Duality Between Feature and Similarity Models, Based on the Reproducing-Kernel Hilbert Space

Matt Jones (University of Colorado) and Jun Zhang (University of Michigan)

December 22, 2016

1 Introduction

There are two longstanding theoretical approaches to learning and concept representation, one based on features and one based on similarity.

The feature approach has its roots in associative learning (e.g., Pavlov, 1927) and the idea that learning involves acquiring associations from individual cues to consequential outcomes or responses. Formal models of this learning process assume that any stimulus is decomposable into a set of features, and that learning involves adjusting the associative strength of each feature (Estes, 1950; Rosenblatt, 1958). More modern versions of these models posit learning rules whereby features compete to make predictions (Rescorla & Wagner, 1972), a mechanism that famously explains blocking (Kamin, 1968) and other forms of cue competition in category learning (Gluck & Bower, 1988).

The similarity approach makes a radically different assumption, that learning is centered on whole stimuli rather than on substituent features. This view has its roots in experimental work on stimulus generalization (Guttman & Kalish, 1956) and the idea that learning can be characterized by the degree to which knowledge acquired about one stimulus is transferred or generalized to another. Formal models of similarity-based learning assume that responses to new stimuli are determined by direct generalization from previously encountered stimuli (Nosofsky, 1986; Shepard, 1987). Learning per se is based on acquisition of new exemplars (Nosofsky, 1986) or on learning configural associations from whole stimuli to outcomes (Kruschke, 1992; Pearce, 1987, 1994). Similarity-based models have shown great success in predicting transfer performance on novel stimuli (Nosofsky, 1992), particularly in cases where behavior is not well captured by an additive function of stimulus features (Medin & Schaffer, 1978; Pearce & Bouton, 2001).

The debate between feature- and similarity-based learning has a long history in both animal conditioning, as a debate between elemental and configural models, and in human category learning, as a debate between examplar models and prototype or connectionist models. The debate seems to get at fundamental questions regarding the nature of knowledge and psychological representation. Are conceptual and causual knowledge encoded in associations among individual features (cues), or in memory for specific instances (exemplars)? Is a stimulus fundamentally represented in an analytic fashion, in terms of its values on various features, or holistically, in terms of its similarity to other stimuli? It seems natural to believe that one of these views must be more correct than the other, and that clever experimental and computation research should eventually be able to determine which set of principles governs the brain's operation.

However, recent theoretical developments in machine learning suggest that the difference between feature and similarity models is not as fundamental as psychologists have believed. Instead, the principles of these two approaches turn out to be compatible, via a deep mathematical connection. This connection is embodied in the *kernel framework*, a mathematical system that has been used to formulate a variety of powerful machine-learning algorithms for classification and prediction (Schölkopf & Smola, 2002). Under the kernel framework, simple linear algorithms, such as linear regression or the support vector machine, are made much more flexible by first projecting the stimulus set into a high (usually infinite) dimensional vector space, and then applying the algorithm in that new space (Shawe-Taylor & Christianini, 2004). The vector space is called a reproducing-kernel Hilbert space (RKHS), and it is generated by a kernel function that assigns a real number to every pair of stimuli in the original space. In principle, learning and optimization are done in the RKHS, which is what gives kernel methods their great flexibility. However, in practice all necessary

calculations can be re-expressed in terms of the kernel function in the original (simpler) stimulus space. This equivalence, known as the kernel trick, allows for learning algorithms that are simultaneously flexible and computationally efficient, as well as having good convergence properties.

The proposal explored in the present paper is that the brain exploits the computational power of the kernel trick, to achieve the advantages of similarity and feature strategies simultaneously. In psychological terms, the kernel and the RKHS can be thought of as separate similarity- and feature-based representations of the same stimuli. That is, the kernel can be interpreted as a similarity function, $\mathrm{sim}\,(x, x')$, and the the RKHS can be equated with a (possibly infinite) set of feature dimensions, $\mathcal{F}$. For every stimulus x in the original space, there is a corresponding vector $\mathbf{x} = (\mathbf{x}_i)_{i \in \mathcal{F}}$ in the RKHS, and the similarity between any two stimuli equals the inner product (or dot product) between their feature representations:

$$\mathrm{sim}\,(x, x') = \sum_{i \in \mathcal{F}} \mathbf{x}_i \mathbf{x}'_i. \tag{1}$$

Fundamental mathematical results from functional analysis show that for any similarity function obeying certain criteria (given below), there always exists a feature representation satisfying Equation 1, and vice versa (Mercer, 1909).

The fact that similarity and feature representations can be constructed satisfying Equation 1 has been shown to imply equivalence between psychological models of learning that had been previously considered antithetical. Jäkel, Schölkopf, and Wichmann (2007) showed that exemplar models of categorization are equivalent to prototype models, if sim defines the exemplar similarity function and prototypes are defined with respect to the features in $\mathcal{F}$. Likewise, Jäkel, Schölkopf, and Wichmann (2009) showed that exemplar models are equivalent to perceptrons (i.e., one-layer neural networks) when the perceptron encodes stimuli by their values on the features in $\mathcal{F}$. Ghirlanda (2015) proved analogous results in the domain of conditioning, namely that elemental models of conditioning, which learn associations from individual features (e.g., Rescorla & Wagner, 1972), are equivalent to configural models, which learn associations from whole exemplars and generalize according to similarity (e.g., Pearce, 1987, 1994). Again, the equivalence holds if sim defines the configural model's similarity function and $\mathcal{F}$ defines the features for the elemental model.

It is important to recognize that, in these cases of equivalent models, the feature representation usually is entirely different from the representation used to derive the similarity function. That is, most similarity-based models assume that the similarity function arises from some underlying stimulus representation together with some theory of similarity, rather than directly positing the similarity function as a primitive. For example, Shepard's (1957, 1987) influential model of similarity and generalization holds that stimuli are represented in a multidimensional Cartesian space, $x = (x_1, \ldots, x_m)$ and that similarity is an exponential function of distance in that space: $\mathrm{sim}\,(x, y) = \exp\left(-\sum |x_j - y_j|\right)$. Likewise, Tversky's (1977) contrast model of similarity holds that stimuli are represented by binary attributes, $x = (x_1, \ldots, x_m)$ with $x_j \in \{0, 1\}$, and similarity is determined by the shared and distinctive attributes of two stimuli. In both cases, the dimensions or attributes $j \in \{1, \ldots, m\}$ that underlie the similarity representation (x) are entirely different from the features $i \in \mathcal{F}$ that constitute the feature representation ($\mathbf{x}$) in Equation 1. The important point here is that, regardless of where a model's similarity function comes from, if that function is related to some feature set $\mathcal{F}$ according to Equation 1, then the set of concepts or behaviors learnable by the similarity model is identical to that learnable by the feature model.

The equivalence between similarity and feature models is an example of the classic tradeoff in psychological modeling between representation and process. In this case, similarity processing on one representation is equivalent to feature processing on a different representation. This finding might be taken as a major problem for psychological learning theory, in particular as a problem of model unidentifiability. The equivalence means that one cannot determine—even in principle—whether the brain learns by feature associations or similarity-based generalization from exemplars. However, we suggest that this is the wrong question. In fact, the equivalence between similarity and feature models implies the question of which is more correct is meaningless. Instead, we propose a view of *duality*, whereby feature and similarity models are viewed as two descriptions of the same system. We recall that models are just that: formal (mathematical or computational) systems that researchers create as a means to describe and understand the modeled system (the mind or brain). When two models are formally equivalent, they can still differ in the insights they afford, much like viewing some complex geometrical or architectural structure from different angles. Following this

viewpoint, we refer to Equation 1 as the *kernel duality*, and we show in this paper how it can yield a deeper understanding of human and animal learning than can feature or similarity models taken separately.

The kernel trick is celebrated in machine learning because it enables one to obtain the best of both worlds: the intuitiveness and optimality properties of linear (feature) methods, and the computational efficiency and flexibility of nonlinear (similarity) methods. Here we advocate an analogous approach in cognitive modeling. Research founded on feature and similarity models has produced a variety of theoretical principles, concerning how each of these learning strategies can explain complex empirical phenomena as well as how they can confer advantages from a normative standpoint. Furthermore, within both of these model classes there are competing models that differ in the details of how knowledge is updated after each learning event, and in the contributions of other mechanisms such as attention. In the present paper we follow the model translation approach of Jones and Dzhafarov (2014), in which the insights, principles, and specific hypotheses from one modeling framework are reinterpreted in another, by exploiting their formal equivalence. We argue that the translation approach enables recognition of deeper connections than possible within one framework alone, integration of ideas that previously appeared incommensurable, and development of more powerful and encompassing theories. Thus we aim to move past the feature-similarity debate to a new view in which these two perspectives mutually inform each other.

The remainder of this paper is organized as follows. Section 2gives formal definitions of the kernel framework and of similarity and feature-based learning models. In Section 2, we formally define feature and similarity models of associative learning and show how, when their representations are related by Equation 1, the set of concepts or output functions they can learn is identical. We then introduce the mathematical framework of kernel methods from machine learning and explain how the equivalence between similarity and feature models fits within that framework, and show how the mathematical tools from the kernel framework can be used to construct feature models that are dual to similarity models and vice versa. In Section 3, we develop the dual-representation model in detail. We start with a similarity model with similarity defined by a generalization gradient in a multidimensional stimulus space, and we derive a dual feature representation based on an infinite family of features all defined by a shared shape of receptive field but at various scales (resolutions) and locations throughout the stimulus space. We then demonstrate connections between this *continuous-features model* and neural population coding, theories of separable versus integral dimensions, and Bayesian models of concept learning. In Section 4, we apply the idea of the kernel duality to theories of attention in learning. We show how two fundamentally different theories of attention, one grounded in feature processing and the other in similarity, can be seen to implement exactly the same mechanism but applied to different (dual) representations. We then use the kernel duality to translate each theory of attention into the opposite modeling language, expressing feature-based attention as a change in similarity and similarity-based attention as a change in features, and show how this translation yields new insights into the implications of these theories. Finally, Section 5 discusses other potential applications of the duality approach, including translation of learning rules between similarity and feature models, and modeling of asymmetric similarity using recent results extending the RKHS construction from Hilbert space to the more general Banach space (Zhang & Zhang, 2012). To some extent, the results reported in this paper represent the low-hanging fruit offered by the kernel duality. There is much more work to be done to develop a complete theory that integrates the similarity and feature approaches. Our primary aim with this initial work is to demonstrate the potential of the duality perspective, both in learning theory and elsewhere in cognitive science.

2 Similarity, Features, and the Kernel Framework

2.1 Equivalence of Similarity and Feature Models

The psychological models falling under the scope of this paper all involve learning to estimate values for stimuli. These values could represent outcome probabilities in a classical conditioning task, category-membership probabilities in a categorization task, or action tendencies or expected rewards in an operant conditioning (reinforcement learning) task. In tasks with multiple possible outcomes or actions, such as categorization with multiple categories, a response rule is also needed to convert the estimated values for the different options into choice probabilities, such as the Luce choice (Luce, 1963) or softmax rule (Sutton & Barto, 1998). However, we set the question of response rules aside for now, because it is independent of the learning

model per se (i.e., any learning model can be combined with any response rule). Thus for present purposes we can restrict to a single outcome and think of any learning model as working to learn a mapping, v, from the space of possible stimuli to the space of real numbers ($\mathbb{R}$).

Similarity models calculate their output values from weighted similarity to stored exemplars. The model assumes a similarity function with $\text{sim}(x, x') \in \mathbb{R}$ for every pair of stimuli x and x' (including $x = x'$). The model maintains a set of exemplars, $\{x^1, \ldots, x^n\}$, either determined in advance or able to grow with experience, and learning involves updating a weight for each exemplar, $\{c^1, \ldots, c^n\}$.[1] In tasks with multiple outcomes or actions, there would be a separate set of exemplar weights for each option, but we omit this from our notation for simplicity. For example, in the generalized context model of categorization (GCM; Nosofsky, 1986), each exemplar has a weight of 1 if it was observed in the category in question, and 0 otherwise. In the attention-learning covering map (ALCOVE; Kruschke, 1992), the weights are continuous-valued and are updated by gradient descent after each trial. In Pearce's (1987, 1994) configural models of conditioning, the weights are continuous but after each trial only the weight for the presented stimulus is updated. Despite these differences in learning rules, the set of value functions these models can in principle represent is the same.[2] At any point in learning, any similarity model has some finite set of exemplars and associated weights. Thus the model's output for any possible stimulus x on the next trial is equal to

$$v_{\text{s}}(x) = \sum_{i=1}^{n} c^i \cdot \text{sim}\left(x^i, x\right). \tag{2}$$

Therefore the set of value functions (v_{s}) learnable in principle by a similarity model is the set of all functions of the form in Equation 2, for any natural number n, any set of stimuli $\{x^1, \ldots, x^n\}$ as the exemplars, and any real numbers $\{c^1, \ldots, c^n\}$ as the weights.

Feature models represent each stimulus as a vector of feature values, $\mathbf{x}$. The model learns a weight vector, $\mathbf{w}$, consisting of a weight for each feature. At any point in learning the model's output for a stimulus $\mathbf{x}$ is equal to the inner product of that stimulus with the current weight vector:

$$v_{\text{f}}(\mathbf{x}) = \sum_{i \in \mathcal{F}} \mathbf{w}_i \mathbf{x}_i. \tag{3}$$

As with similarity models, feature models vary in their rules for updating the feature weights (e.g., Kruschke, 2001; Mackintosh, 1975; Rescorla & Wagner, 1972). Setting aside those differences for now, the set of value functions (v_{f}) learnable in principle by a feature model is the set of all functions of the form in Equation 3, for any real values of the weights.

Assume now that Equation 1 holds for all pairs of stimuli. This equation entails a strong relationship between the similarity function of the similarity model and the feature representation of the feature model. Assume also that the stimulus vectors $\mathbf{x}$ in the feature model span the feature space, meaning the features are all linearly independent so there is no degeneracy in the feature representation (otherwise, we could simply remove any redundant features). Under these conditions, the set of learnable value functions for the two models is exactly the same (see Ghirlanda, 2015, for a similar result). To see this, first consider any value function predicted by the similarity model, as given in Equation 2. If we define the feature model's weight vector by

$$\mathbf{w} = \sum_{i=1}^{n} c^i \cdot \mathbf{x}^i, \tag{4}$$

[1] We use superscript to indicate different stimuli, and subscript to indicate different features of a stimulus. Dummy variables such as i and j are used to index stimuli or to index features; we always indicate which is being indexed if it is not clear from context.

[2] The GCM might appear more limited than the other two, because of its discrete weights. However, because exemplars can be encountered multiple times in different categories, and because response rules normalize the output values across categories, the GCM's exemplar weights can effectively take on any rational values.

then the feature model's output is

$$\begin{aligned} v_{\mathrm{f}}(\mathbf{x}) &= \sum_{j \in \mathcal{F}} \sum_{i=1}^{n} c^i \mathbf{x}_j^i \mathbf{x}_j \\ &= \sum_{i=1}^{n} c^i \mathrm{sim}\left(x^i, x\right) \\ &= v_{\mathrm{s}}(x). \end{aligned} \tag{5}$$

Conversely, for any choice of feature weight vector $\mathbf{w}$, the assumption that the stimuli span the feature space implies there exist coefficients $\left\{c^1, \ldots, c^n\right\}$ and stimuli $\left\{\mathbf{x}^1, \ldots, \mathbf{x}^n\right\}$ that satisfy Equation 4. Adopting these coefficients and the corresponding exemplars $\left\{x^1, \ldots, x^n\right\}$ for the similarity model, we see once again that $v_{\mathrm{f}}(\mathbf{x}) = v_{\mathrm{s}}(x)$ by the same reasoning as in Equation 5.

Thus, any similarity model is equivalent to a feature model, and any feature model is equivalent to a similarity model, provided there exists a representation for the other model (i.e., a feature set or a similarity function) satisfying Equation 1. This equivalence is in terms of value functions that the models can in principle produce; we address the question of specific learning rules later in this paper. The demonstrated here is a fairly elementary result from a mathematical perspective, and it has been observed previously in the psychological literature (Ghirlanda, 2015; Jäkel et al., 2007, 2009). Nevertheless, we believe it has not received the appreciation it merits. The proposal advanced here is that the equivalence points to a deeper mathematical understanding of biological learning than is possible from either class of models alone. To provide the groundwork for that deeper understanding, we turn now to a summary of the kernel framework from machine learning.

2.2 The Kernel Framework

The kernel framework is a mathematical system defined by a dual pair of representations on some input set (i.e., means of imposing structure on that set). The first representation is a kernel: a two-place, real-valued function on the inputs. The second representation is the RKHS: a space of real-valued (one-place) functions on the inputs.[3] The representations are dual in that the inner product in the RKHS reproduces the kernel. In this and the following two subsections, we summarize the essential properties of the kernel framework that have been established in the machine-learning literature and offer psychological interpretations of these properties. The interested reader is referred to Aronzajn (1950) and Schölkopf and Smola (2002) for a more thorough treatment of the mathematical details.

Let $\mathcal{X}$ be an arbitrary set, which we can think of as comprising all potential stimuli relevant to some psychological task. Assume we have a function that assigns a real number to every pair of elements of $\mathcal{X}$. This is the kernel function, which we can think of as a measure of similarity:

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}.$$

($\mathbb{R}$ represents the real numbers.) Given a kernel, we can also define the *evaluation function* for each stimulus $x \in \mathcal{X}$, which represents the similarity of x to all other stimuli:

$$k(x, \cdot) : \mathcal{X} \to \mathbb{R}.$$

Assume also that we have a space $\mathcal{H}$ of one-place functions, each of which assigns a real number to every member of $\mathcal{X}$:

$$\forall f \in \mathcal{H}, f : \mathcal{X} \to \mathbb{R}.$$

We can think of each $f \in \mathcal{H}$ as a possible value function that a learning model might be capable of representing. Now assume that $\mathcal{H}$ has the structure of a Hilbert space, which generalizes Euclidean space ($\mathbb{R}^m$) to also include spaces with infinite dimensionality. Formally, a Hilbert space is defined as a real vector space with an inner product, meaning that it is closed under addition (if $f \in \mathcal{H}$ and $g \in \mathcal{H}$, then $f + g \in \mathcal{H}$) and

[3] The framework naturally extends to the case where the kernel and the functions in the RKHS map to the complex numbers, but for present purposes we specialize to the real case.

scalar multiplication (if $f \in \mathcal{H}$ and $c \in \mathbb{R}$ then $cf \in \mathcal{H}$), and it has an inner product that we denote by $\langle \cdot, \cdot \rangle$ ($\langle f, g \rangle \in \mathbb{R}$ for all $f, g \in H$). The inner product generalizes the notion of dot product to vector spaces of possibly infinite dimensionality. We can loosely think of the inner product of two functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ as a measure of their correlation or degree of agreement across the stimulus space.

Given a kernel and Hilbert space as just defined, we say that k is a reproducing kernel for $\mathcal{H}$ (or $\mathcal{H}$ is a RKHS for k) if the evaluation functions $k(x, \cdot)$ are elements of $\mathcal{H}$ for all $x \in X$, and furthermore for all $x \in X$ and $f \in \mathcal{H}$:

$$\langle f, k(x, \cdot) \rangle = f(x). \tag{6}$$

An immediate consequence of this condition is that the inner product reproduces the kernel:

$$\langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x'). \tag{7}$$

The relationship in Equation 6 is commonly referred to as the *reproducing property*, and it links each evaluation function $k(x, \cdot)$ with the corresponding function-evaluation operator $f_x^* : f \mapsto f(x)$ (which is a map from $\mathcal{H}$ to $\mathbb{R}$). The reproducing property is the defining characteristic of a RKHS, and it is equivalent to the kernel duality as defined in Equation 1. Indeed, given a kernel and Hilbert space satisfying Equation 6, we can choose any orthonormal basis $\mathcal{F}$ for $\mathcal{H}$ and treat it as a set of feature dimensions. The feature set then provides provides a coordinate system for $\mathcal{H}$, and in particular the evaluation functions can be expressed as

$$\begin{aligned} k(x, \cdot) &= \sum_i \langle f_i, k(x, \cdot) \rangle f_i \\ &= \sum_i f_i(x) f_i \end{aligned} \tag{8}$$

If we define the feature representation for each stimulus x as the vector of its values on the features in $\mathcal{F}$, $\mathbf{x} = (f_i(x))_{i \in \mathcal{F}}$, then the kernel duality follows:

$$\begin{aligned} \sum_i \mathbf{x}_i \cdot \mathbf{x}'_i &= \left\langle \sum_i f_i(x) f_i, \sum_j f_j(x') f_j \right\rangle \\ &= \langle k(x, \cdot), k(x', \cdot) \rangle \\ &= k(x, x'). \end{aligned} \tag{9}$$

Conversely, if a kernel k and feature set $\mathcal{F}$ satisfy the kernel duality, then the evaluation functions satisfy $k(x, \cdot) = \sum_i f_i(x) f_i$. If we define $\mathcal{H}$ as the Hilbert space generated by the feature functions in $\mathcal{F}$ (i.e., with $\mathcal{F}$ as an orthonormal basis), then any element of $\mathcal{H}$ can be written as $f = \sum_i w_i f_i$ and the reproducing property follows:

$$\begin{aligned} \langle f, k(x, \cdot) \rangle &= \left\langle \sum_i w_i f_i, \sum_j f_j(x) f_j \right\rangle \\ &= \sum_i w_i f_i(x) \\ &= f(x). \end{aligned} \tag{10}$$

Psychologically, we interpret the kernel and Hilbert space as constituting dual representations, one in terms of similarity and the other in terms of features. The bridge between these representations is the evaluation function, $k(x, \cdot)$, which encodes the similarity of x to all other stimuli and is also an element of the vector space $\mathcal{H}$ (which we have also written above as $\mathbf{x}$ to emphasize the vector interpretation). Thus the inner product of any two stimuli in their vector representations equals their similarity.

Under the kernel duality, the kernel and the Hilbert space determine each other uniquely. These properties of unique determination in turn provide two means of constructing RKHS systems: One can define a kernel (satisfying certain necessary and sufficient conditions) and construct the corresponding Hilbert space, or one can define a Hilbert space (again satisfying certain necessary and sufficient conditions) and derive the corresponding kernel. These two methods, which correspond to deriving a feature model equivalent to any similarity model, and deriving a similarity model equivalent to any feature model, are described next.

2.3 From Similarity to Features

Consider first a finite stimulus set $\mathcal{X} = \left\{x^1, \ldots, x^n\right\}$. Any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be written as a similarity matrix K with $K_{ij} = k\left(x^i, x^j\right)$. The function is said to be an admissible kernel if the similarity matrix is symmetric (meaning $k\left(x^i, x^j\right) = k\left(x^j, x^i\right)$) and positive semidefinite (meaning $v^T K v \geq 0$ for any column vector v, where ^{T} indicates transposition). Under these conditions, a standard result from linear algebra implies that the similarity matrix can be decomposed as

$$K = F^T F. \tag{11}$$

The matrix F can be thought of as a feature matrix, with F_{ij} indicating the value of stimulus j on feature i. This matrix is not unique, but one way to construct it is to set the rows of F (i.e., the features) equal to the eigenvectors of K, scaled such that $\sum_j F_{ij}^2$ is equal to the corresponding eigenvalue. Given such a decomposition of the similarity matrix, we can idenfity each stimulus x^i with the vector of its feature values $\mathbf{x}^i = F_{\cdot i}$. Equation 11 can then be rewritten as

$$k\left(x^i, x^j\right) = \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle, \tag{12}$$

where $\left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle = \mathbf{x}^i \cdot \mathbf{x}^j$ is the inner product (or dot product) between the vectors. Thus the feature representation F and the similarity function k satisfy the kernel duality (cf. Ghirlanda, 2015).

This result generalizes to the case of an infinite stimulus set. Let $\mathcal{X}$ now be a multidimensional space, for example a region in $\mathbb{R}^m$ for some integer m. A similarity function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an admissible kernel if it is continuous and symmetric and if the similarity matrix for any finite set of stimuli is positive-semidefinite. Most symmetric similarity functions used in psychological models (e.g., exponential or Gaussian functions of distance) satisfy this property. A classical result in functional analysis known as Mercer's theorem (Mercer, 1909) implies that for any admissible kernel function, there exists a (possibly infinite) set $\mathcal{F}$ of feature functions $f_i : \mathcal{X} \to \mathbb{R}$ such that

$$k\left(x, x'\right) = \sum_i f_i\left(x\right) f_i\left(x'\right) \tag{13}$$

for any stimuli x and x'. We can thus identify each stimulus x with its vector of feature values, $\mathbf{x} = \left(f_i\left(x\right)\right)_i$, and the kernel duality is satisfied for this case of an infinite stimulus set (and typically infinite feature set).

Given the set of features that is dual to a given kernel, one can also define a Hilbert space $\mathcal{H}$ comprising all functions of the form $v = \sum_i w_i f_i$. These constitute all the value functions that can be expressed as linear combinations of the features (Equation 3).[4] The inner product is defined by $\langle f_i, f_i \rangle = 1$ for all i and $\langle f_i, f_j \rangle = 0$ for all $i \neq j$, so that the features form an orthonormal basis (or axis system) for $\mathcal{H}$. From Equation 13, the evaluation function for any stimulus can be written as an element of $\mathcal{H}$ with the coefficients given by its feature values:

$$k\left(x, \cdot\right) = \sum_i f_i\left(x\right) f_i \tag{14}$$

Therefore the evaluation functions satisfy the reproducing property (Equation 6) asderived above in Equation 10. Thus the kernel duality is achieved by identifying each stimulus with its evaluation function (conceptualized as a vector in the Hilbert space $\mathcal{H}$). Notice that identifying a stimulus x with its vector of features values, $\left(f_i\left(x\right)\right)_i$, is essentially the same as identifying it with its evaluation function, $k\left(x, \cdot\right) = \sum_i f_i\left(x\right) f_i$, because the features f_i are a coordinate system (i.e., orthonormal basis) for $\mathcal{H}$. To illustrate in the two-dimensional case, where the feature space is a Cartesian plane, the difference is just that between the pair of numbers $(\mathbf{x}_1, \mathbf{x}_2)$ and the point in the plane with coordinates $(\mathbf{x}_1, \mathbf{x}_2)$. The former emphasizes the analytic represention of a stimulus as a sequence of feature values, and the latter emphasizes the stimulus's relation to all other stimuli via the inner product over all features.

As an alternative to constructing the RKHS from the features guaranteed by Mercer's theorem, one can define it directly from the kernel's evaluation functions, $k\left(x, \cdot\right)$. Specifically, given any stimulus set $\mathcal{X}$ (finite or infinite) and a symmetric and positive-semidefinite kernel k, we can define $\mathcal{H}$ as the space of all functions of

[4] In the case of an infinite feature set, the coefficients are constrained to satisfy $\sum_i w_i^2 < \infty$, which ensures that the inner product in $\mathcal{H}$ is always finite.

the form $\sum_{i=1}^{n} c^i k\left(x^i, \cdot\right)$ for any finite set of stimuli $\left\{x^1, \ldots, x^n\right\}$ and real coefficients $\left\{c^1, \ldots, c^n\right\}$.[5] Thus $\mathcal{H}$ comprises all value functions v_s that can be learned based on similarity (Equation 2). Under this conception of $\mathcal{H}$, the inner product can be directly defined by the reproducing property (Equation 7). Therefore by linearity of the inner product, the inner product of any two elements of $\mathcal{H}$ is given by

$$\left\langle \sum_{i=1}^{n} c^i k\left(x^i, \cdot\right), \sum_{j=1}^{m} d^j k\left(y^j, \cdot\right) \right\rangle = \sum_{i,j} c^i d^j k\left(x^i, y^j\right). \tag{15}$$

This operation is well-defined because, if $f = \sum_{i=1}^{n} c^i k\left(x^i, \cdot\right)$, then for any $g \in \mathcal{H}$:

$$\langle f, g \rangle \quad = \quad \sum_i c^i g\left(x^i\right). \tag{16}$$

Therefore the value of $\langle f, g \rangle$ is independent of how g is written as a sum of evaluation functions (if multiple such sums exist). The same argument applies to f. The operator $\langle \cdot, \cdot \rangle$ is linear in both arguments (Equation 16 shows this for the second argument), and it is symmetric because k is. The positive-semidefiniteness of k can be shown to imply that $\langle f, f \rangle > 0$ for any nonzero $f \in \mathcal{H}$. These properties imply $\langle \cdot, \cdot \rangle$ is an inner product.

As explained below in Section 2.5, any admissible kernel has a unique RKHS satisfying Equation 6. Therefore the construction of $\mathcal{H}$ based on Mercer's theorem (or the finite version in Equation 11) and the construction based on the evaluation functions yield one and the same Hilbert space. In particular, the set of features $\mathcal{F} = \{f_i\}$ guaranteed by Mercer's theorem always forms a basis for $\mathcal{H}$, implying the evaluation functions can always be written according to Equation 14 and thus that any stimulus x can always be identified with its feature values $\left(f_i\left(x\right)\right)_i$. Furthermore, the two equivalent constructions of $\mathcal{H}$ show that it is both the space of all value functions learnable from similarity (Equation 2) and the space of all functions learnable as linear combinations of features (Equation 3).

2.4 From Features to Similarity

Assume we are given a set $\mathcal{F}$ of features, each of which can be thought of as a function from the stimulus space to the real numbers: $f_i : \mathcal{X} \to \mathbb{R}$. Every stimulus x can be identified with its feature values, $\mathbf{x} = \left(f_i\left(x\right)\right)_i$. If there are finitely many features, then it is straightforward to define a kernel or similarity function as the inner product of these feature vectors:

$$k\left(x, x'\right) = \langle \mathbf{x}, \mathbf{x}' \rangle = \sum_i f_i\left(x\right) f_i\left(x'\right). \tag{17}$$

This definition immediately implies that k is symmetric and positive-semidefinite. The definition works just as well if $\mathcal{F}$ is infinite, provided the features satisfy what can be called an l_2 constraint, namely that $\sum_i f_i\left(x\right)^2 < \infty$ for all stimuli x. In other words, every stimulus's feature vector $\mathbf{x}$ must have a finite Euclidean norm within the feature space. If this constraint is satisfied, then the inner product and hence the kernel is guaranteed to be finite for all stimulus pairs, by the Cauchy-Schwarz inequality. Psychologically, the l_2 constraint is a reasonable restriction to assume for infinite-dimensional feature models, because if it is violated then it can be shown that standard updating rules based on error correction (e.g., Rescorla & Wagner, 1972) will lead the model to predict infinite outcome values—effectively, the model learns an infinite amount in a single trial. Conversely, if l_2 is satisfied then a feature learning model will be well-behaved even with an infinite number of features.

Given such a feature set, one can also define a Hilbert space $\mathcal{H}$ comprising all functions of the form $\sum_i w_i f_i$ for real coefficients satisfying $\sum_i w_i^2 < \infty$. Then the evaluation functions $k\left(x, \cdot\right) = \sum_i f_i\left(x\right) f_i$ lie in $\mathcal{H}$, and $\mathcal{H}$ comprises all value functions learnable as linear combinations of features in $\mathcal{F}$, and equivalently all value functions learnable by similarity using k.

A kernel can also be constructed directly from a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, even without a given set of features. For each $x \in \mathcal{X}$, define a functional from $\mathcal{H}$ to $\mathbb{R}$ by $f_x^*\left(f\right) = f\left(x\right)$. If f_x^* is bounded

[5] Technically, to obtain a true Hilbert space we must also consider all convergent (i.e., Cauchy) sequences of such functions, and include the functions that are limits of those sequences. In mathematics, this is the process of *completing* the vector space.

(meaning $f(x) \leq M$ whenever $\langle f, f \rangle \leq 1$, for some bound M), then another classic result in mathematical analysis, the Riesz representation theorem (Frechet, 1907; Reisz, 1907), implies that f_x^* is equivalent to taking the inner product with some element of $\mathcal{H}$. In other words, there is an element f_x of $\mathcal{H}$ satisfying

$$\langle f_x, f \rangle = f_x^*(f) = f(x) \tag{18}$$

for all $f \in \mathcal{H}$. Now f_x, which is a function from $\mathcal{X}$ to $\mathbb{R}$, can be taken to be the function indicating similarity of all stimuli to x. In other words, we can define the kernel by taking f_x as its evaluation function:

$$k(x, \cdot) = f_x. \tag{19}$$

Equations 18and 19 together imply the reproducing property (Equation 6), and thus $\mathcal{H}$ is a RKHS for k. Furthermore, if $\mathcal{F}$ is any orthonormal basis for $\mathcal{H}$, then each evaluation function can be written as

$$\begin{aligned} k(x, \cdot) &= \sum_{i \in \mathcal{F}} \langle k(x, \cdot), f_i \rangle f_i \\ &= \sum_{i \in \mathcal{F}} f_i(x) f_i \end{aligned} \tag{20}$$

and moreover the assumption above that each f_x^* is bounded can be shown to be equivalent to the assumption that the feature set $\mathcal{F}$ satisfies the l_2 constraint.

2.5 Uniqueness and Extension

The previous two subsections have shown how any similarity function has a dual feature space, and vice versa. The essense of these constructions is that any similarity function (if it is symmetric and positive semidefinite) can be factored into the product of a feature set with itself, in the sense of Mercer's theorem or Equation 11, and any feature set (if it satisfies the l_2 constraint) defines an inner product that can be interpreted as a similarity function. In this section we explain the extent to which these constructions are unique. We first summarize the established results from machine learning that the kernel is uniquely determined for any Hilbert space or feature set, and that for any kernel the Hilbert space is uniquely determined and the features are determined up to rotation within the Hilbert space. We then explore some potentially psychologically meaningful ways in which the Hilbert space and feature set can change, with a corresponding change in the kernel.

Given a set of features $\mathcal{F}$, the kernel duality (Equation 1) immediately determines the kernel. Likewise, given a Hilbert space $\mathcal{H}$ for which the evaluation functionals f_x^* are bounded, the element f_x provided by the Riesz representation theorem must be unique. This is because f_x is defined by its inner product with all other elements of $\mathcal{H}$ (Equation 18), and if two members of a Hilbert space match everywhere in terms of the inner product then they must be identical. Uniqueness of f_x in turn implies uniqueness of the kernel (Equation 19). Thus the kernel is uniquely determined by the inner product, as provided either by $\mathcal{H}$ or by $\mathcal{F}$.

Going in the other direction, a kernel does not uniquely determine a dual feature set, but it does determine a unique feature space. The only flexibility comes from rotating the features, which is easy to see in the finite-dimensional case: If U is any rigid rotation matrix, then its transpose equals its inverse, and therefore Equation 11 can be replaced by

$$K = F^T U^T U F = (UF)^T UF. \tag{21}$$

In general, flexibility in choice of $\mathcal{F}$ corresponds to choice of an orthonormal basis for $\mathcal{H}$, which generalizes the idea of choosing an axis system in finite-dimensional Euclidean space.

To see that $\mathcal{H}$ is uniquely determined by k, that is, that every kernel has a unique RKHS, consider the construction of $\mathcal{H}$ in Section 2.3 based on the evaluation functions (e.g., Equation 15). This construction is easily seen to yield the minimal Hilbert space that is dual to k, because it is the minimal Hilbert space that contains the evaluation functions $k(x, \cdot)$. Moreover, the inner product is uniquely determined on $\mathcal{H}$, from (7). Assume now that $\mathcal{H}'$ is another Hilbert space satisfying the reproducing property (Equation (6)). $\mathcal{H}$ nust be a subspace of $\mathcal{H}'$, so for any function f that lies in $\mathcal{H}'$ but not in $\mathcal{H}$, we can define g as the projection of f into $\mathcal{H}$. Since $k(\cdot, x) \in \mathcal{H}$ for any x, its inner products with f and g are the same, implying

$$f(x) = \langle f, k(x, \cdot) \rangle = \langle g, k(x, \cdot) \rangle = g(x). \tag{22}$$

Therefore $f = g$ and $f \in \mathcal{H}$. This implies $\mathcal{H}' = \mathcal{H}$ and thus $\mathcal{H}$ is unique. This result is known as the Moore–Aronszajn theorem (Aronzajn, 1950).

Despite this uniqueness result, in general it is possible to extend an RKHS $\mathcal{H}$ with kernel k to a larger Hilbert space $\mathcal{H}^+$, with a new kernel k^+ that differs from the original in interesting ways. To see how the two kernels would be related, note that for any $f \in \mathcal{H}$ its inner products with the two kernels are equal:

$$\langle f, k(x, \cdot)\rangle = f(x) = \left\langle f, k^+(x, \cdot)\right\rangle. \tag{23}$$

Therefore k and k^+ differ only by functions that are orthogonal to $\mathcal{H}$. Thus for all x we can write

$$k^+(x, \cdot) = k(x, \cdot) + g_x$$

for some unique g_x in $\mathcal{H}^\perp$, the subspace of $\mathcal{H}^+$ that is orthogonal to $\mathcal{H}$. Now for any $g \in \mathcal{H}^\perp$ and $x \in \mathcal{X}$,

$$\begin{aligned} g(x) &= \left\langle g, k^+(\cdot, x)\right\rangle \\ &= \langle g, k(\cdot, x) + g_x\rangle \\ &= \langle g, g_x\rangle. \end{aligned} \tag{24}$$

Therefore $(g_x)_{x\in X}$ defines the reproducing kernel for $\mathcal{H}^\perp$, via $k^\perp(x, x') = g_x(x')$. In conclusion, any extension of $\mathcal{H}$ must take the form $\mathcal{H}^+ = \mathcal{H} \oplus \mathcal{H}^\perp$ with $k^+ = k + k^\perp$. The direct sum, $\oplus$, essentially indicates a Cartesian product (e.g., $\mathbb{R}^n \oplus \mathbb{R}^m = \mathbb{R}^{n+m}$), and it implies here that bases or feature sets for these Hilbert spaces would be related according to $\mathcal{F}^+ = \mathcal{F} \cup \mathcal{F}^\perp$. Psychologically the extension from $\mathcal{H}$ to $\mathcal{H}^+$ could be interpreted as learning new features (i.e., dimensions of the feature space), whose contribution to similarity (the kernel) is simply added to the existing similarity function. Stated differently, the sum of any two similarity functions corresponds, via the kernel duality, to the union of their corresponding feature sets.

A special case arises when the original (i.e., smaller) Hilbert space is defined by the kernel evaluation functions on one stimlus set, and the new (larger) Hilbert space is defined by an expanded stimulus set. Let k_1 be a kernel on stimulus set $\mathcal{X}_1$, and $\mathcal{H}_1$ the corresponding RKHS. Let $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ (with $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$) be an expansion of the stimulus set, with a kernel k that extends k_1 (i.e., $k(x, x') = k_1(x, x')$ for any $x, x' \in \mathcal{X}_1$), and let $\mathcal{H}$ be the RKHS corresponding to k. In other words, the stimulus set has expanded but similarity among the original stimuli is unchanged.

To understand how the two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}$ relate, we first define a new space, $\tilde{\mathcal{H}}_1$, that is a minimal extension of $\mathcal{H}_1$ from a space of functions on $\mathcal{X}_1$ to a space of functions on $\mathcal{X}$. We do this by considering the evaluation functions for all stimuli in $\mathcal{X}_1$ but treated as functions on $\mathcal{X}$; that is, $k(x_1, \cdot) : \mathcal{X} \to \mathbb{R}$ for all $x_1 \in \mathcal{X}_1$. Let $\tilde{\mathcal{H}}_1$ be the Hilbert space generated by these functions, with inner product defined by

$$\langle k(\cdot, x_1), k(\cdot, x_1')\rangle_{\tilde{H}_1} = k_1(x_1, x_1') = k(x_1, x_1') \tag{25}$$

for all $x_1, x_1' \in X_1$. Note that the inner product in $\tilde{\mathcal{H}}_1$ is the pullback of the inner product in $\mathcal{H}_1$, via restriction from $\mathcal{X}$ to $\mathcal{X}_1$. That is, for all $f, f' \in \tilde{\mathcal{H}}_1$,

$$\langle f, f'\rangle_{\tilde{\mathcal{H}}_1} = \langle f|_{\mathcal{X}_1}, f'|_{\mathcal{X}_1}\rangle_{\mathcal{H}_1}. \tag{26}$$

Furthermore, restriction to $\mathcal{X}_1$ constitutes a 1-1 mapping between $\tilde{\mathcal{H}}_1$ and $\mathcal{H}_1$. This is because, if any function $f = \sum_i c^i k\left(x^i, \cdot\right)$ (with $x^i \in \mathcal{X}_1$) is equal to zero everywhere on $\mathcal{X}_1$, then

$$\langle f, f\rangle_{\tilde{\mathcal{H}}_1} = \sum_i c^i f\left(x^i\right) = 0, \tag{27}$$

which implies f must be zero everywhere. Therefore any two functions in $\tilde{\mathcal{H}}_1$ that agree on $\mathcal{X}_1$ must agree everywhere on $\mathcal{X}$. Consequently, $\mathcal{H}_1$ and $\tilde{\mathcal{H}}_1$ are isomorphic; that is, the RKHS on $\mathcal{X}_1$ can be generalized to an identically structured RKHS on $\mathcal{X}$. From the standpoint of the feature representation, the feature set is unchanged. The only change is that the embedding of the stimulus set $\mathcal{X}_1$ into the feature space (via $x_1 \mapsto k(x_1, \cdot)$) has been extended to cover the larger stimulus set $\mathcal{X}$.

The definition of the inner product in $\tilde{\mathcal{H}}_1$ implies that $\langle f, f'\rangle_{\mathcal{H}} = \langle f, f'\rangle_{\tilde{\mathcal{H}}_1}$ for all $f, f' \in \tilde{\mathcal{H}}_1$, and thus $\mathcal{H}$ is an extension of $\tilde{\mathcal{H}}_1$ (i.e. $\tilde{\mathcal{H}}_1$ is a subspace of $\mathcal{H}$). Therefore, using the results above, $\mathcal{H}$ can be decomposed

as $\tilde{\mathcal{H}}_1 \oplus \tilde{\mathcal{H}}_2$ with $k = \tilde{k}_1 + \tilde{k}_2$, where $\tilde{k}_1$ and $\tilde{k}_2$ are the reproducing kernels for $\tilde{\mathcal{H}}_1$ and $\tilde{\mathcal{H}}_2$ (and $\tilde{\mathcal{H}}_1$, $\tilde{k}_1$, $\tilde{k}_2$, $\tilde{\mathcal{H}}_2$ are all defined on the full stimulus set $\mathcal{X}$). For all $x_1 \in \mathcal{X}_1$, $k(x_1, \cdot)$ satisfies the reproducing property in $\tilde{\mathcal{H}}_1$, and thus $\tilde{k}_1(x, x_1) = k(x, x_1)$ for all $x \in \mathcal{X}$ and $x_1 \in \mathcal{X}_1$. Therefore $\tilde{k}_1$ and k agree everywhere except on $\mathcal{X}_2 \times \mathcal{X}_2$, and thus $\tilde{k}_2 = 0$ everywhere except $\mathcal{X}_2 \times \mathcal{X}_2$. By the same argument as above, $\tilde{\mathcal{H}}_2$ is isomorphic to $\mathcal{H}_2$, the RKHS on $\mathcal{X}_2$ defined by the restricted kernel $k_2 = \tilde{k}_2|_{\mathcal{X}_2 \times \mathcal{X}_2}$.

In summary, given an original stimulus set ($\mathcal{X}_1$) and kernel (k_1), and extensions thereof ($\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$, $k|_{\mathcal{X}_1 \times \mathcal{X}_1} = k_1$), the extended RKHS ($\mathcal{H}$) is a direct sum of the original RKHS ($\mathcal{H}_1$) and an RKHS on the new stimuli ($\mathcal{H}_2$), with the latter comprising all functions orthogonal to those generated by the original set. Psychologically, the sequence $\mathcal{H}_1 \rightarrow \tilde{\mathcal{H}}_1 \rightarrow \mathcal{H}$ can be interpreted as follows: When a new group of stimuli ($\mathcal{X}_2$) is encountered, they can initially be represented using the class of features already known for the original stimuli ($\mathcal{F}_1$). The class of learnable concepts is thus far unchanged ($\tilde{\mathcal{H}}_1 \cong \mathcal{H}_1$). However, additional features unique to the new stimuli (i.e., that do not vary among the original stimuli) can then expand the representation ($\mathcal{H} = \tilde{\mathcal{H}}_1 \oplus \tilde{\mathcal{H}}_2$ and $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$). Under the kernel duality, incorporating these new features corresponds to adding a new component to the similarity function ($\tilde{k}_2$) that has no impact on the original stimuli.

2.6 Psychological applications

In machine learning, the RKHS or feature space representation is not explicitly calculated, because the learning algorithms used with kernel methods require calculation of only the inner products, which can be evaluated directly using the kernel function. This is the kernel trick, which is considered a primary advantage of the approach in that field. In psychology, however, we are often interested in models that deal directly with feature representations, which in the kernel framework correspond to the dimensional structure of the Hilbert space. Understanding this structure is thus of potential value for cognitive modeling. Unfortunately, standard methods for defining RKHSs do not provide means for determining a set of features that form an orthonormal basis for the space. Mercer's theorem guarantees these feature functions exist (Equation 13), but it is not constructive. The construction of the RKHS based on evaluation functions (Equation 15) provides an explicit representation of every function in the space, but it also does not provide an orthonomal basis. Such a basis would be important for explicitly deriving feature-based models (e.g., perceptrons) that are dual to certain similarity-based models (e.g., exemplar models), and for understanding how the feature representations of the former depend on mechanisms in the latter such as attention learning.

The new results of Section 2.5 offer some ideas about psychological representations and representation change. Although these results to not provide a full basis for a RKHS, they provide some structure by showing how it can be decomposed into subspaces defined by different classes of features or by different subsets of stimuli. They also show that such vector-space decompositions are concomitant with additive decompositions of the kernel, with potentially useful relationships between an original kernel and the additive components of a new, expanded kernel.

The construction in Section 2.3 implies that the evaluation functions constitute a frame or Riesz basis for the RKHS, meaning that every function in the space can be written as a linear combination of evaluation functions (see also Equations 2 & 3). This property has applications for example in wavelet analysis in vision. However, it is important to recognize that the evaluation functions are not orthogonal, except in the degenerate case where similarity between any distinct stimuli is zero. Mistaking the evaluation functions for an orthonormal basis encourages one to consider a perceptron-style model in which the kernel directly defines a set of feature dimensions. That is, the kernel is viewed as embodying a receptive field, with $k(x, \cdot)$ defining an activation function for a feature "centered" at x. For example, if the stimulus space were equipped with an m-dimensional Euclidean geometry and k were a Gaussian function of distance in that geometry, then one could imagine a basis function centered on each stimulus (or on a lattice of stimuli) with a m-variate Gaussian receptive field. Although we see promise in an approach that equates basis functions (i.e., feature dimensions) with receptive fields, and that identifies each basis function with a stimulus that lies at the center of its receptive field, the function defining the receptive field in such a model is not the same as the kernel.

In Section 3 we develop and analyze models of this type, and derive the correct relationship between receptive fields, feature distributions, and the kernel. Briefly, because the kernel is the inner product over all feature dimensions, it essentially equals the self-convolution of the function defining the receptive field.

To see this, assume $\mathcal{X}$ has a vector-space structure (e.g., $\mathbb{R}^m$) and a set of basis functions is defined by

$$f_x(x') = r(x' - x).$$

Here we think of r as a tuning curve and b_x as an activation function for a feature centered on x with shape determined by r.[6] Then the feature representation of any stimulus x' is given by

$$\mathbf{x}' = (f_x(x'))_{x \in \mathcal{X}} = (r(x' - x))_{x \in \mathcal{X}} \tag{28}$$

and the kernel is given by

$$\begin{aligned} k(x', x'') &= \langle \mathbf{x}', \mathbf{x}'' \rangle \\ &= \int_{\mathcal{X}} r(x' - x) \cdot r(x'' - x) \, \mathrm{d}x, \end{aligned} \tag{29}$$

where the integral is with respect to whatever measure determines the distribution of the basis functions b_x (i.e., the density of stimuli x on which the basis functions are centered). If r is a Gaussian then k is also a Gaussian, with double the variance. However, if r is more complex, as might be the case for more structured stimuli, then the difference between r and k could be more significant. This exposition illustrates the utility of the kernel duality perspective as applied to psychology, as that perspective explicitly acknowledges the two representations involved (viz., similarity and feature-based) as well as the different roles of the kernel within these two representations.

3 Translating similarity models into continuous-features models

3.1 Feature models for continuous stimulus spaces

Many similarity-based models, especially those in the tradition of multidimensional scaling, are founded on continuous stimulus spaces (Nosofsky, 1992; Shepard, 1962). The stimulus space is typically modeled as $\mathcal{X} \subseteq \mathbb{R}^m$ for some integer m, and similarity as a function of vectors in this space, $\mathrm{sim}(x, x')$. Most often the similarity function is translation-invariant (Shepard, 1957), meaning it can be written as

$$\mathrm{sim}(x, x') = \Gamma(d), \tag{30}$$

where Γ is a generalization gradient and $d = x - x'$ is the vector indicating the difference between two stimuli. Here we investigate how these models can be recast as feature-based models, using the kernel duality. Thus the goal is to translate the similarity function (treated as a kernel) into a corresponding feature space.

To avert possible confusion, we stress that the dimensional representation $x \in \mathbb{R}^m$ that determines similarity in Equation 30 is entirely different from the feature representation $\mathbf{x} \in \mathcal{H}$. The goal here is to derive the Hilbert space $\mathcal{H}$, and associated basis of feature functions $\mathcal{F}$, such that the two representations satisfy the kernel duality.

As explained in Section 2.5 earlier, the Hilbert space associated with any kernel is unique. However, interpreting a RKHS as a feature space requires selecting an orthonormal basis for that Hilbert space, with each basis element interpreted as a feature. Because the choice of basis will not be unique, neither will the feature representation. That is, for any feature representation that reproduces the kernel in question, arbitrary "rotations" of that representation (e.g., replacing features f_1 and f_2 with $\sin\theta \cdot f_1 + \cos\theta \cdot f_2$ and $\cos\theta \cdot f_1 - \sin\theta \cdot f_2$) will also reproduce the kernel. Therefore we must constrain the problem further.

Our approach here is to assume that all features have a shared, predetermined shape, and that they differ only in scale and in location within the stimulus space. Specifically, we define a *tuning curve*, $r : \mathbb{R}^m \to \mathbb{R}$, which is taken to integrate to unity and to be symmetric, $r(-d) = r(d)$. Then for all stimuli $z \in \mathbb{R}^m$ and scalars $s \in \mathbb{R}^+$ (i.e., positive real numbers), we define a feature $f_{s,z} : \mathbb{R}^m \to \mathbb{R}$ by

$$f_{s,z}(x) = \tfrac{1}{s^m} r\left(\tfrac{x - z}{s}\right). \tag{31}$$

[6] One could also assume r is symmetric: $r(x' - x) = r(x - x')$. This assumption is unnecessary for symmetry of k, but it does make the self-convolution interpretation more exact.

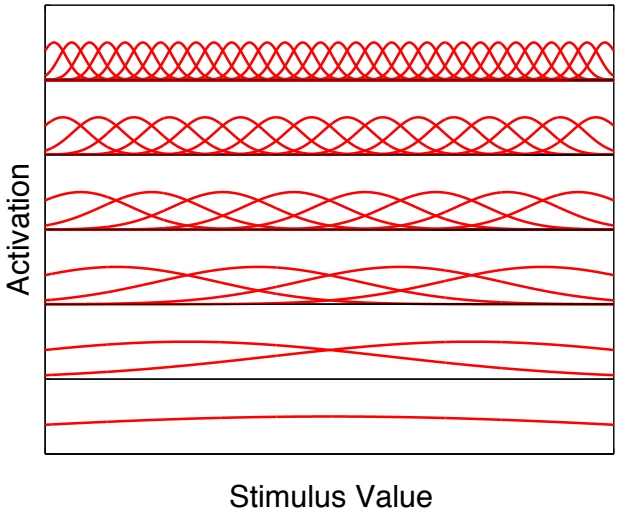

Figure 1: Schematic population of features using a Gaussian tuning curve. Each curve indicates the activation function of one representative feature. Each tier in the diagram contains features with different resolutions (separated and renormalized for visual clarity). Features at each level of resolution are uniformly distributed throughout the stimulus space. The similarity kernel or generalization gradient generated by the model is determined by the relative density of features at different resolutions.

Thus the feature $f_{s,z}$ can be thought of as having a receptive field centered on z, with activation function instantiating the tuning curve r with scale s. (The scaling coefficient $\frac{1}{s^m}$ ensures every feature's activation function integrates to unity, a property that will be useful later.) For example, r could be the standard m-variate Gaussian, and then the set of features comprises Gaussian receptive fields at all scales and locations (Figure 1). Thus the model admits a loose interpretation as a population of sensory neurons, although it also could apply at higher stages of processing, where stimulus representations are more complex.

Because the set of features is infinite, we require a measure on that set.[7] We assume the measure is translation-invariant (i.e., independent of z), which will imply translation-invariance of the corresponding kernel. Thus the question of the measure reduces to one of specifying the density of features at each scale s, which we represent as a weighting function, $p(s)$. We take the weighting function to integrate to unity, and therefore it can be thought of as determining the proportion of features at any (range of) spatial scale. Under this model, any stimuli x and x' have feature representations $\mathbf{x}$ and $\mathbf{x}'$, corresponding to their values on all features, and the inner product of these feature representations is given by:

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \int \int f_{s,z}(x) f_{s,z}(x') p(s) \, \mathrm{d}z \mathrm{d}s. \tag{32}$$

The model will be equivalent to the similarity model in Equation 30 if $\langle \mathbf{x}, \mathbf{x}' \rangle = \mathrm{sim}(x, x') = \Gamma(x - x')$ for all x and x'. In that case we will have translated the similarity model into a feature-based model, one with an infinity of features with a regular organization. We refer to the latter as the *continuous-features model*, because features are continuously distributed throughout the stimulus space.

One property of the continuous-features model is that, given a stimulus x, its similarity integrated over all other stimuli equals unity. This follows from the assumptions that r and $p(s)$ integrate to unity:

$$\begin{aligned} \int \mathrm{sim}(x, x') \, \mathrm{d}x' &= \int \int \int \tfrac{1}{s^m} r\left(\tfrac{x-z}{s}\right) \tfrac{1}{s^m} r\left(\tfrac{x'-z}{s}\right) p(s) \, \mathrm{d}z \mathrm{d}s \mathrm{d}x' \\ &= \int p(s) \left[\int \tfrac{1}{s^m} r\left(\tfrac{x-z}{s}\right) \left[\int \tfrac{1}{s^m} r\left(\tfrac{x'-z}{s}\right) \mathrm{d}x' \right] \mathrm{d}z \right] \mathrm{d}s \\ &= 1. \end{aligned} \tag{33}$$

[7] A technical issue here concerns whether the feature set is countably or uncountably infinite. The present construction implies it is uncountable, although Mercer's theorem guarantees a countable feature set. We set aside this subtle inaccuracy for the present exposition, because it enables more convenient derivations in terms of integrals.

In psychological applications, the scaling of similarity is usually arbitrary, so that $c \cdot \text{sim}$ is essentially the same as sim. Therefore a given similarity function can always be rescaled to satisfy $\int \text{sim}(x, x')\,\mathrm{d}x' = 1$ (provided the integral is finite). Note that this offers an alternative to the common convention of defining self-similarity to equal unity, $\text{sim}(x, x) = \Gamma(0) = 1$.

In summary, we define a feature-based model with a continuous set of features, with receptive fields all having the same shape and varying only in resolution and location. The features at a given scale are assumed to be uniformly distributed throughout the stimulus space. The flexibility in the model lies in the density of features at each scale. The goal of translating a similarity model like that in Equation 30 into this feature-based framework is to determine the resolution-density function ($p(s)$) that will reproduce the given kernel or generalization gradient (Γ).

3.2 Unidimensional Stimuli

Consider first a unidimensional stimulus space, isomorphic to $\mathbb{R}$. Thus the tuning curve is a function $r : \mathbb{R} \to \mathbb{R}$ that is symmetric about zero, and each feature $f_{s,z}$ has an activation function replicating r with width s and location z. Let x and y be any two stimuli, separated by a distance $d = |x - y|$. The inner product of their associated feature representations is equal to

$$\begin{aligned} \langle \mathbf{x}, \mathbf{x}' \rangle &= \int\int \tfrac{1}{s^2} r\left(\tfrac{z}{s}\right) r\left(\tfrac{d-z}{s}\right) p(s)\,\mathrm{d}z\mathrm{d}s \\ &= \int \tfrac{1}{s} \left[r * r\right]\left(\tfrac{d}{s}\right) p(s)\,\mathrm{d}s, \end{aligned} \tag{34}$$

where $r * r$ denotes the convolution of the tuning curve with itself:

$$\left[r * r\right](d) = \int r(z)\, r(z - d)\,\mathrm{d}z. \tag{35}$$

Therefore the model yields a generalization gradient given by

$$\Gamma(d) = \int \tfrac{1}{s} \left[r * r\right]\left(\tfrac{d}{s}\right) p(s)\,\mathrm{d}s. \tag{36}$$

The generalization gradient contributed by any fixed feature resolution (i.e., value of s) equals the convolution of the tuning curve with itself, scaled by s. When resolution is variable, the overall generalization gradient, or kernel, is a mixture of the kernels contributed by individual resolutions, with mixture distribution $p(s)$. This mixture property relates to work in machine learning on learning of a kernel via convex optimization (e.g., Micchieli & Pontil, 2007). This connection will become useful later when we explore mechanisms for attention learning, which can be characterized as learning $p(s)$.

As a detailed example, we work through the simple case where features are all-or-none intervals. The tuning curve is thus a boxcar distribution,

$$r(d) = \begin{cases} 1 & -\frac{1}{2} \le d \le \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{37}$$

and its self-convolution $r * r$ is a triangular distribution,

$$\left[r * r\right](d) = \begin{cases} 1 - d & -1 \le d \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

Thus the generalization gradient is a mixture of these triangular distributions, expressed as

$$\begin{aligned} \Gamma(d) &= \int_{\frac{d}{s} \le 1} \tfrac{1}{s} \left(1 - \tfrac{d}{s}\right) p(s)\,\mathrm{d}s \\ &= \int_{s=d}^{\infty} \tfrac{s-d}{s^2} p(s)\,\mathrm{d}s. \end{aligned} \tag{39}$$

By differentiating twice with respect to d, we obtain a complementary relationship giving p in terms of Γ:

$$\Gamma'(d) = \int_{s=d}^{\infty} -\tfrac{1}{s^2} p_s(s)\,\mathrm{d}s, \tag{40}$$

$$\Gamma''(d) = \tfrac{1}{d^2} p_s(d), \tag{41}$$

and therefore

$$p(s) = s^2 \Gamma''(s). \tag{42}$$

Thus the density of feature resolutions can be uniquely determined for any generalization gradient, provided it is twice-differentiable and convex.[8]

For example, if generalization obeys the empirically supported exponential law, $\Gamma(d) = \frac{\alpha}{2} e^{-\alpha d}$ (Shepard, 1987), then $p(s) = \frac{\alpha^3}{2} s^2 e^{-\alpha s}$; that is, $s \sim \text{Gamma}(3, \alpha)$. Scale-free or power-law generalization, $\Gamma(d) \propto d^{-a}$, corresponds to scale-free feature density, $p(s) \propto s^{-a}$ (with a truncated tail at zero or infinity, depending on whether $a \geq 1$ or $a \leq 1$). If we switch from a boxcar to a Gaussian tuning curve, then numerical calculations (available from the first author) show that $s \sim \text{Gamma}(3, \alpha)$ still yields a close approximation to an exponential generalization gradient. Thus if the continuous-features model is taken literally as a model of neural processing, then the model together with Shepard's (1987) theory of exponential generalization gradients offers a precise prediction about the distribution of the sizes of tuning curves in sensory neurons.

For general choices of the tuning curve r, Equation 36 is a Fredholm integral equation (Fredholm, 1903) for determining the function $p(s)$ from a given generalization gradient Γ. By introducing the changes of variables $l = \ln(d)$ and $u = \ln(s)$, and the functions $\Theta(l) = \Gamma\left(e^l\right)$, $q(u) = p(e^u)$, $\beta(x) = [r * r](e^x)$, we obtain a new Fredholm equation that admits an analytic solution:

$$\begin{aligned}
\Theta(l) &= \Gamma(d) \\
&= \int \tfrac{1}{s} [r * r]\left(\tfrac{d}{s}\right) p(s)\,\mathrm{d}s \\
&= \int \beta(l-u)\, q(u)\,\mathrm{d}u \\
&= [\beta * q](l).
\end{aligned} \tag{43}$$

Thus Θ is the convolution of β and q, which in many cases will enable q—and hence $p(s)$—to be explicitly derived using Fourier transforms. In other words, Equation 43 provides a simple connection between the generalization gradient (Θ, as a transform of Γ), the tuning curve (β, as a transform of r), and the resolution-density function (q, as a function of p). This connection offers a means for developing a general theory of continuous-features models that are dual to various similarity models under various tuning curves.

3.3 Multidimensional Stimuli and Dimensional Separability

We now consider the case of a multidimensional stimulus space, modeled as $\mathcal{X} = \mathbb{R}^m$ with $m > 1$. A longstanding axiom in psychological modeling (originating with Medin & Schaffer, 1978), is that similarity is multiplicative across dimensions:

$$\text{sim}(x, x') = \prod_i \text{sim}_i(x_i, x_i').$$

Within the present framework, there are two natural ways to achieve this relationship, differing in how the features combine across dimensions. The first is a tensor-product approach (cf. Smolensky, 1990), wherein the joint model contains a feature for every possible combination of features on the individual dimensions. That is, we assume for each dimension i a set of pre-features $\mathcal{F}_i$ and define the unidimensional similarity on that dimension as

$$\text{sim}_i(x_i, x_i') = \int_{f \in \mathcal{F}_i} f(x_i)\, f(x_i')\,\mathrm{d}f. \tag{44}$$

[8] The requirement that Γ be convex follows from the fact that $p(s)$ cannot be negative. Note, however, that other choices of r produce different relationships between $p(s)$ and Γ and can yield nonconvex generalization gradients. For example, if r is Gaussian then Γ is a mixture of Gaussians at different scales, which can be nonconvex.

For example, the elements of $\mathcal{F}_i$ could be distributed as in Section 3.2, and then similarity on each dimension would be as in Equation 36. Given a set of pre-features for each dimension, we then define the set of multidimensional features as the Cartesian product across dimensions,

$$\mathcal{F} = \bigotimes \mathcal{F}_i. \tag{45}$$

Thus each feature $f \in \mathcal{F}$ is defined by a set of one receptive field per dimension, $f = (f_1, \ldots, f_m)$, and we define its activation function as the product of its constituents':

$$(f_1, \ldots, f_m)(x) = \prod_i f_i(x_i). \tag{46}$$

Then the multiplicative similarity rule holds:

$$\begin{aligned}
\operatorname{sim}(x, x') &= \int_{f \in \mathcal{F}} f(x) f(x') \, \mathrm{d}f \\
&= \int_{f_1 \in \mathcal{F}_1} \cdots \int_{f_m \in \mathcal{F}_m} \prod_{i=1}^{m} \left[f_i(x_i) f_i(x_i') \right] \mathrm{d}f_1 \cdots \mathrm{d}f_m \\
&= \prod_{i=1}^{m} \left[\int_{f_i \in \mathcal{F}_i} f_i(x_i) f_i(x_i') \, \mathrm{d}f_i \right] \\
&= \prod_{i=1}^{m} \operatorname{sim}_i(x_i, y_i).
\end{aligned} \tag{47}$$

Alternatively, we could assume that there are no multidimensional features but that instead similarity is determined separately on each dimension and then directly multiplied together to determine joint similarity. That is, the dimension-specific feature sets $\mathcal{F}_i$ are primary, and the relationship

$$\operatorname{sim}(x, x') = \prod_{i=1}^{m} \left[\int_{f_i \in \mathcal{F}_i} f_i(x_i) f_i(x_i') \, \mathrm{d}f_i \right] \tag{48}$$

is computed directly rather than via joint features as defined in Equations 45 and 46.

The distinction between these two possible feature representations for multidimensional stimuli can be taken as a theory of the psychological distinction between integral and separable dimensions (Garner, 1974; Shepard, 1964). The tensor-product representation models the psychological representation of integral dimensions, because there is no explicit dimensional structure to the set of features $\mathcal{F}$. Although the joint features were mathematically constructed as m-tuples of unidimensional pre-features, psychologically we assume the joint features are primary. The set of features having a particular choice of f_i for dimension i (i.e., the set $\mathcal{F}^1 \times \cdots \times \{f^i\} \times \cdots \times \mathcal{F}^n$) has no coherent psychological identity. Likewise, stimuli matching on one dimension have nothing psychologically special in common, and thus a value x_i on dimension i has no particular meaning beyond the whole stimulus x that exhibits that value. Thus the representation affords no basis for dimension-specific rules or dimensional selective attention. Nevertheless, the stimulus space is not entirely unstructured. In particular, the representation provides enough structure to explain the results of Jones and Goldstone (2013), which ruled out a purely topological model of integral dimensions (i.e., a representation based solely on local similarity). Indeed, the distribution of features in $\mathcal{F}$, as inherited from the Cartesian product in Equation 45, defines a local geometry sufficient for learning of orthogonal constituent dimensions of the sort implied by Jones and Goldstone's experiments.

In contrast, the factorial representation of Equation 48 models the psychological representation of separable dimensions, because it makes the dimensional structure psychologically explicit. The features in $\mathcal{F}_i$ enable recognition of stimuli matching on dimension i, in turn enabling learning of rules and other higher-order concepts. Factorial representations for separable dimensions are also normatively defensible, because of the geometric explosion of features that would otherwise be required under a joint-feature representation. Maintaining joint features that simultaneously encode values on multiple dimensions makes sense only when stimuli defined by those dimensions are better processed holistically, as is arguably the defining property of integral dimensions.

A primary theory of the integral/separable distinction in the literature concerns the shapes of multidimensional generalization gradients (more precisely, of their isoclines), or equivelently of the difference between Euclidean (L2) and city-block (L1) dissimilarity metrics. Integral dimensions tend to exhibit circular generalization gradients and are better modeled with L2 metrics in multidimensional scaling. Separable dimensions tend to exhibit diamond-shaped generalization gradients and are better modeled with L1 metrics. The tensor-product representation automatically produces circular generalization consistent with an L2 metric—the only assumption required is the natural one that the joint tuning curve is a radial basis function (i.e., is isotropic). Although the factorial representation does not necessarily generate L1 generalization, it does under the additional assumption that generalization is an exponential function of dissimilarity (Shepard, 1987). In that case, we have

$$\text{sim}\left(x_i, x_i'\right) \quad = \quad \tfrac{\alpha_i}{2} e^{-\alpha_i \left| x_i - x_i' \right|}, \tag{49}$$

where α_i determines the scale of the generalization gradient along dimension i, often interpreted as an attention parameter (see Section 4). Then overall similarity is given by

$$\begin{aligned} \text{sim}\left(x, x'\right) &= \prod_{i=1}^{m} \text{sim}_i\left(x_i, x_i'\right) \\ &= \tfrac{1}{Z} e^{-\sum_i \alpha_i \left| x_i - x_i' \right|}, \end{aligned} \tag{50}$$

where $Z = 2^m / \prod_{i=1}^{m} \alpha_i$ is a normalization constant. Thus similarity is a monotonic function of the L1 (i.e., city-block) distance $\left\| x - x' \right\|_1$ (with each dimension i scaled by α_i), yielding diamond-shaped isoclines.[9]

Finally, the joint and factorial representations introduced here naturally offer an explanation of the well-known findings that dimensional selective attention is easier with separable than with integral dimensions (e.g., Garner, 1974). Dimensional selective attention has been modeled in the literature as a change in the generalization gradient, becoming narrower along attended dimensions and broader along unattended dimensions (Kruschke, 1992; Nosofsky, 1986). Fits of these models show this mechanism operates more effectively with separable than with integral dimensions (Jones, Maddox, & Love, 2005; Kruschke, 1992; Nosofsky, 1987). This empirical dissociation concords nicely with the distinction proposed here between factorial and joint-feature representations. Under a factorial representation, attention to individual dimensions is easily achieved, because each dimension has its own bank of features that can be weighted independently of how other dimensions are processed, enabling independent control of sim_i for each dimension i. As we show below in Section 4, the principle of narrower generalization gradients for attended dimensions and broader gradients for unattended dimensions corresponds to emphasizing fine-resolution features in the former and coarse-resolution features in the latter (i.e., shifting the distribution $p\left(s\right)$ toward smaller and larger values of s, respectively). In contrast, under a joint-feature representation, the set of features has no dimensional decomposition. Changing the weighting of fine- versus coarse-scale features would change the overall shape and breadth of the generalization gradient, but in an isotropic manner.

3.4 Bayesian Interpretation

Shepard (1987) proposed a Bayesian model of generalization based on consequential regions (CRs) of stimuli that reliably predict a given meaningful outcome. Generalization between stimuli x and x' is assumed to reflect the posterior probability that x' belongs to the CR, given that x was sampled from the CR. We show that this model is equivalent to a special case of the continuous-features model proposed above, where the tuning curve r is taken to be an all-or-none (i.e., 0/1) function with shape matching that of the CR. Furthermore, we show that this equivalence can be extended to provide a Bayesian interpretation of the continuous-features model under any tuning curve (not just all-or-none ones).

Under the CR model, the subject assumes the shape of the CR is known (in the unidimensional case, it is a connected interval), and its size is unknown with some given prior distribution. The location of the CR is

[9] The astute reader will realize that multiplicative similarity together with exponential generalization gradients always implies an L1 metric. That is, this conclusion does not depend on the factorial representation but would also hold under the joint-feature representation. We do not pursue this issue in detail here because it is not specific to the present model. Rather, the issue is that the axioms of multiplicative similarity and exponential generalization taken together are at odds with the empirical finding of circular isoclines for integral dimensions.

uniformly distributed over the stimulus space (i.e., an improper prior distribution). The uniformity assumption and the assumption that the shape of the CR is independent of its location imply the representation is translation-invariant, just as in the continuous-features model. Shepard (1987) suggests this homogeneity property could reflect prior learning or evolution of the stimulus space, a proposal that applies equally to our model.

Under the Bayesian framing, each possible CR can be identified with a hypothesis $h_{s,z}$, paralleling the notation used in the continuous-features model, with the scaling defined so that the volume of the CR is equal to s^m. Denote the prior distribution on the true CR's size by $\rho(s)$. On observing a stimulus x paired with the outcome in question (e.g., reward), the subject assumes that x was randomly sampled from the true CR. Letting x_{R} indicate that x was presented with reward, the likelihood $p(x_{\mathrm{R}}|h_{s,z})$ for any candidate CR equals s^{-m} for $x \in h_{s,z}$ and zero for $x \notin h_{s,z}$. The model's posterior probability that another stimulus x' lies in the true CR—that is, that it will lead to reward (R)—can then be calculated as:

$$\begin{aligned}
\Pr(R|x_{\mathrm{R}}, x') &= \frac{\int\int \Pr(R|h_{s,z}, x')\, p(x_{\mathrm{R}}|h_{s,z})\, \rho(s)\, \mathrm{d}z\mathrm{d}s}{\int\int p(x_{\mathrm{R}}|h_{s,z})\, \rho(s)\, \mathrm{d}z\mathrm{d}s} \\
&= \frac{\int\int_{x,x' \in h_{s,z}} s^{-m} \rho(s)\, \mathrm{d}z\mathrm{d}s}{\int\int_{x \in h_{s,z}} s^{-m} \rho(s)\, \mathrm{d}z\mathrm{d}s} \\
&= \int \Gamma_s(x, x')\, \rho(s)\, \mathrm{d}s \qquad (51)
\end{aligned}$$

where $\Gamma_s(x, x') = s^{-m} \int_{x,x' \in h_{s,z}} \mathrm{d}z$ is the proportion of CRs of size s containing x that also contain x'. In particular, when $m = 1$, Γ_s is the triangular function $\Gamma_s(x, x') = \max\left\{1 - \frac{|x-x'|}{s}, 0\right\}$. Equation 51 shows that the CR model's generalization gradient is a mixture of Γ_s for different s, with mixture weights given by the size distribution $\rho(s)$.

Shepard's (1987) CR model can be shown to be a special case of the continuous-features model, with the tuning curve defined by the shape of the CR. Indeed, define a tuning curve $r : \mathbb{R}^m \to \mathbb{R}$ as the indicator function for a generic CR of size $s = 1$:

$$r(d) = \begin{cases} 1 & z + d \in h_{1,z} \\ 0 & \text{otherwise,} \end{cases} \qquad (52)$$

where z is arbitrary. Defining features $f_{s,z}$ according to Equation 31, these features then satisfy

$$f_{s,z}(x) = \begin{cases} s^{-m} & x \in h_{s,z} \\ 0 & x \notin h_{s,z}. \end{cases} \qquad (53)$$

Next, let $c = \int s^m \rho(s)\, \mathrm{d}s$, and define a scale distribution for the continuous-features model by $p(s) = s^m \rho(s)/c$. We then obtain the following equivalence between the two models:

$$\begin{aligned}
\Pr(R|x_{\mathrm{R}}, x') &= \int\int_{x,x' \in h_{s,z}} s^{-m} \rho(s)\, \mathrm{d}z\mathrm{d}s \\
&= \int\int f_{s,z}(x)\, f_{s,z}(x')\, s^m \rho(s)\, \mathrm{d}z\mathrm{d}s \\
&= c \cdot \mathrm{sim}(x, x'). \qquad (54)
\end{aligned}$$

As dicussed above, the scalar c can be ignored when modeling similarity or generalization, because it can be absorbed into whatever operational measure is used to assess these psychological quantities. In conclusion, the CR model can be reinterpreted as a continuous-features model in which the features are all-or-none, with receptive fields corresponding to each candidate CR. The CR model's generalization gradient then matches the similarity function implied by that set of features. The transformation of the scale distribution between models, from $\rho(s)$ to $p(s)$, reflects the *size principle* (Tenenbaum & Griffiths, 2001): the fact that the likelihood of sampling x depends inversely on the size of the CR, thus making smaller CRs relatively more likely in the Bayesian model.

Building on this equivalence, we can offer a Bayesian interpretation of the continuous-features model for an arbitrary tuning curve r. Without loss of generality, we assume $\max r(x) = 1$.[10] For each feature $f_{s,z}$, define a hypothesis $h_{s,z}$ governing the probability of reward according to

$$\begin{aligned} \Pr(R|h_{s,z}, x) &= s^m f_{s,z}(x) \\ &= r\left(\tfrac{x-z}{s}\right). \end{aligned} \tag{55}$$

That is, if the environment operates according to $h_{s,z}$, then every time stimulus x is encountered the probability of reward is $s^m f_{s,z}(x)$. In the special case above corresponding the to CR model, the probabiliy of reward is always one or zero (Equation 53). Next, let the prior distribution for the hypotheses be defined by a prior $\rho(s) = cs^{-m}p(s)$ on the size and an improper uniform distribution on the location. When the subject encounters a stimulus x paired with reward, (s)he assumes that x was sampled with probability proportional to its likelihood of generating a reward, $\Pr(R|x)$. For example, the experimenter could have generated stimuli completely at random until one arose that produced a reward. Under this sampling process, the likelihood of any given stimulus being the one sampled is equal to

$$\begin{aligned} p(x_{\mathrm{R}}|h_{s,z}) &= \frac{\Pr(R|x, h_{s,z})}{\int \Pr(R|y, h_{s,z})\,\mathrm{d}y} \\ &= \frac{s^m f_{s,z}(x)}{\int s^m f_{s,z}(y)\,\mathrm{d}y} \\ &= f_{s,z}(x). \end{aligned} \tag{56}$$

When the subject later encounters a test stimulus, x', the posterior probability that it will produce a reward conditioned on the earlier observation of x_{R} is given by

$$\begin{aligned} \Pr(R|x_{\mathrm{R}}, x') &= \frac{\int\int \Pr(R|h_{s,z}, x')\, p(x_{\mathrm{R}}|h_{s,z})\, \rho(s)\,\mathrm{d}z\mathrm{d}s}{\int\int p(x_{\mathrm{R}}|h_{s,z})\, \rho(s)\,\mathrm{d}z\mathrm{d}s} \\ &= \frac{\int\int s^m f_{s,z}(x')\, f_{s,z}(x)\, \rho(s)\,\mathrm{d}z\mathrm{d}s}{\int\int f_{s,z}(x)\, \rho(s)\,\mathrm{d}z\mathrm{d}s} \\ &= c \cdot \mathrm{sim}(x, x'). \end{aligned} \tag{57}$$

In conclusion, the continuous-features model's similarity function can be interpreted as a Bayesian posterior probability that one stimulus will trigger reward given that another did, under a generating process wherein reward probability follows a known tuning curve (r) with unknown scale and location (s, z).

One question raised by this extension of Shepard's (1987) model is whether his universal law of generalization holds under arbitrary tuning curves. Shepard found that, for a wide variety of plausible size distributions $\rho(s)$, the generalization gradient resulting from the CR model is nearly exponential: $\Pr(R|x_{\mathrm{R}}, x') \simeq e^{-\alpha\|x-y\|}$. This finding was for all-or-none consequential regions and hence for all-or-none tuning curves in the continuous-features model (e.g., a boxcar for $n = 1$). It would be interesting to explore whether a similar result holds for other tuning curves. If so, then Shepard's universal law would be a robust property of the continuous-features model as well.

Finally, we note that the continuous-features model also has a Bayesian interpretation under another sampling procedure. Tenenbaum and Griffiths (2001) distinguished two assumptions a Bayesian model of generalization could make regarding how the initial (training) stimulus is sampled. The first possibility is that the stimulus was chosen to be one that produces the outcome in question. This is the assumption used in Shepard's (1987) model and in our analyses above, and it is the assumption under which Tenenbaum and Griffiths' size principle arises. The second possibility is that the training stimulus was determined fully at random and just happened to produce the outcome. Under this latter sampling process, the probability of observing reward following x is simply:

$$\Pr(x_{\mathrm{R}}|h_{s,z}) = \Pr(R|h_{s,z}, x). \tag{58}$$

[10] We can always replace $r(x)$ by $s^{-m}r(x/s)$ for any chosen s, and r will still integrate to unity and the resulting feature set $\{f_{s,z}\}$ will be unchanged. Thus the only assumption needed is that r is bounded, i.e. that the maximal activation of any feature is finite.

If we then define hypotheses by

$$\Pr\left(R|h_{s,z},x\right) \quad = \quad f_{s,z}\left(x\right), \tag{59}$$

as an alternative to (55), and we identify $\rho\left(s\right)=p\left(s\right)$, then the posterior becomes

$$\begin{aligned}\Pr\left(R|x_{\mathrm{R}},x'\right) &= \frac{\int\int \Pr\left(R|h_{s,z},x'\right) p\left(x_{\mathrm{R}}|h_{s,z}\right) \rho\left(s\right) \mathrm{d}z\mathrm{d}s}{\int\int p\left(x_{\mathrm{R}}|h_{s,z}\right) \rho\left(s\right) \mathrm{d}z\mathrm{d}s} \\ &= = \frac{\int\int f_{s,z}\left(x'\right) f_{s,z}\left(x\right) \rho\left(s\right) \mathrm{d}z\mathrm{d}s}{\int\int f_{s,z}\left(x\right) \rho\left(s\right) \mathrm{d}z\mathrm{d}s} \\ &= \mathrm{sim}\left(x,x'\right). \end{aligned} \tag{60}$$

Thus under the second sampling assumption the model has an even more direct Bayesian interpretation: similarity is equal to the posterior probability for generalization (with no scaling factor), the features are the same as the hypotheses' likelihood functions, and the scale density for the features is the same as that for the hypotheses.

4 Attention Learning

We have thus far established that the kernel duality provides an equivalence between learning models based on similarity and on features, and we have demonstrated with the continuous-features model how a feature representation can be explicitly derived that is dual to a similarity model based on generalization in a continuous stimulus space. The main thesis of this paper is that these equivalences provide a means for translating theoretical principles between similarity and feature-based modeling frameworks, thus affording greater insight into biological learning than is possible in either framework alone.

To demonstrate the utility of the duality and translation approach, this section presents a theoretical analysis of attention in learning. The proposal that associative learning is moderated by concomitant learning of attention has a long history within both similarity and feature frameworks (Mackintosh, 1975; Nosofsky, 1986). However, theories and formal models developed in these two frameworks differ fundamentally in the psychological mechanisms by which attention acts. Feature-based models of attention learning posit that attention affects learning rates, determining how rapidly associations can be learned for individual cues (Mackintosh, 1975). Similarity-based models of attention posit that attention affects generalization gradients, determining how learning about one stimulus influences responding to other stimuli (Nosofsky, 1986; Sutherland & Mackintosh, 1971). Despite these apparently fundamental differences, we show here that the kernel duality offers an interpretation wherein both forms of attention operate in exactly the same way, just acting on different representations. Specifically, both theories of attention learning can be interpreted as changing the scaling of stimulus representations, with attended dimensions being stretched and unattended dimensions shrunken. Rescaling the similarity space ($\mathcal{X}$) manifests in changes to the generalization gradient, and rescaling the feature space ($\mathcal{H}$) manifests in changes to learning rates. Moreover, we show in the continuous-features model how similarity-based attention to a dimension of the stimulus space is equivalent to shifting attention from the coarse to the fine-scale features associated with that dimension.

4.1 Attention in Feature Models

Consider a standard perceptron-style feature model of associative learning. The model maintains a weight vector $\mathbf{w}$ and, for a given input stimulus $\mathbf{x}$, generates a prediction equal to

$$v_{\mathrm{f}}\left(x\right) = \left\langle \mathbf{w},\mathbf{x}\right\rangle. \tag{61}$$

After feedback (t) is given, a prediction error is computed as the difference between actual and predicted outcomes,

$$\delta = t - v_{\mathrm{f}}\left(\mathbf{x}\right), \tag{62}$$

and $\mathbf{w}$ is updated according to gradient descent with a learning rate ε:

$$\begin{aligned}\Delta\mathbf{w} &= -\frac{\varepsilon}{2}\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\left(\delta^2\right) \\ &= \varepsilon\delta\mathbf{x}. \end{aligned} \tag{63}$$

Thus each weight is updated in proportion to the model's overall prediction error and to the activation of the corresponding feature (which determines how much that weight contributed to the prediction).

This learning rule shows why the inner product between stimuli determines the generalization function for feature models (Ghirlanda, 2015; Jones & Sieck, 2003). Given two stimuli presented in sequence, $\mathbf{x}$ then $\mathbf{x}'$, consider how learning following $\mathbf{x}$affects responding to $\mathbf{x}'$:

$$\begin{aligned}\Delta v_{\mathrm{f}}\left(\mathbf{x}'\right) &= \left\langle \Delta\mathbf{w}, \mathbf{x}'\right\rangle \\ &= \left\langle \varepsilon\delta\mathbf{x}, \mathbf{x}'\right\rangle \\ &= \varepsilon\delta\left\langle \mathbf{x}, \mathbf{x}'\right\rangle. \end{aligned} \tag{64}$$

Therefore generalization between stimuli is proportional to their inner product, and hence proportional to the kernel that is dual to the model's feature representation.

The learning rule in Equation 63 was first introduced in Rescorla & Wagner's (1972) model of associative learning. That model also assumed the features could vary in salience, with faster learning for more salient features. With some simplification of notation, each feature f_i is associated with a different learning rate, ε_i, and the update for the weight component $\mathbf{w}_i$ is given by

$$\begin{aligned}\Delta\mathbf{w}_i &= \frac{\varepsilon_i}{2}\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}_i}\left(\delta^2\right) \\ &= \varepsilon_i\delta\mathbf{x}_i. \end{aligned} \tag{65}$$

Subsequent work by Mackintosh (1975) reinterpreted the ε_i as reflecting attention, and proposed that attention itself is learnable from feedback. [11]Thus once a subject learns that a given cue is relevant in a task environment, he or she can shift attention to that cue and thus acquire new associations for that cue more rapidly (see Le Pelley, Mitchell, Beesley, George, & Wills, in press, for a review of empirical evidence).

Although this theory of attention is traditionally interpreted in terms of learning rates, it has a formally equivalent interpretation based purely on the stimulus representation. To see this, we can reparameterize the model with a new stimulus representation $\tilde{\mathbf{x}}$ such that

$$\tilde{\mathbf{x}}_i = \sqrt{\varepsilon_i}\mathbf{x}_i. \tag{66}$$

In other words, each feature (viewed as a function on the stimulus space) is scaled up by a factor of $\sqrt{\varepsilon_i}$. Define the reparameterized weight vector, $\tilde{\mathbf{w}}$, by

$$\tilde{\mathbf{w}}_i = \frac{1}{\sqrt{\varepsilon_i}}\mathbf{w}_i. \tag{67}$$

Then the new weight and feature vectors generate the same prediction as before:

$$\left\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}\right\rangle = \left\langle \mathbf{w}, \mathbf{x}\right\rangle = v_{\mathrm{f}}\left(x\right). \tag{68}$$

The learning rule under the new parameterization is

$$\begin{aligned}\Delta\tilde{\mathbf{w}}_i &= \frac{1}{\sqrt{\varepsilon_i}}\Delta\mathbf{w}_i \\ &= \sqrt{\varepsilon_i}\delta\mathbf{x}_i \\ &= \delta\tilde{\mathbf{x}}_i. \end{aligned} \tag{69}$$

[11]In Mackintosh's model, learning is based on the prediction from each feature alone, rather than on joint prediction error. This difference is important for understanding the relationship between associative learning in feature vs. similarity models, but it does not affect our analysis of attention learning here.

Thus the learning rate equals 1 for all features. Instead of modifying the learning rates, attention acts to rescale the feature space, such that attended feature dimensions are stretched out and unattended feature dimensions are compressed (by factors of $\sqrt{\varepsilon_i}$). In terms of the stimulus representations (i.e., feature vectors), differences between stimuli on attended feature dimensions become magnified whereas differences on unattended feature dimensions become attenuated.

It is also instructive to derive the same equivalence from the opposite direction, in part because it shows how the theory can be naturally generalized. Begin with a model with a learning rate of 1 for all features (i.e., obeying Equation 69), and denote its stimulus representation and weight vector respectively by $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{w}}$. Let T be any invertible self-adjoint linear operator on the feature space, meaning that $\langle T\mathbf{y}, \mathbf{z}\rangle = \langle \mathbf{y}, T\mathbf{z}\rangle$ and $\left\langle T^{-1}\mathbf{y}, \mathbf{z}\right\rangle = \left\langle \mathbf{y}, T^{-1}\mathbf{z}\right\rangle$ for any feature vectors (i.e., stimulus or weight vectors) $\mathbf{y}$ and $\mathbf{z}$. In the finite-dimensional case, this means that T is an invertible symmetric matrix. Now define a reparameterized model with stimulus representation $\mathbf{x} = T^{-1}\tilde{\mathbf{x}}$, meaning

$$\tilde{\mathbf{x}} = T\mathbf{x}, \tag{70}$$

and weight vector

$$\mathbf{w} = T\tilde{\mathbf{w}}. \tag{71}$$

In the construction above (Equations 66 & 67), T is a diagonal matrix with

$$T_{ij} = \begin{cases} \sqrt{\varepsilon_i} & i = j \\ 0 & i \neq j. \end{cases} \tag{72}$$

Given these relationships between the two parameterizations, the $(\mathbf{x}, \mathbf{w})$ representation generates the same prediction as the $(\tilde{\mathbf{x}}, \tilde{\mathbf{w}})$ representation,

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}\rangle &= \langle T\tilde{\mathbf{w}}, \mathbf{x}\rangle \\ &= \langle \tilde{\mathbf{w}}, T\mathbf{x}\rangle \\ &= \langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}\rangle, \end{aligned} \tag{73}$$

and the learning rule for $\mathbf{w}$ is given by

$$\begin{aligned} \Delta\mathbf{w} &= \Delta\left[T\left(\tilde{\mathbf{w}}\right)\right] \\ &= T\left(\Delta\tilde{\mathbf{w}}\right) \\ &= T\left(\delta\tilde{\mathbf{x}}\right) \\ &= T^2\delta\mathbf{x}. \end{aligned} \tag{74}$$

Thus we can think of T as an attention operator, with dual interpretations. Under the $(\mathbf{x}, \mathbf{w})$ representation, T can be interpreted as directly modifying the learning process, such that the update is multiplied by T^2 (Equation 74). That is, the learning rate along each eigenvector of T is scaled by the square of its corresponding eigenvalue. Under the $(\tilde{\mathbf{x}}, \tilde{\mathbf{w}})$ representation, T can be interpreted as operating on the feature representation, effecting a transformation of the feature space via Equation 70, with concomitant changes to the metric of the space. That is, each eigenvector of T is rescaled by its eigenvalue. Under either interpretation, T acts to emphasize (stretch) all eigenvectors having eigenvalues greater than unity, and to de-emphasize (compress) all eigenvectors having eigenvalues less than unity. Under the classical formulation (Equations 65 & 72), T is diagonal and the eigenvectors are the features themselves. In the general case, the eigenvectors can be an arbitrary set of (mutually orthogonal) linear combinations of features, and T implements attention to those linear combinations.

4.2 Attention in Similarity Models

Formal theories of attention learning in similarity models are founded in multidimensional-scaling approache to stimulus representation (Nosofsky, 1992). Each stimulus is modeled as a point $x = (x_1, \ldots, x_m)$ in an m-dimensional stimulus space $\mathcal{X}$, and these coordinate representations determine similarity. (Again, we stress that this multidimensional representation should not be confused with the feature representation

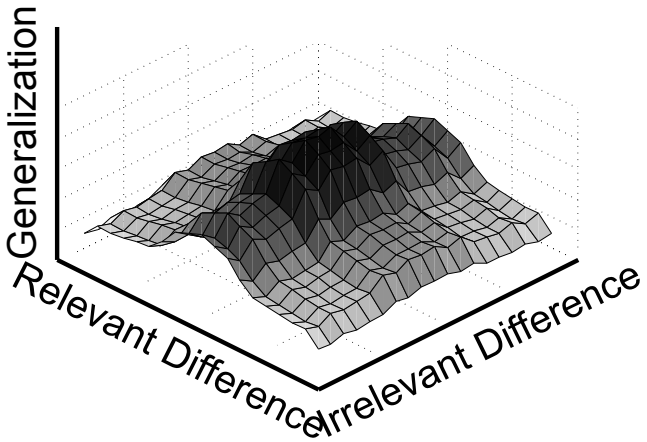

Figure 2: Generalization gradient from a two-dimensional categorization task, in which only one dimension is relevant to category membership. Generalization is seen to be broader along the irrelevant dimension, in line with the generalization theory of attention (Nosofsky, 1986). Based on data from Jones, Maddox, and Love (2005) and the sequential-effects technique for measuring generalization from Jones, Love, & Maddox (2006).

$\mathbf{x} = (\mathbf{x}_i)_{i \in \mathcal{F}}$ that is equivalent to the similarity function under the kernel duality.) Similarity is assumed to decrease with distance along each dimension, usually following a Gaussian or an exponential function. The overall similarity between two stimuli is the product of the contributions from all dimensions:

$$\operatorname{sim}(x, x') = \prod_{i=1}^{m} \exp\left(-\alpha_i \left(x_i - x_i'\right)^p\right) = \exp\left(-\sum_{i=1}^{m} \alpha_i \left(x_i - x_i'\right)^p\right), \tag{75}$$

where p equals 1 (exponential generalization) or 2 (Gaussian). The attention parameters, α_i, determine how much each dimension contributes to similarity, or equivalently how rapidly similarity drops off with differences on each dimension. A large value of α_i produces little generalization across dimension i, because stimuli differing on that dimension are highly dissimilar, whereas a value of α_i close to zero produces a broad generalization gradient along that dimension. These predictions accord with empirical generalization gradients in animals (see Honig & Urcuioli, 1981, for a review) and in humans (Jones, Maddox, & Love, 2005; Figure 2).

A common way to conceptualize this theory of attention is as a rescaling of the stimulus space (e.g., Kruschke, 1992). That is, similarity is thought of as a function of distance,

$$\operatorname{sim}(x, x') = e^{-d(x,x')^p}, \tag{76}$$

where overall distance is either a city-block ($p = 1$) or Euclidean ($p = 2$) function of distance on the individual dimensions:

$$d(x, x') = \left[\sum_{i=1}^{m} d_i (x, x')^p\right]^{1/p}. \tag{77}$$

Under this view, attention determines the metric on each dimension:

$$d_i(x, x') = \alpha_i \left|x_i - x_i'\right|. \tag{78}$$

Therefore increasing attention to a stimulus dimension amounts to stretching that dimension, and decreasing attention amounts to shrinking it.

4.3 Translating Theories of Attention between Frameworks

The kernel duality offers two means to understand the relationship between feature- and similarity-based theories of attention in learning. First, the two theoretical mechanisms can be viewed as operating on two

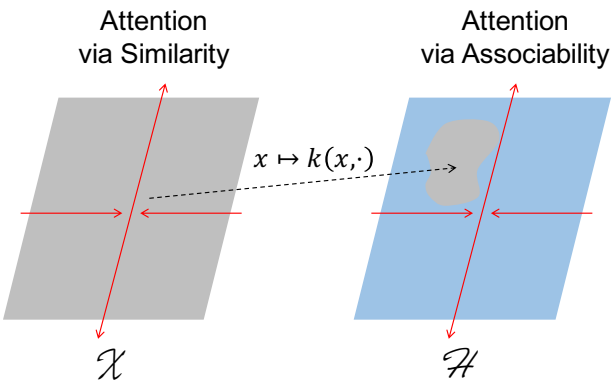

Figure 3: Schematic summary of the parallel between generalization and associability theories of attention, as revealed by the kernel duality. Attention acting on similarity or generalization is equivalent to rescaling the similarity space (grey space at left). Under the kernel framework, the similarity space can be embedded in the reproducing-kernel Hilbert space (blue space at right) by a nonlinear mapping that identifies each stimulus x with its evaluation function $k(x,\cdot)$. Attention acting on associability is equivalent to rescaling of the Hilbert space, by rescaling the features that form a basis for that space. Under both mechanisms, attended dimensions are stretched and unattended dimensions are compressed (red arrows). Thus the two theories of attention correspond to identical operations acting on different stimulus representations.

different (yet dual) representations. Attention via associability acts on the feature representation ($\mathcal{H}$ and $\mathcal{F}$), and attention via generalization gradients acts on the similarity representation ($\mathcal{X}$ and k). As the previous two subsections have shown, both of these mechanisms amount to rescaling their respective representations, by stretching the attended dimensions and compressing the unattended ones. Figure 3 offers a schematic illustration of this connection. We consider this result fairly remarkable, because on the surface these two theories of attention seem entirely different, one concerning learning rates and other other concerning breadth of generalization gradients.

Second, the kernel duality can be used to translate each attention mechanism into the other framework. That is, the associability theory of attention can be recast as a change in the kernel, and the generalization theory of attention can be recast as a change in the features.

One advantage of the reparameterization in Section 4.1, recasting the associability theory of attention as rescaling of the feature space, is that it enables derivation of the equivalent kernel. Under the original feature representation (i.e., before a shift of attention), the corresponding kernel is defined by

$$k(x,x') = \langle \mathbf{x}, \mathbf{x}' \rangle . \tag{79}$$

Under the influence of attention, we can consider the stimulus representation to be modified according to Equation 66, and thus the new kernel is defined by

$$\begin{aligned} k(x,x') &= \langle \tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \rangle \\ &= \sum_{i \in \mathcal{F}} \varepsilon_i \mathbf{x}_i \mathbf{x}'_i. \end{aligned} \tag{80}$$

This result can also be seen by following the reasoning above (Equation 64), to calculate the amount of generalization from $\mathbf{x}$ to $\mathbf{x}'$ when they are presented successively:

$$\begin{aligned} \Delta v_{\mathrm{f}}(\mathbf{x}') &= \langle \Delta \mathbf{w}, \mathbf{x}' \rangle \\ &= \delta \sum_{i \in \mathcal{F}} \varepsilon_i \mathbf{x}_i \mathbf{x}'_i. \end{aligned} \tag{81}$$

Thus the associability theory of attention is equivalent to weighting each feature (f_i) by its learning rate (ε_i) to determine overall similarity between stimuli.

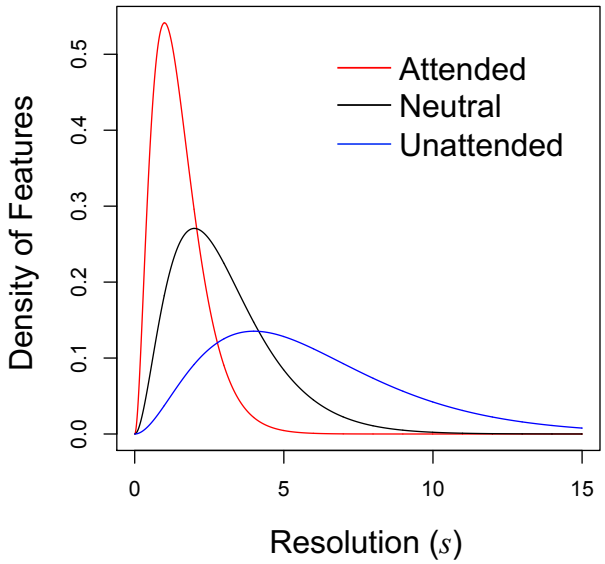

Figure 4: Distribution of features at different resolutions, for different levels of similarity-based attention. Similarity is defined by an exponential generalization gradient, and features are from a continuous-features model that is dual to the similarity model. Attention parameter is set at 1 for a neutral feature, 2 for an attended feature, and $\frac{1}{2}$ for an unattended feature.

Going in the other direction, we can use the continuous-features model to explore how similarity-based changes in attention translate to changes in the dual feature space. As shown in Section 3.2, given a family of features of a given shape (the tuning curve, r) distributed throughout the stimulus space, a unidimensional generalization gradient can be translated into a density distribution for features at different scales. Therefore changes in the generalization gradient can be modeled as changes in the feature density distribution. Take for example the exponential generalization gradient of Shepard (1957, 1987), given in Equation 75 with $p = 1$. From Equation 42 we saw that this similarity function is dual to a continuous-features model with a boxcar tuning curve and feature resolutions (s) following a Gamma distribution:

$$s \sim \text{Gamma}\,(3, \alpha)\,. \tag{82}$$

The rate parameter in this distribution, α, is the attention parameter in the generalization gradient. As this parameter is increased, the distribution shifts toward finer scales, and as it is decreased the distribution shifts towards coarser scales (Figure 4).

In a model of separable dimensions, we would assume a separate family of features for each dimension of the stimulus space ($\mathcal{X}$), with overall similarity given by Equation 48. Increasing attention to some dimensions and decreasing it to others would correspond to weighting the fine-scale features for the former and the coarse-scale features for the latter. Because the continuous-features model embodies a mathematical idealization of a continuous set of features (i.e., a feature $f_{s,z}$ for all z), the reweighting can be interpreted in several ways. First, it can be viewed as changing the sampling density of features, with $p\,(s)$ indicating the number of features $f_{s,z}$ within any given interval of values of z. Thus attention changes the population of features, with more features at more densely sampled resolutions and fewer features at less densely sampled resolutions. Second, the reweighting can be viewed as keeping a fixed population of features and rescaling their activation functions. Thus the value of each feature $f_{s,z}$ would be multiplied by $\sqrt{p\,(s)}$ (see Equation 32). Third, the features could be entirely unchanged, and attention could act on their learning rates. Thus the learning rate for each feature weight $\mathbf{w}_{s,z}$ would be proportional $p\,(s)$. In summary all four of these views of attention—rescaling the input dimensions to similarity, shifting the sampling density of fine- and coarse-scale features on those dimensions, rescaling the activation values of those features, and adapting their associative learning rates—are equivalent. Under the duality approach advocated here, they are all different ways of modeling the same underlying biological system.

5 Discussion

- summary of contribution
- other applications
  - translation of associative learning rules [I have several results that could be summarized]
  - translation of attention learning rules [several open questions here]
- extensions
  - asymmmetric similarity functions
  - Banach space representation and semi-inner product
  - additional duality of reference vs. comparison stimulus

References

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

[2] Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94-107.

[3] Fréchet, M. (1907). Sur les ensembles de fonctions et les opérations linéaires. *Comptes rendus de l'Académie des sciences, 144*, 1414-1416.

[4] Fredholm, E. I. (1903). Sur une classe d'equations fonctionnelles. *Acta Mathematica, 27*, 365-390.

[5] Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

[6] Ghirlanda, S. (2015). On elemental and configural theories of associative learning. *Journal of Mathematical Psychology, 64-65*, 8-16.

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

[8] Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology, 51*, 79-88.

[9] Honig, W. K., & Urcuioli, P. J. (1981). The legacy of Guttman and Kalish (1956): 25 years of research on stimulus generalization. *Journal of the Experimental Analysis of Behavior, 36*, 405-445.

[10] Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51*, 343-358.

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

[12] Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review, 121*, 1-32.

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

[14] Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*, 316-332.

[15] Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 405-410.

[16] Jones, M. & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*, 626-640.

[17] Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior, 1967: Aversive stimulation* (pp. 9-31). Miami: University of Miami Press, 1968.

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

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

[20] Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103-189). New York: Wiley.

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

[22] Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

[23] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A, 209*, 415-446.

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

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

[26] Nosofsky, R.M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43*, 25-53.

[27] Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87-108.

[28] Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

[29] Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, 61–73.

[30] Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review, 101*, 587–607.

[31] Pearce, J. M. and Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52*, 111–139.

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

[33] Riesz, F. (1907). Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *Comptes rendus de l'Académie des sciences, 144*, 1409–1411.

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

[35] Scholköpf, B., & Smola, A. J. (2002). *Learning with kernels*. MIT Press.

[36] Shawe-Taylor, J. & Cristianini N. (2004). *Kernel methods for pattern analysis*. Cambridge, MA: Cambridge University Press.

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

[38] Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika, 27*, 125–140.

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

[40] Shepard, R. N. (1987). Toward a universal law of general- ization for psychological science. *Science, 237*, 1317-1323.

[41] Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence, 46*, 159-216.

[42] Sutherland, N., & Mackintosh, N. (1971). *Mechanisms of animal discrimination learning*. New York, NY: Academic Press.

[43] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

[44] Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629-641.

[45] Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

[46] Zhang, H. and Zhang, J. (2012). Regularized learning in Banach space as an optimization problem: Representer theorems. Journal of Global Optimization, 54: 235-250