs. Banach Spaces: A measure embedding viewpoint. Advances in Neural Information Processing Systems, *25*, 1773–1781.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum Associates.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*, 404–409.

Giles, J. R. (1967). Classes of semi-inner-product spaces. *Transactions of the American Mathematical Society*, *129*, 436–446.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*, 116–139.

Goodman, N. (1972). Seven structures on similarity. In N. Goodman (ed.), *Problems and projects*. New York, NY: Bobbs-Merrill.

Green, D. M. & Swets, J. A. (1966). Signal detection theory and Psychophysics. Vol. 1. New York, NY: Wiley.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008a). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review*, *15*, 256–271.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008b). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*, 297–303.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*, 381–388.

James, R. C. (1964). Characterizations of reflexivity. *Studia Mathematica*, *23*, 205–216.

Jones, M., & Goldstone, R. L. (2013). The structure of integral dimensions: Contrasting topological and Cartesian representations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 111–132.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. Proceedings of the 27th Annual Meeting of the Cognitive Science Society (pp. 1066–1071).

Kavukcuoglu, K., Sermanet, P., Boureau, Y-L., Gregor, K., Mathieu, M., & LeCun, Y. (2010). Learning convolutional feature hierachies for visual recognition. *Advances in Neural Information Processing Systems*, *23*, 1090–1098.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., & Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72.

Lavery, J. E. (2000). Shape-preserving, multiscale fitting of univariate data by cubic L_1 smoothing splines. *Computer Aided Geometric Design*, *17*, 715–727.

Li, S. (1995). On general frame decompositions. *Numerical Functional Analysis and Optimization*, *16*, 1181–1191.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: Wiley.

Lumer, G. (1961). Semi-inner-product spaces. *Transactions of the American Mathematical Society*, *100*, 29–43.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*, 49–70.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Micchelli, C. A. & Pontil, M. (2005a). Learning the kernel function via regularization. *Journal of Machine Learning Research*, *6*, 1099–1125.

Micchelli, C. A. & Pontil, M. (2005b). On learning vector-valued functions. *Neural Computation*, *17*, 177204.

Micchelli, C. A., & Pontil, M. (2007). Feature space perspectives for learning the kernel. *Machine Learning*, *66*, 297–319.

Nath, B. (1971/72). On a generalization of semi-inner product spaces. *Mathematical Journal of Okinawa University*, *15*, 1–6.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, *14*, 849–856.

Noles, N. S. & Gelman, S. A. (2012). Effects of categorical labels on similarity judgments: A critical analysis of similarity-based approaches. *Developmental Psychology*, *48*, 890–896.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*, 94–140.

Nosofsky, R. M. & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Olshausen, B. A. & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, *7*, 333–339.

Poggio, T. & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the AMS*, *50*, 537–544.

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. *Cognition*, *82*(3), B75–B88.

Rao, M. M. & Ren, Z. D. (1991). *Theory of Orlicz spaces*. New York, NY: Marcecl Dekker.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Roweis, S. & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Schölkopf, B. & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. Cambridge, MA: The MIT Press.

Shalev-Shwartz, S. & Tewari, A. (2011). Stochastic methods for ℓ_1-regularized loss minimization. *Journal of Machine Learning Research*, *12*, 1865–1892.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shi, L., Feng, Y.-L., & Zhou, D.-X. (2011). Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, *31*, 286–302.

Smale, S. & Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, *26*, 153–172.

Smale, S., Rosasco, L., Bouvrie, J., Caponnetto, A., & Poggio, T. (2010). Mathematics of neural response. *Foundations in Computational Mathematics*, *10*, 67–91.

Smith, E. E. (1995). Concepts and categorization. In E. E. Smith & O. Daniel (eds.), *Thinking*, Vol. 3 (pp. 3–33) Cambridge, MA: MIT Press.

Smith, J. D. & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Smith, J. D. & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.

Song, G., Zhang, H., & Hickernell, F. J. (2013). Reproducing kernel Banach spaces with the ℓ^1 norm. *Applied and Computational Harmonic Analysis*, *34*, 96–116.

Srebro, N., Sridharan, K., & Tewari, A. (2011). On the universality of online mirror descent. *Advances in Neural Information Processing Systems*, *24*, 2645–2653.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Scholkopf, B., & Lanckriet, G.R.G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, *11*, 1517–1561.

Sriperumbudur, B., Fukumizu, K., & Lanckriet, G.R.G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, *12*, 2389–2410.

Sutherland, N. S. & Mackintosh, N. J. (1971). *Mechanisms of Animal Discrimination Learning*, NY: Academic Press.

Tenenbaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for non-linear dimensionality reduction. *Science*, *290*, 2319–2323.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, *58*, 267–288.

Tong, H., Chen, D.-R. & Yang, F. (2010). Least square regression with ℓ^p-coefficient regularization. *Neural Computation*, *22*, 3221–3235.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Tversky, A. & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*, 123–154.

Vanpaemel, W. & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin and Review*, *15*, 732–749.

von Luxburg, U. (2004). Statistical learning with similarity and dissimilarity functions. Max Planck Institute for biological cybernetics, Tübingen, Germany. PhD thesis.

Wu, Q. & Zhou, D.-X. (2008). Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, *56*, 2896–2907.

Weiss, Y. (1999). Segmentation using eigenvectors: a unifying view. In Proceedings IEEE International Conference on Computer Vision, pp. 975–982.

Xu, Z., Zhang, H., Wang, Y., Chang, X., & Liang, Y (2010). $L^{1/2}$ regularization. *SCIENCE China Information Sciences*, *53*, 1159–1169.

Zhang, H. & Zhang, J. (2010). Generalized semi-inner products with applications to regularized learning. *Journal of Mathematical Analysis and Applications*, *372*, 181–196.

Zhang, H. & Zhang, J. (2011). Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products. *Applied and Computational Harmonic Analysis*, *31*, 1–25.

Zhang, H. & Zhang, J. (2012). Regularized learning in Banach spaces as an optimization problem: representer theorems. *Journal of Global Optimization*, *54*, 235–250.

Zhang, H. & Zhang, J. (2013). Vector-valued reproducing kernel Banach spaces with applications to multi-task learning. *Journal of Complexity*, *29*, 195–215.

Zhang, H., Xu, Y., & Zhang, J. (2009). Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, *10*, 2741–2775.