Some Theoretical Aspects of Human Categorization Behavior: Similarity and Generalization

Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der Fakultät für Biologie
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von
Frank Jäkel
aus Iserlohn, Deutschland

Tag der mündlichen Prüfung:	9. November 2007
Dekan der Fakultät für Biologie:	Prof. Dr. F. Schöffl
Dekan der Medizinischen Fakultät:	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. F. A. Wichmann
2. Berichterstatter:	Prof. Dr. R. Ulrich
3. Berichterstatter:	Prof. Dr. A. Diederich
Prüfungskommission:	Prof. Dr. A. Diederich
	Prof. Dr. H. A. Mallot
	Prof. Dr. B. Schölkopf
	Prof. Dr. R. Ulrich
	Prof. Dr. F. A. Wichmann

To my parents.

I am indebted to many people who have provided support, help and encouragement over the last couple of years—without them this thesis would not have materialized. Above all I need to thank Felix Wichmann for his supervision and thoughtful advice on all aspects of this work, for teaching me how to run proper psychophysical experiments, and for the right encouragement at the right time. Very special thanks go to Bernhard Schölkopf for introducing me to kernel methods. Without his vision and without his support I would not have started working on this topic. It was a great pleasure and a great opportunity to work in his department that offers a unique and truly interdisciplinary environment for studying learning and perception in man and machine. Many thanks go to all members of AGBS for making the department a place where ideas and insights are happily shared and for making it a fun place to work at. Theresa Cooke, Dilan Görür, Malte Kuss, and I worked together on various projects. Jan Eichhorn, Bruce Henning, Jakob Macke, and Florian Steinke were kind enough to read large parts of this thesis at various stages and provided many helpful comments.

Contents

Summary

Explanations of human categorization behavior often invoke similarity. Stimuli that are similar to each other are grouped together whereas stimuli that are very different are kept separate. Despite serious problems in defining similarity, both conceptually and experimentally, this is the prevailing view of categorization in prototype theories (Posner & Keele, 1968; Reed, 1972) and exemplar theories (Medin & Schaffer, 1978; Nosofsky, 1986). This is also the prevailing approach in machine learning. A popular class of methods in machine learning is based on the idea of modeling the similarity of patterns by a kernel (Schölkopf & Smola, 2002). Many of these methods are akin to exemplar models in psychology, as they also base the categorization on a comparison with stored examples with known category labels. In this thesis, we re-examine the notion of similarity as it is used in models for human categorization behavior from a machine learning perspective.

Our current understanding of many machine learning methods has been deepened considerably by the realization that similarity can be modeled as a so-called positive definite kernel. One of the most commonly used similarity measures in psychology, Shepard's universal law of generalization (Shepard, 1987), is shown to be such a positive definite kernel. This observation opens up the possibility to use tools from functional analysis, that are also used in machine learning, in the analysis of psychological similarity. Two important theoretical insights about similarity are gained from such an analysis.

First, early models of similarity introduced the notion of a psychological space with a Euclidean metric that represented the similarity of stimuli (Torgerson, 1952; Ekman, 1954). Shepard's early work on multidimensional scaling can be understood as an effort to overcome the assumption that the similarity of stimuli is captured by a Euclidean metric (Shepard, 1962). The later introduction of the universal law of generalization was the culmination of work that happened over several decades and summarized the relationship between similarity and metrics in many psychological spaces (Shepard, 1987). Ironically, however, this thesis demonstrates that the universal law leads to an embedding of similarity into a Euclidean space and therefore means a return to those roots of multidimensional scaling that Shepard tried to overcome.

Second, models for similarity that are based on multidimensional scaling have been heavily criticized by Tversky and coworkers (Beals, Krantz, & Tversky, 1968; Tversky, 1977; Tversky & Gati, 1982). The most severe criticism concerns the triangle inequality which all metric models of similarity assume. Despite this criticism scaling methods have been used with great success, especially in categorization research. Even if the criticism is acknowledged researchers usually proceed with scaling without much hesitation (Nosofsky, 1986). Still, Tversky and Gati (1982) reported data that seemed to show that multidimensional scaling cannot capture many human similarity judgments. However, their tests of the triangle inequality also assumed segmental additivity. For Tversky and Gati segmental additivity was an essential property of any geometric model of similarity and therefore also for multidimensional scaling. Here, it is shown that there are theoretically well-motivated metrics—induced by Shepard's law of generalization and implicitly used in many multidimensional scaling scenarios—that do not have the property of segmental additivity. These metrics are therefore not affected by Tversky's criticism and provide a post-hoc justification for the use of multidimensional scaling for data that seem to violate the triangle inequality. In fact, these metrics provide a theoretically well-justified model for stimulus similarity that are also bounded from above, thereby implementing the intuition that stimulus similarity is best defined locally (Indow, 1994).

As Shepard's law is used extensively in psychological models of categorization (Nosofsky, 1986; Kruschke, 1992; Love, Medin, & Gureckis, 2004) the insight that similarity can be modeled as a positive definite kernel can also benefit a theoretical analysis of categorization behavior. Exemplar theories in particular make heavy use of positive definite kernels. Here, it is shown that exemplar models in psychology are closely related to kernel logistic regression (Hastie, Tibshirani, & Friedman, 2001). The link between kernel logistic regression and exemplar theories is their use of radial-basis-function neural networks (Poggio & Girosi, 1989; Poggio, 1990).

A traditional concern against exemplar theories is their lack of an abstraction mechanism that seemingly limits their generalization performance (Smith & Minda, 1998, 2000). However, kernel logistic regression is used successfully in many applications in machine learning. Using insights from kernel methods a first analysis of the generalization ability of exemplar models is provided. It is found that exemplar theories in psychology are indeed prone to overfitting, i.e. they show poor generalization performance. However, like their relatives in machine learning exemplar models can be equipped with regularization mechanisms that are known to improve generalization performance under real-world category learning conditions. Hence, despite concerns from prototype theorists about the generalization ability of exemplar models, exemplar models can be made to reliably extract the structure inherent in real-world categories by using techniques from machine learning.

CHAPTER 1

Introduction

Categorization is arguably one of the most fundamental cognitive processes. It serves, for example, as a crucial link between perception and high level cognition. For many cognitive skills, like reasoning, induction, and language, categorization is a prerequisite. In fact, the meaning of simple nouns, such as ‘animal’, ‘dog’, ‘apple’, or ‘psychologist’ seems to be defined by their respective categories. Even very basic object properties, like ‘being red’, are already a categorical response to an analogue perceptual input. Our ability to categorize allows us to partition the world into groups of objects that can be treated alike, if only for a certain purpose. By categorizing an animal into the category ‘dog’, many perceived details like the exact size, color or fur are discarded and a more compact, easily transmittable representation is obtained. Without seeing the dog one can infer that it probably barks, has teeth, and can bite. By only knowing an object’s category label a wealth of useful information about the object is available and predictions about its behavior can be made.

As categorization is apparently fundamental to our cognitive world—“concepts seem to be the very stuff of which cognitions are made” (Rey, 1983/1999, p. 279)—it is not a big surprise that research on categorization has received a lot of attention from philosophers, linguists and psychologists alike. Despite this attention we still do not have a clear understanding of the psychological (not to speak of the neural) processes underlying categorization behavior.

One of the problems of categorization research is that categorization happens at many, if not all, cognitive levels. Even at the perceptual level there are already effects of category boundaries, e.g. in color perception or speech perception (Harnad, 1987). There are perceptual categories that seem almost atomic, not analyzable into more basic categories—‘being red’ would be a prime example for this sort of category. At the other extreme are ad-hoc categories such as ‘things to take with you on a camping trip’ (Barsalou, 1983). Formation of these categories depends on many high-level cognitive abilities, like language, imagination, knowledge, memory, and common-sense reasoning.

In the so-called classical theory of concepts complex concepts are made up from atomic perceptual terms by logical combination (Laurence & Margolis, 1999), e.g. an ‘apple’ is ‘round’ and ‘red’, but some apples are ‘green’, and so on. A possibly very long list of perceptual primitives that are combined by logical operations defines a category. One big problem of this approach is that for many concepts it proved to be incredibly hard to really specify necessary and sufficient conditions that determine category membership. Of course, once you have defined some concepts nothing is stopping you from using those to define more complex concepts—and in fact it seems obvious that many concepts are composed in this way. The principle of compositionality is at the heart of the classical theory and it is clear that any complete theory of concepts also needs to address this aspect of categorization. Therefore, early psychological research on categorization has often concentrated on concepts that can be defined by logical formulas and this approach is still alive today (e.g. Bourne & Restle, 1959; Shepard, Hovland, & Jenkins, 1961; Feldman, 2000).

There is, however, more to the relation between concepts than the fact that they can be combined in an arbitrary way to compose new concepts. Many concepts seem to form hierarchies that reflect our knowledge about the structure of the world. Take for example biological taxonomies: Dalmatians and poodles can be grouped into the category dogs and together with cats, dolphins, apes, etc. they form the superordinate category of mammals. The basis for grouping all mammals together is similar to the basis for grouping 'things to take on a camping trip' together: We can find reasons why they should be grouped together. In fact, the more reasons we can find and the better these reasons are embedded into a network of related concepts—a theory of the objects under consideration—the more coherent a concept appears. This account of categorization is called theory-theory (Murphy & Medin, 1985/1999).

Within a taxonomy—for example the following hierarchy of concepts: sports car, car, vehicle—the intermediate level stands out psychologically. This is the level at which things are usually named, where objects have many properties in common, a similar shape, and similar motor actions associated with them. Categories at this level have been termed basic-level categories by Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976). At the level of basic categories it seems to be perception, rather than theoretical reasoning, that drives categorization. Things are grouped together because they are perceptually similar to each other. In this thesis we will mostly deal with categorization at this level. We will analyze fairly reduced, perceptually mediated categorization behavior (not all of the studies we will discuss are strictly at the basic level but all of them are based on perceptual similarity). While concepts that are not basic categories are certainly important for cognition it does not seem unreasonable to us that more complicated concepts are built on top of perceptually mediated, basic-level categories. Furthermore, very different mechanisms, including conscious reasoning, might be needed to explain the compositionality and the theoretical coherence of these concepts. To avoid confusions right from the start, in this thesis we will concentrate on the perceptual aspects of similarity and how similarity might give rise to categorization behavior.

1. Generalization and categorization

The reason why basic-level categorization is so useful—and why it works in the first place—is probably that there are natural categories in the world (Rosch, 1973; Anderson, 1991). Basic-level categories seem special because they are not just mediated by our perception but also reflect the structure of the world. Basic-level categories not only form clusters in our heads but also in the world. Of course, it can happen that categories in our heads do not match the categories in the world. However, a category is the more useful the more it is possible to transfer experience with one object of a class to another object of the same class—and this is only possible if there is some reality to the categories that we form. Also superordinate biological categories, like mammals, are real and are therefore useful for structuring our knowledge about the world. But mammals are not easily perceived as being mammals (a dolphin looks like a fish and not a mammal). Instead there seems to be a theory-driven insight at work that allows us to group all mammals together. This is in contrast to categories at the basic level where the perceptual experiences generalize effortlessly to other objects of the same kind. In this sense, basic-level categorization is really about generalization of perceptual experiences.

There is a second reason why categorization and generalization are deeply intertwined. Let us assume for simplicity that learning of a category proceeds by seeing examples of objects that either belong or do not belong to the category. The learner tries to assign the right category label, for example by uttering the category name. There might be a teacher who immediately corrects any wrong utterances. Feedback might also be provided by the environment directly, for example if a poisonous fruit is mistaken for an edible fruit. In any case the learner has to decide which features of the stimuli she encounters are crucial and which are accidental. She has to find the structure that is common to all the objects in a category. But the learner does never encounter all instances of a category. So how can she generalize from a possibly very small number of instances to the full category? This is of course the century-old problem of induction.

2. Categorization and similarity

A common-sense answer to the problem of category induction involves similarity. The reason why objects are grouped together in categories is that they are similar to each other. However, the appeal of invoking similarity in the explanation of categorization behavior merely stems from the need to generalize. A response will generalize to a new stimulus if the stimulus is similar enough to a stimulus with this response. Similar stimuli give rise to the same response. But how do we know that stimuli are similar? They are similar because the response to one stimulus will generalize to the other stimulus. Such a definition is of course circular but it seems hard to define the similarity of stimuli without invoking the response that they illicit. A definition of similarity is needed that does not depend on the generalization of responses (Bush & Mosteller, 1951; Shepard, 1987).

Despite its intuitive appeal as an explanatory construct similarity is a slippery concept and may be too flexible to provide the "glue that makes a category learnable and useful" (Murphy & Medin, 1985/1999, p. 427). One problem is that similarity can be extremely context and task dependent. If asked for the similarity of two arbitrary stimuli human participants will usually wonder "in what respect" (Medin, Goldstone, & Gentner, 1993). Still, problematic as similarity may be we can investigate how exactly assessment of similarity changes with context and task.

In fact, similarity based models of categorization have been extremely successful and while controversial the notion of similarity proved to be a fruitful concept to start with (e.g. the work collected in Hahn & Ramscar, 2001). Two classes of models about categorization stand out as relying almost completely on similarity: Prototype models and exemplar models. In prototype models it is assumed that subjects extract a summary representation, the prototype, from all the instances of a category that they encounter. When categorizing a new stimulus the similarity to this prototype is assumed to be the crucial factor. There is a plethora of studies that used the similarity of stimuli to a learned prototype as an explanation for categorization behavior (e.g. Posner & Keele, 1968; Franks & Bransford, 1971; Reed, 1972; Smith & Minda, 1998; Minda & Smith, 2001). Exemplar models, on the other hand, assume that there is no abstract representation like a prototype but instead suggest that categorization is mediated by memorization of exemplars. Exemplars with known category membership are stored in memory and new stimuli are categorized by assessing their similarity to the stored exemplars (e.g. Medin & Schaffer, 1978; Nosofsky, 1986; Kruschke, 1992). Prototype and exemplar models both rely on similarity. One crucial ingredient for the success of exemplar models, however, has been the realization that similarity is not a fixed concept. It varies with task and experimental contexts. Hence, experimental studies that described such changes were a major influence (Tversky, 1977; Nosofsky, 1986; Medin et al., 1993). Recently, there has also been some interest in how object similarity is mediated by the different senses and how the different sensory modalities affect similarity judgments and categorization (Cooke, Jäkel, Wallraven, & Bülthoff, 2007).

3. Similarity and kernels

The starting point for this thesis were two observations. First, in machine learning many methods for categorization also depend on similarity. Second, a popular mathematical tool for describing similarity in machine learning are so-called kernels. Incidentally, one of the most popular similarity measures in psychology, Shepard's universal law of generalization (Shepard, 1987), is also such a kernel. Therefore, some of the mathematical machinery and, more importantly, some statistical insights of machine learning might be applicable to the many psychological models that are built on Shepard's law (e.g. Nosofsky, 1986; Kruschke, 1992; Love et al., 2004). Briefly, it turned out that especially the notions of a positive definite kernel and of a reproducing kernel Hilbert space were useful in providing insights about similarity. Chapter 2 of this thesis provides an in-depth introduction to the mathematics involved while trying to motivate their use from a psychological and neural networks perspective.

In order to explain what kernels can add to the understanding of similarity we have to be a bit more specific about the actual models of similarity that we consider. Unfortunately, there are many ways how psychological similarity can be modeled and we are far from having reached a consensus on which model is to be preferred in what situation (Navarro, 2002). However, we think it is fair to say that the most influential approach has been geometrical in nature and is deeply connected with the method of multidimensional scaling (MDS). In early studies on MDS it was assumed that stimuli are represented as points in a multidimensional space and that the similarity between stimuli can be modeled as the Euclidean distance between the respective points (Torgerson, 1952). The closer two points are in space the more similar are the respective stimuli. From the beginning the assumptions of a Euclidean space seemed overly restrictive but they were overcome by the development of ordinal scaling methods (Shepard, 1962; Kruskal, 1964). These methods could deal with non-Euclidean spaces and also with a non-linear relationship between the distance in the embedding space and the measured similarity. In a very influential paper Shepard (1987) could unify many data-sets that were analyzed by ordinal scaling methods by postulating an exponential relationship between the distance in the embedding space and measured generalization performance. As generalization performance is a popular similarity measure, especially in categorization research, the exponential law is as much a law of generalization as it is of stimulus similarity.

In this thesis we analyze how the exponential law enters categorization models as a measure for stimulus similarity. We find that the way similarity is modeled gives rise to a positive definite kernel. As the similarity is a positive definite kernel it follows that any measured similarity matrix is assumed to be positive semidefinite. This means that the similarity matrix can be embedded in a Euclidean space. Ironically, Shepard's finding of the exponential as a link between similarity and distance has led to a theory that adheres to the same restrictions of Euclidean space that he tried to overcome by ordinal scaling methods. These theoretical results are explained in detail in Chapter 3 of this thesis.

Geometric models of similarity based on MDS have been very popular but also very much criticized. The most fundamental criticism has been put forward by Tversky and coworkers (Beals et al., 1968; Tversky, 1977; Tversky & Gati, 1982). Especially the triangle inequality—essential to MDS procedures—has been met with considerable skepticism. By using the finding that Shepard's law gives rise to a positive definite kernel we give this debate a new twist. Briefly, the issue is that Tversky and colleagues always assumed the triangle inequality in conjunction with a second property called segmental additivity. Our theoretical analysis shows that there is a natural metric associated with Shepard's law (the distance in the respective reproducing kernel Hilbert space) that does not have this property of segmental additivity and therefore avoids the serious criticism of the triangle inequality. On top of this we also find that this metric is bounded from above. This is a highly desirable property for a psychological metric because stimuli far apart in a psychological space can probably not get more dissimilar than "completely different" (Indow, 1994). Furthermore, we find that our analysis leads to an interpretable representation where all stimuli are represented by their similarity to all other stimuli (Edelman, 1998). These results are also described in Chapter 3.

4. Kernels and exemplar models

The analogy between categorization models in psychology and categorization algorithms in machine learning can be taken even a bit further than just saying that both build on similarity. In fact, there is a formal correspondence between exemplar models and a kernel method called kernel logistic regression.

At the very least a quantitative model for human categorization behavior needs to be able to predict the probability that a participant will respond to a stimulus with a certain category label. As mentioned before, a participant is assumed to assess the similarity of a stimulus to a prototype or stored exemplars. But this is not enough. Based on this similarity assessment a categorization decision needs to be reached. This decision is potentially probabilistic, be it because of noise in the similarity representation or noise in the decision process itself. In any case, a decision making model has to be built on top of the model for stimulus similarity. Traditionally, in exemplar theories this decision model has been implemented by Luce's choice rule (Shepard, 1957; Luce, 1959, 1961; Nosofsky, 1986). Luce's choice rule is related to logistic regression which is frequently used in psychophysics to assess a subject's responses (Recent methodological studies using logistic regression and Luce's choice rule are Jäkel & Wichmann, 2006; Kuss, Jäkel, & Wichmann, 2005). It has long been known that Luce's choice rule can be problematic (Debreu, 1960; Tversky, 1972; Luce, 1977) especially in similarity choice situations (Krantz, 1967). However, an extension like Tversky's elimination-by-aspects model (Tversky, 1972)—that can deal with these problems—is a lot more complicated and only recently have suggestions for its use in practice been put forward (Görür, Jäkel, & Rasmussen, 2006).

With Luce's choice rule in place it becomes obvious that exemplar models perform a logistic regression on the exemplar similarities. As the exemplar similarities can be modeled by kernels there is a close correspondence between exemplar models and kernel logistic regression. The basic model underlying ALCOVE (Kruschke, 1992), a well-known exemplar model, is even formally equivalent to kernel logistic regression. Interestingly, this model can be also seen as a radial-basis-function (RBF) neural network. RBF-networks have repeatedly been advocated as models for brain function by Poggio and coworkers (Poggio, 1990; Poggio & Edelman, 1990; Poggio & Bizzi, 2004). The relationship between different exemplar models, kernel logistic regression and RBF-networks is explained in the first part of Chapter 4.

5. Exemplar models and generalization

While the basic model underlying ALCOVE and kernel logistic regression is the same there are crucial differences in how the parameters of the model are adapted through learning. These differences can make a significant difference with regard to the generalization performance of the model. Exemplar models usually have too many free parameters and any learning algorithm that such a model might possibly implement is therefore prone to overfitting. As exemplar models, by definition, are rote learners there are strong concerns about the generalization ability of these models (Smith & Minda, 1998, 2000; Minda & Smith, 2001, 2002). If a model only learns the labels of observed exemplars by heart how will this mechanism explain generalization to new exemplars? As it turns out appealing to similarity is not enough in many of the models that are usually considered. While proponents of prototype theories argue that some abstraction mechanism accounts for the generalization, here we demonstrate that exemplar models can be made to generalize well without an explicit abstraction mechanism. For this analysis we use regularization techniques as they are used in machine learning. These techniques are explained in Chapter 2 and are subsequently applied to exemplar models in Chapter 4.

6. Preview

The core of this thesis consists of three chapters. Chapter 2 introduces the mathematical apparatus of positive definite kernels and reproducing kernel Hilbert spaces that will be used in subsequent chapters. This apparatus greatly deepened our understanding of categorization methods in machine learning and, as it turns out, is also useful for psychological theorizing. Chapter 3 shows that Shepard's universal law of generalization leads to a positive definite kernel and discusses consequences of this observation. This leads to an interpretation of perceptual spaces where stimuli are represented by their similarity to all other stimuli, the distance between stimuli is bounded and the worrisome property of segmental additivity is not needed. Chapter 4 draws parallels between kernel methods for categorization and exemplar models. The generalization ability of exemplar models is analyzed with the help of regularization techniques.

CHAPTER 2

Kernels

Machine learning is occupied with inventing computer algorithms that are capable of learning. For example, a machine that is equipped with a digital camera is shown instances of handwritten digits. Imagine an application where postal codes on letters have to be recognized so that the letters can be sorted automatically. The machine is shown many instances of each digit and has to learn to classify new instances based on the experience with the old ones. The prospect of not having to program a machine explicitly but rather having a machine learn from examples has attracted engineers to study learning since the early days of artificial intelligence. In their quest for intelligent machines early research was inspired by neural mechanisms and ideas from reinforcement learning. However, for practical applications researchers in machine learning also need to take technical constraints (like scalability, robustness and speed) into account. Furthermore, a good understanding of what the algorithm does, perhaps even with performance guarantees, would be very desirable if the algorithm was to be used in practice. Usually these constraints require techniques and insights from statistics, optimization and complexity theory that make the algorithm implausible as a psychological model of learning. Nevertheless, some of the methods used in machine learning are still strikingly similar to models that are discussed in psychology. Many of the ideas about learning that can be found in the machine learning literature are certainly based on the same intuitions that psychologists have.

Kernel methods, in particular, can be linked to neural network models and exemplar theories of categorization. Psychologically speaking, a kernel can often be thought of as a measure for stimulus similarity. In a category learning task it seems natural to assume that the transfer from old to new stimuli will depend on their similarity. In fact, this idea can be found throughout machine learning and psychology. As categorization is an important cognitive ability it has received a lot of attention from machine learning and psychology. It is also in categorization models that the similarity between machine learning methods and psychological models becomes most obvious.

This chapter is a tutorial on kernel methods for categorization. These methods try to tackle the same problems that human category learners face when they try to learn a new category. Hence, we think that the mathematical tools that are used in machine learning show a great potential to be also useful for psychological theorizing. Even if most of the solutions that machine learning offers turned out to be psychologically implausible, psychologists should still find it interesting to see how a related field deals with similar problems—Especially as machine learning methods are increasingly used for the analysis of neural and behavioral data. At the very least, this chapter provides an introduction to these new tools for data analysis. We find, however, that some of the methods in machine learning are closely related to categorization models that have been suggested in psychology. Briefly, some kernels in machine learning are akin to a class of similarity measures considered in psychology. This class of similarity measures is based on Shepard's *universal law of generalization* and has been used extensively in exemplar models of categorization

(Nosofsky, 1986; Kruschke, 1992). Kernel methods are also like exemplar models in other respects: They usually store all the exemplars they have encountered during the course of learning and they can be implemented in a neural network. In Chapters 3 and 4 we will use some of the results presented in this chapter to clarify the relationship between similarity and generalization in categorization models and to resolve some conceptual problems with popular models of similarity. The other two chapters focus on the psychological aspects of kernels, whereas in this chapter we concentrate more on the mathematical aspects.

There are several useful introductions to kernel methods in the machine learning literature but none of them is addressing psychological issues directly—hence this chapter. Most of the technical material we present is based on two recent books on kernel methods (Christianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002) and standard results in linear algebra (e.g. Strang, 1988). We will assume that the reader has had some previous exposure to linear algebra, for example in the context of artificial neural networks or psychometrics. However, in order to make the chapter accessible to a larger audience we included reminders of relevant results throughout the text.

1. Inner products

So what is a kernel? Kernels can be regarded as a non-linear generalization of inner products. We will take a little detour before explaining kernels and discuss the relationship between inner products, perceptrons and prototypes. This will set the stage on which kernels appear naturally to solve non-linear classification problems.

1.1. Perceptrons.

The perceptron can be considered the most basic of all pattern recognition algorithms. It was conceived as a simple model for learning in the brain (Rosenblatt, 1958). A pattern x, in the form of n real numbers x_1 to x_n, is fed into a neuron. The inputs are weighted by the synaptic strengths w of the n connections to the neuron. There are n real numbers w_1 to w_n that represent the synaptic strength of each input. The neuron integrates all its weighted inputs by summing them up:

$$\langle w, x \rangle = \sum_{i=1}^{n} w_i x_i. \tag{1}$$

If the excitation of the neuron is greater than a threshold value θ the neuron fires. The excitation is a linear combination of the inputs. For this reason the perceptron is also referred to as a linear classifier. Mathematically speaking, the neuron calculates the standard inner product, denoted with brackets $\langle \cdot, \cdot \rangle$, of the vector x with the vector w, both of which are elements in a n-dimensional vector space. Inner products are also called dot products or scalar products in linear algebra.

A vector space with an inner product $\langle \cdot, \cdot \rangle$ is a very rich representation and has a natural measure of length and angle that conforms to intuitions about Euclidean space. The length or norm $\|\cdot\|$ of any vector w can naturally be defined with the help of the inner product as:

$$\|w\|^2 = \sum_{i=1}^{n} w_i^2 = \langle w, w \rangle . \tag{2}$$

By using Pythagoras' theorem one can find that this is in agreement with Euclidean intuitions. All the familiar properties of Euclidean space can be expressed in terms of the standard inner product. The distance $d(\cdot, \cdot)$ between two points x and y in the space can then be defined as the length of their difference vector

$$d(x,y) = \|x-y\| = \sqrt{\langle x-y, x-y\rangle}. \tag{3}$$

This distance can be used to define a metric on the space. Moreover, the angle α between two vectors v and w can be expressed as

$$\cos\alpha = \frac{\langle v,w\rangle}{\|v\|\,\|w\|}. \tag{4}$$

In particular, two vectors are perpendicular whenever their inner product is zero.

Geometrically speaking, the weight vector together with the threshold can be interpreted as the normal vector of a hyperplane (an ordinary plane if the dimension of the space is three and a straight line if the dimension of the space is two). Checking whether the inner product is bigger than the threshold is equivalent to checking which side of the hyperplane a pattern vector x falls on. In this way a simple classification can be implemented by separating the vector space into the two parts on both sides of the hyperplane. This is illustrated in Figure 1. The figure shows a two-dimensional space and each point in the space defines a possible pattern. There are two classes of patterns (circles and crosses) and several instances of each class are shown. A vector w pointing away from the origin is depicted together with its hyperplane $\langle w,x\rangle = 0$, that is the set of all points x that are perpendicular to w. If the inner product between w and x is greater than zero the two vectors form an angle that is less than 90 degrees, hence w and x lie on the same side of the hyperplane. It is possible to shift the hyperplane along the vector w by changing the threshold parameter θ. In this example we have chosen w and θ such that the hyperplane that they define can correctly separate the circles from the crosses. In general, the learning problem for the perceptron is to find a vector w and a threshold θ that separates two classes of patterns as well as possible. It is a very common view to see learning as adapting weights in a neural network. There is a long list of learning algorithms that try to accomplish this task of which the perceptron learning algorithm is just one.

1.2. Prototypes. Take the psychologically rather than neurally motivated example of a prototype learner (Posner & Keele, 1968; Reed, 1972). The learning machine is given a set of patterns A that are known to belong to one class and a set of patterns B that are known to belong to another class. The prototype learner is usually understood as trying to extract the central tendency of the two classes. Hence, to separate A from B the arithmetic means of all examples in A and B are calculated: $\overline{a} = \frac{1}{|A|}\sum_{a\epsilon A} a$ and $\overline{b} = \frac{1}{|B|}\sum_{b\epsilon B} a$. A new pattern x is classified as belonging to class A if it is closer to $\overline{a}$ (the mean of A) than to $\overline{b}$ (the mean of B). We can take 'closer' to mean Euclidean distance in the vector space in which the patterns are given. The Euclidean distance between two points x and y is given by the square root of the inner product of the difference vector with itself: $\langle x-y, x-y\rangle$ (Eq. 3). Therefore, a Euclidean prototype classifier decides that a new stimulus x belongs to class A whenever

$$\begin{align*}
\langle x-\bar{b}, x-\bar{b}\rangle &> \langle x-\bar{a}, x-\bar{a}\rangle \\
\langle x,x\rangle - 2\langle \bar{b},x\rangle + \langle \bar{b},\bar{b}\rangle &> \langle x,x\rangle - 2\langle \bar{a},x\rangle + \langle \bar{a},\bar{a}\rangle \\
2\langle \bar{a},x\rangle - 2\langle \bar{b},x\rangle &> \langle \bar{a},\bar{a}\rangle - \langle \bar{b},\bar{b}\rangle \\
\langle \bar{a}-\bar{b},x\rangle &> \frac{1}{2}\left(\langle \bar{a},\bar{a}\rangle - \langle \bar{b},\bar{b}\rangle\right) \\
\langle \bar{a}-\bar{b},x\rangle &> \theta. \tag{5}
\end{align*}$$

Remember that the definition of the inner product $\langle\cdot,\cdot\rangle$ involves a sum and therefore it is linear in both arguments: For all x, y and z it holds that $\langle x, y+z\rangle =$

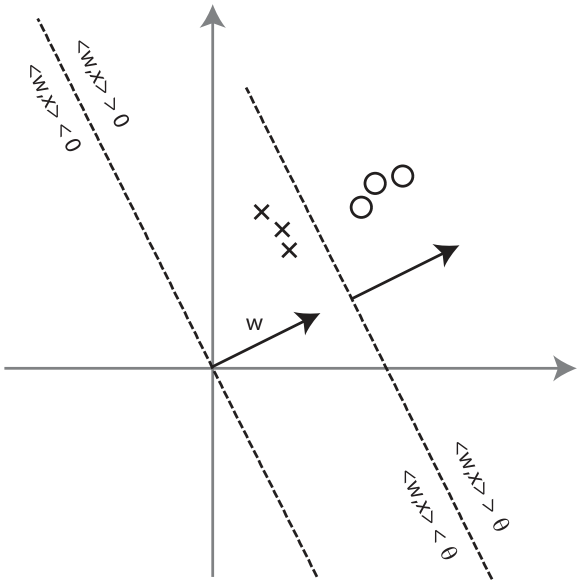

Figure 1. Each perceptron defines a hyperplane in a vector space. The weight vector w is the normal vector of this hyperplane and the threshold θ defines the offset from the origin. On one side of the hyperplane (the side that w points to) the inner product of all points with w is greater than θ. On the other side the inner product is smaller than θ. In this example we have chosen w and θ so they can separate the circles from the crosses.

$\langle x, y\rangle + \langle x, z\rangle$ and also that $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$. We have used this property extensively in the above derivation. From the last line of (5) it can be seen that the prototype classifier defines a hyperplane in the input space just as the perceptron in Eq. (1) does. The weight vector w is given by the difference of the means $\bar{a} - \bar{b}$ and the threshold θ by the right hand side $\frac{1}{2}\left(\langle \bar{a}, \bar{a}\rangle - \left\langle \bar{b}, \bar{b}\right\rangle\right)$. However, θ is really just a bias parameter that determines which of the two category responses is preferred and might be chosen differently. The crucial fact is that for the prototype classifier we take an inner product with the difference vector of the means.

1.3. Positive definite matrices. The inner product defined in Eq. (1) is called the standard inner product because it naturally arises in the context of Euclidean spaces. In general, an inner product $\langle \cdot, \cdot\rangle$ has to fulfill three formal properties that ensure that the norm, distance and angle will behave as in Euclidean space. First, it has to be symmetric: For all real-valued vectors w and v it holds that $\langle w, v\rangle = \langle v, w\rangle$. This reflects the fact that the (absolute) angle between two vectors does not depend on whether it is measured from w to v or from v to w. Second, an inner product has to be linear in its arguments, that is for a real number a and three vectors u, v and w it holds that $\langle au, v\rangle = a\,\langle u, v\rangle$ and $\langle u + w, v\rangle = \langle u, v\rangle + \langle w, v\rangle$. Because of the symmetry an inner product is linear in both arguments. Third, an inner product has to be positive definite. By positive it is meant that $\langle w, w\rangle \geq 0$ for all w. Definiteness refers to $\langle w, w\rangle = 0$ if and only if $w = 0$. Positive definiteness is a natural requirement for a length measure. Remember that the inner product of a vector with itself $\langle w, w\rangle$ defines the square of the length of the vector $\|w\|^2$ and the squared length always has to be positive and is only zero for the zero vector. It is easily verified that the standard inner product (Eq. 1) fulfills all three axioms.

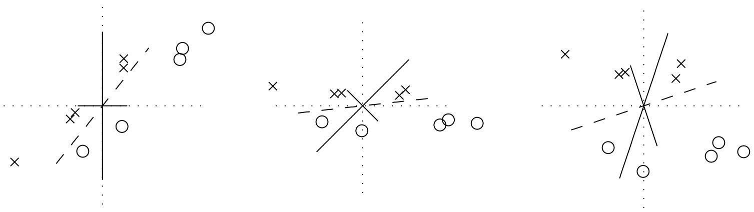

FIGURE 2. Whether a prototype classifier can separate two classes depends also on the inner product that is chosen. The left panel shows two classes (circles and crosses) with highly correlated dimensions—in this case the standard inner product is not appropriate. The short solid line connects the means of the category distributions and the long solid line is the corresponding decision bound when the standard inner product is used. The dashed line depicts a decision bound with a different inner product. This inner product corresponds to the standard inner product in the space depicted on the right that can be obtained by first rotating (central panel) and then rescaling the original space (right panel).

The standard inner product is by no means the only interesting inner product. A generalization that will be very important in the following is given by

$$\langle w, v\rangle_K = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i v_j k_{ij}. \tag{6}$$

Taking K to be a matrix with entries k_{ij} and ^{T} to denote the transpose of a matrix (rows and columns exchanged) the equation can be written more elegantly as a matrix multiplication

$$\langle w, v\rangle_K = w^T K v. \tag{7}$$

If K is the identity matrix the standard inner product (1) is recovered. In order for this definition to result in an inner product the three axioms have to be fulfilled. Symmetry depends on the symmetry of K. If $k_{ij} = k_{ji}$ for all i and j then the definition in (6) will be symmetric. As a consequence of the linearity of the sum it is immediately clear that (6) is always linear. It remains to demand positive definiteness. The inner product defined in Eq. (6) is positive definite if the matrix K is positive definite, that is for all vectors w the quadratic form that defines the squared length of the vector $\|w\|^2$ is positive,

$$w^T K w \geq 0, \tag{8}$$

and zero if and only if w is zero. It is a standard result in linear algebra that symmetric positive definite matrices can be decomposed into principal components. Principal component analysis (PCA) is used frequently in the analysis of psychological data, for example covariance matrices are positive definite and they are often subjected to PCA. There is a rotation matrix Ψ and a diagonal matrix Λ that contains only positive eigenvalues such that $K = \Psi^T \Lambda \Psi$. With this result the inner product (7) can be rewritten as

$$\langle w, v\rangle_K \quad = \quad w^T K v$$

$$
\begin{aligned}
&= w^T \left(\Psi^T \Lambda \Psi\right) v \\
&= w^T \left(\sqrt{\Lambda}\Psi\right)^T \left(\sqrt{\Lambda}\Psi\right) v \\
&= \left(\sqrt{\Lambda}\Psi w\right)^T \left(\sqrt{\Lambda}\Psi v\right) \\
&= \left\langle \sqrt{\Lambda}\Psi w, \sqrt{\Lambda}\Psi v \right\rangle .
\end{aligned}
$$

Remember that for any two matrices A and B that can be multiplied $(AB)^T = B^T A^T$. By using Ψ the vectors v and w are rotated such that they coincide with the principal components. After that they are rescaled using the diagonal matrix $\sqrt{\Lambda}$. In this coordinate system the inner product $\langle w, v\rangle_K$ amounts to a standard inner product. In order for K to implement an inner product all eigenvalues have to be positive. Otherwise there could be vectors with a squared length smaller than zero—clearly in contradiction with Euclidean intuitions. Note also that if some eigenvalues were zero than there would be vectors, other than the zero vector, with a length zero. All this illustrates the close connections between the standard inner product, Euclidean space and positive definite matrices. Positive definite matrices can define an inner product. If the coordinate axes are rotated and rescaled appropriately this inner product becomes a standard inner product and therefore can induce a norm, a metric and angles that behave like the familiar Euclidean ones.

1.4. Prototypes and orthogonality. As an example consider the following classification problem. The left panel of Figure 2 shows two categories drawn from two Gaussian distributions. Each point is a stimulus that is described by two dimensions. The dimensions are highly correlated for both stimulus classes. On a first glance a prototype learner will not find a good decision bound to separate the two classes. The two means for the two classes are connected with a solid line and the decision bound resulting from a prototype classifier is also shown as a solid line. As the decision bound is orthogonal to the shortest connection between the two means it cannot pay due respect to the correlations in the classes. On a first glance one could think that the problem is that the prototype classifier cannot deal with correlation and therefore such a problem cannot be solved by a prototype classifier. However, the problem does not actually lie in the prototype classifier as such, it lies in the inner product that is used to define orthogonality. A prototype learner that does not fail for even the simplest category structures should take the covariance of the stimulus dimensions into account (Reed, 1972; Fried & Holyoak, 1984; Ashby & Gott, 1988). If we take the inner product $\langle \cdot, \cdot \rangle_K$ to be given by a positive definite matrix K that is the inverse of the covariance matrix of the classes we get the *right* definition of orthogonality. The resulting decision bound is depicted as a dashed line. K is the inverse of a positive definite matrix (covariance matrices are always positive definite) and is therefore also a positive definite matrix. Hence, it corresponds to the standard inner product after rotating and scaling the space with the matrices Ψ and $\sqrt{\Lambda}$. The middle panel in Figure 2 shows the rotated space and the right panel shows the space after scaling. In the transformed space the two classes do not have correlated axes anymore and therefore the prototype classifier with the standard inner product in this transformed space can classify all stimuli correctly.

1.5. Non-linear classification problems. A linear classifier like the perceptron is a very attractive method for classification because it builds on strong

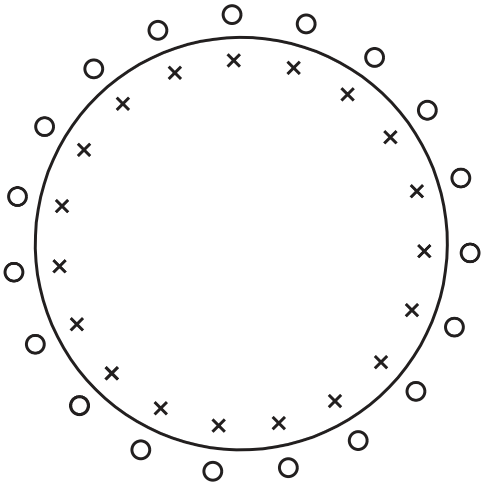

FIGURE 3. The crosses and the circles cannot be separated by a linear perceptron in the plane.

geometric intuitions and the extremely well-developed mathematics of linear algebra. However, there are problems that a linear classifier cannot solve—at least not directly. As several psychological theories of categorization are based on linear classifiers this issue has also attracted some attention in the psychological literature (Medin & Schwanenflugel, 1981; Smith, Murray, & Minda, 1997). One example of a problem that cannot be solved with a linear classifier can be seen in Figure 3. For a long time the most popular approach to solve non-linear problems like this one was to use a multi-layer perceptron. Multi-layer perceptrons are known to be able to approximate any function (Hornik, Stinchcombe, & White, 1989) and can be trained efficiently by using the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986). The approach that will be presented here is fundamentally different. The strategy is to use a non-linear function to map the input patterns into a space where the problem can be solved by a linear classifier. The following toy-example illustrates this approach.

Figure 3 shows examples from two classes (crosses and circles) that cannot be separated by a hyperplane in the input space (i.e. a straight line in two dimensions). Instead of trying to classify the examples in the input space that is given by the values x_1 and x_2 the data are transformed in a non-linear way. Linear classification of the data is then attempted in the transformed space. In machine learning such a non-linear transform is called a feature map and the resulting space a feature space. The term 'feature' is already heavily overloaded in psychology. Therefore we will use the more neutral terms linearization function and linearization space instead. The term linearization space was used in an early paper on kernel methods (Aizerman, Braverman, & Rozonoer, 1964b). For example, consider the following linearization function $\Phi : \mathbb{R}^2 \mapsto \mathbb{R}^3$

$$\Phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \end{pmatrix} = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}. \tag{9}$$

This transformation maps the example patterns to a three-dimensional space that is depicted in Figure 4. The examples live on a two-dimensional manifold of this three-dimensional space. In this space the two classes become linearly separable, that is it is possible to find a two-dimensional plane such that the circles fall on one

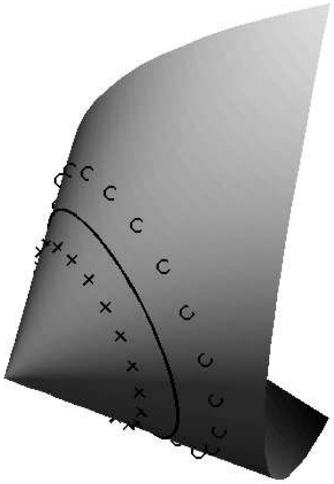

Figure 4. The crosses and circles from Figure 3 can be mapped to a three-dimensional space in which they can be separated by a linear perceptron.

side and the crosses on the other side. This shows that with an appropriate non-linear transformation of the input a simple linear classifier can solve the problem. The linearization approach is akin to transforming the data in data-analysis before fitting a linear model. In the current example each hyperplane in the linearization space defines a quadratic equation in the input space. Hence, it is possible to deal with quadratic (i.e. non-linear) functions by only using linear methods. In general, the strategy is to *preprocess* the data with the help of a function Φ such that a linear perceptron model is likely to be applicable. Formally this can be expressed as

$$\langle w, \Phi(x) \rangle = \sum_{i=1}^{n} w_i \phi_i(x), \tag{10}$$

where n is now the dimension of the linearization space and w is a weight vector in the linearization space. It is clear that there is a wide variety of non-linear functions that can be used to preprocess the input. In fact, this approach was very popular in the early days of machine learning (Nilsson, 1965). The problem is of course that the function Φ has to be chosen before learning can proceed. In our toy example we have only shown how one can use linear methods to deal with quadratic functions but usually one will not know in advance whether it is possible to separate the data with a quadratic function. However, if Φ is chosen to be sufficiently flexible, for example instead of a quadratic function with only three coefficients one could choose a high order polynomial with many coefficients, then it may be possible to approximate even very complicated decision functions. This comes at the cost of increasing the dimensionality of the linearization space and the number of free parameters. Therefore early machine learning research has tried to avoid this.

2. Kernels

The next section will introduce the *kernel trick* that makes it possible to work with high dimensional and flexible linearization spaces.

2.1. The kernel trick. There is an interesting observation about the linearization function that was used in the foregoing example. The standard inner product between two input vectors in the linearization space can be calculated without having to explicitly map the data into the linearization space. For two points x and y in $\mathbb{R}^2$ it holds that

$$\langle \Phi(x), \Phi(y) \rangle = x_1^2 y_1^2 + 2 x_1 x_2 y_1 y_2 + x_2^2 y_2^2 = \langle x, y \rangle^2 .$$

A more general result can be proved. For an n dimensional input space a class of popular and flexible linearization functions is given by all monomials of degree d. A monomial of degree d takes the product of d components of an input vector x. E.g., for $n = 5$ the following are monomials of degree $d = 3$: x_1^3, $x_1 x_2 x_5$ and $x_2^2 x_4$. The possible number of monomials is given by choosing d out of n with replacement. The order does not matter because of the commutativity of the product. However, for simplicity let us consider a linearization function that takes all n^d possible ordered monomials. Thus, $x_1 x_2 x_3$ is a dimension in the new space but $x_2 x_3 x_1$ would be another dimension. For the linearization function $\Phi' : \mathbb{R}^n \mapsto \mathbb{R}^{n^d}$ that computes all ordered monomials it holds that

$$\begin{aligned} \langle \Phi'(x), \Phi'(y) \rangle &= \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} ... \sum_{i_d=1}^{n} x_{i_1} x_{i_2} ... x_{i_d} y_{i_1} y_{i_2} ... y_{i_d} \\ &= \sum_{i_1=1}^{n} x_{i_1} y_{i_1} \sum_{i_2=1}^{n} x_{i_2} y_{i_2} ... \sum_{i_d=1}^{n} x_{i_d} y_{i_d} \\ &= \left(\sum_{i=1}^{n} x_i y_i \right)^d = \langle x, y \rangle^d . \end{aligned}$$

Calculating the inner product in the linearization space is the same as taking the inner product in the original space and taking it to the power of d. Computationally, this is an extremely attractive result. Remember that a high number of dimensions is needed to make the linearization space sufficiently flexible to be useful. If calculated naively the computational effort of the inner product in the linearization space scales with its dimensions. However, this result shows that, in the case of a monomial linearization function, it is not necessary to explicitly map the vectors x and y to the n^d dimensional linearization space to calculate the dot product of the two vectors in this space. It is enough to calculate the standard inner product in input space and take it to the power of d.

The function $k(x, y) := \langle \Phi(x), \Phi(y) \rangle = \langle x, y \rangle^d$ is our first example of a kernel, the so-called polynomial kernel. Intuitively, kernels can provide a way to efficiently calculate inner products in higher dimensional linearization spaces. They also provide a convenient non-linear generalization of inner products. With the help of a kernel, it is easy to build non-linear variants of simple linear algorithms that are based on inner products. This is called the *kernel trick* in the machine learning literature.

Take as an example the prototype classifier, again. Instead of taking the mean in input space, like in Eq. (5), one can construct a prototype classifier in the linearization space. We will take the threshold θ to be a free parameter that we can tune to account for biases. For the left hand side we now want to take the mean in the linearization space, that is the mean after we applied the mapping Φ. To decide whether x belongs to class A we also map x to the linearization space and check that

$$
\begin{aligned}
\left\langle \frac{1}{|A|}\sum_{a\epsilon A}\Phi(a) - \frac{1}{|B|}\sum_{b\epsilon B}\Phi(b), \Phi(x)\right\rangle &> \theta \\
\frac{1}{|A|}\sum_{a\epsilon A}\langle\Phi(a),\Phi(x)\rangle - \frac{1}{|B|}\sum_{b\epsilon B}\langle\Phi(b),\Phi(x)\rangle &> \theta \\
\frac{1}{|A|}\sum_{a\epsilon A}k(a,x) - \frac{1}{|B|}\sum_{b\epsilon B}k(b,x) &> \theta, \qquad (11)
\end{aligned}
$$

where as before A and B are sets of patterns from two classes and the linearity of the inner product and the sum were used.

The input space could be a 16×16 matrix of pixel values, that is a 256 dimensional space. The linearization space could be all monomials of degree 10. Mapping the inputs to the linearization space and calculating the mean there would be prohibitive as the mapping of each input to this space takes a large number of dimensions. Despite the high number of dimensions it is possible to use a prototype classifier in the linearization space by taking advantage of the kernel trick. By using the kernel trick it is not necessary to calculate the mean in the linearization space but a prototype classifier can still be used by using Eq. (11).

2.2. Reproducing kernel Hilbert space. Every linearization function Φ defines a kernel function via

$$k(x,y) = \langle\Phi(x),\Phi(y)\rangle\,. \qquad (12)$$

It is always possible to define a kernel by choosing a linearization function Φ and an inner product. The function $k(\cdot,\cdot)$ can be evaluated by explicitly mapping patterns to the linearization space and calculating the inner product in the linearization space. However, as the example of the polynomial kernel has shown, sometimes it is not necessary to actually compute Φ. It is natural to ask under what circumstances does a function $k(\cdot,\cdot)$ implement an inner product in a linearization space and what does the corresponding linearization space and linearization function look like. As it turns out there is a well-developed branch of mathematics that deals with these questions: Functional analysis. In short the answer is that if $k(\cdot,\cdot)$ is a symmetric and positive definite kernel then k implements an inner product in a linearization space. Constructing a linearization space and an inner product for a positive definite kernel is the purpose of this section.

First, the introduction of some notation is required. For a set of patterns x_1 to x_N and a function $k(\cdot,\cdot)$ of two arguments the kernel matrix is the matrix that collects all pairwise applications of k to the patterns. Let us denote this $N \times N$ matrix with K and denote the entry in the i^{th} row and j^{th} column with k_{ij} then

$$K \text{ with } k_{ij} = k(x_i, x_j)$$

is called the kernel matrix or Gram matrix for the patterns x_1, ..., x_N. A real and symmetric function $k(\cdot,\cdot)$, that is a function with the property $k(x,y) = k(y,x)$, is called a *positive definite* kernel if for all choices of N points the corresponding kernel matrix K is *positive semi-definite*, that is for all N-dimensional vectors w

$$w^T K w \geq 0. \qquad (13)$$

Note that for a matrix to be positive semi-definite we do not require that equality only holds for $w = 0$ (as opposed to the definition of a positive definite matrix, see Eq. 8). As K is only positive semi-definite it can have eigenvalues that are zero and does not have to be full rank. This definition of a positive definite kernel seems confusing because for a kernel to be positive definite we require the corresponding kernel matrices to be positive semi-definite. However, the definition we give is the usual definition used in machine learning and therefore we will use it, too (Schölkopf & Smola, 2002).

With the definition of a positive definite kernel in mind, it is possible to construct a vector space, an inner product, and a linearization function such that the kernel condition (12) is fulfilled. In the following, these three steps are demonstrated in a purely formal way. After that, the formal steps are illustrated by an example, using the Gaussian kernel.

2.2.1. *Step 1: Constructing a vector space.* The vector space will be a space of functions constructed from the kernel. Let $k(\cdot, x)$ denote a function that is taken to be a function of its first argument with a fixed second argument. The vector space is then defined as all functions of the form

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i). \tag{14}$$

Each function in the space is a linear combinations of kernel functions $k(\cdot, x_i)$ and can be expressed by some set of N patterns $x_1, ..., x_N$ with real coefficients $w_1, ..., w_N$. It is important to realize that these N patterns could be different for different functions. All functions are linear combinations of kernel functions given by k and because they are linear combinations they define a vector space—functions can be added and multiplied with scalars. When functions are added potentially all the kernel functions of the two added functions need to be included in the expansion of the summed function but the sum will still be in the vector space.

The expansion of f given in Eq. (14) might not be unique. There is no requirement in the definition of f that the kernel functions need to be linearly independent. If they are not independent then the same function can be expressed in different ways. The function space is the span of the generating system of functions. If there is an infinite number of potential independent kernel functions then the vector space is infinite dimensional, even though each function f can be expressed by a finite sum.

2.2.2. *Step 2: Constructing an inner product.* Next we will equip this vector space with an inner product. A possibly infinite dimensional vector space with an inner product is called a *pre-Hilbert space.* If certain limit points are included in the space it is completed and turned into *Hilbert space* proper. We will ignore these technicalities here (but see Schölkopf & Smola, 2002) and simply note that Hilbert spaces can be thought of as the infinite dimensional generalization of Euclidean spaces. Take a function f with an expansion given by Eq. (14) and let $g(x) = \sum_{i=1}^{M} v_i k(x, y_i)$ be another function from this space then we can define the inner product between the two functions f and g as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{N} \sum_{j=1}^{M} w_i v_j k(x_i, y_j). \tag{15}$$

In order to distinguish the inner product in Hilbert space from the normal inner product in Euclidean space we have added the little index $\mathcal{H}$. We have to show that this definition is indeed an inner product. First we have to show that it is well-defined. The particular expansions of f and g that are used in the definition might not be unique, as mentioned above. Fortunately, the definition (15) does not depend on the particular expansions of f and g that are used to calculate the inner product. To see this, let $f(x) = \sum_{i=1}^{N'} w_i' k(x, x_i')$ and $g(x) = \sum_{i=1}^{M'} v_i' k(x, y_i')$ be two new expansions of f and g that are different from the ones used in the definition of the inner product (15). They will, however, result in the same inner product because

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{M} w_i v_j k(x_i, y_j) &= \sum_{i=1}^{N} w_i g(x_i) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M'} w_i v'_j k(x_i, y'_j) \\ &= \sum_{j=1}^{M'} v'_j f(y'_j) \\ &= \sum_{i=1}^{N'}\sum_{j=1}^{M'} w'_i v'_j k(x'_i, y'_j).\end{aligned} \tag{16}$$

Therefore, (15) is indeed well-defined. To show that it is an inner product it also has to be symmetric, linear in its arguments and positive definite. As k is symmetric in both arguments the above definition is also symmetric. It is obviously linear because of the linearity of the sum. Positive definiteness means that $\langle f, f\rangle_{\mathcal{H}} \geq 0$ where equality only holds for $f = 0$. Note that $\langle f, f\rangle_{\mathcal{H}} = w^T K w$ by definition. As the defining property of a positive definite kernel is that the kernel matrix K is always positive semi-definite (Eq. 13), it is immediately clear that $\langle f, f\rangle_{\mathcal{H}} \geq 0$. Definiteness is a bit more tricky but it can be proved that for all positive definite kernels definiteness of (15) holds (Schölkopf & Smola, 2002). Hence, all positive definite kernels can define an inner product in the above way. This may also justify calling these kernels positive definite.

2.2.3. *Step 3: Constructing a linearization function.* Each kernel $k(\cdot, x)$ with a fixed x is trivially contained in the vector space. It is simply an expansion with only one kernel function and a weight of one. Therefore, the inner product (15) of this function with a function f that has N coefficients w_i and kernel functions $k(\cdot, x_i)$ is

$$\langle k(\cdot, x), f\rangle_{\mathcal{H}} = \sum_{i=1}^{N} w_i k(x, x_i) = f(x), \tag{17}$$

by the definition of the function space (Eq. 14). This is a remarkable fact: The inner product with the function $k(\cdot, x)$ evaluates the function f at point x. Therefore $k(\cdot, x)$ is also called the representer of evaluation. Another remarkable property directly follows from the definition of the inner product (Eq. 15)

$$\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y) \tag{18}$$

because each of the two kernel functions has a simple expansion with just one summand and a coefficient of one. Due to these two properties the linear space of functions as given in Eq. (14) with the above dot product $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ is called a *reproducing kernel Hilbert space* (RKHS) in functional analysis (if it is completed).

Now, a linearization function can be defined in the following way $\Phi(x) := k(\cdot, x)$. Because of the reproducing property the kernel condition $k(x, y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}$ holds for this linearization function. The linearization space is a space of functions over the x. The linearization function that was constructed maps each point x in the input space to a function $k(\cdot, x)$ in the linearization space.

Remember what is accomplished by this. Starting from a positive definite kernel a vector space, an inner product and a linearization function were constructed such that the kernel condition (12) holds. If the kernel is easy to calculate then by means of the kernel it is possible to calculate inner products in the linearization space without actually mapping the points x and y into it. Understanding this trick opens up a box of new non-linear tools for data analysis. Any method where the linearization space only occurs in inner products of the form (12) can benefit. In fact, there is now a long list of familiar linear methods that have been kernelized. This list includes kernel principal component analysis (Schölkopf, Smola, & Müller, 1998) and many others (Schölkopf & Smola, 2002).

It is often helpful (but sometimes misleading) to sharpen one's intuitions about Hilbert spaces by considering the finite dimensional case which reduces functional analysis to linear algebra. We illustrate the above construction for two finite dimensional examples at the end of this chapter. One of the examples uses the quadratic kernel that we used as a motivating example for the introduction of kernels. The other example considers the case where there are only a finite number of patterns. Readers who do not feel comfortable with the above derivation are encouraged to look at both examples but especially the second example might prove helpful. In any case, the example of the Gaussian kernel may clarify the construction of a linearization space from a kernel.

2.3. Gaussian kernel example. The Gaussian kernel has frequently been used in psychology to model the similarity between two mental representations x and y (Nosofsky, 1986, 1990; Ashby & Maddox, 1993). It is defined as

$$k(x, y) = \exp^{-\|x-y\|^2} . \tag{19}$$

A standard result in functional analysis is that the Gaussian kernel is a positive definite function and that any kernel matrix K resulting from the Gaussian kernel is always full rank (Schoenberg, 1938; Schölkopf & Smola, 2002). We will not prove these two facts but we will take them for granted in what follows. The fact that the kernel matrices are always full rank and therefore positive definite (and not just semi-definite) is important and should be kept in mind.

2.3.1. *Step 1: Constructing a vector space.* As a vector space we take all functions that can be expressed as a linear combination of Gaussian kernel functions (see Eq. 14). One example of such a function is shown in Figure 5. The Gaussian functions are depicted with dotted lines. Their height is scaled with the weight w_i that each function receives in the sum. Summing the Gaussian functions results in the solid functions. It is easy to imagine that by changing the weights and adding more Gaussian functions very different functions can be implemented or at least approximated. While each function is a finite sum, the space includes all functions that can be expressed in this way with infinitely many different choices for Gaussian functions. In fact, there are uncountably many choices because each point on the axis is a potential candidate for a Gaussian functions centered on this point. For this reason this vector space does not have a finite dimensional basis. It is not possible to describe all functions that are spanned by the Gaussian functions with a finite number of basis functions. We have an example of an infinite dimensional space—that however in many respects is similar to the ordinary finite dimensional vector spaces that are the subject of linear algebra. For example, this infinite dimensional space can also be equipped with an inner product.

2.3.2. *Step 2: Constructing an inner product.* Eq. (15) defines an inner product that can be used. With the inner product on the function space it is possible to define a norm and a metric on the space in the same way as it is done in Euclidean spaces. One can even calculate angles between functions. Let us, for illustration, calculate the angle between two Gaussian functions. This is done in the same way as in Euclidean spaces (see Eq. 4). First note that each Gaussian function has a trivial expansion in the function space. It is simply itself with a weight of one. The

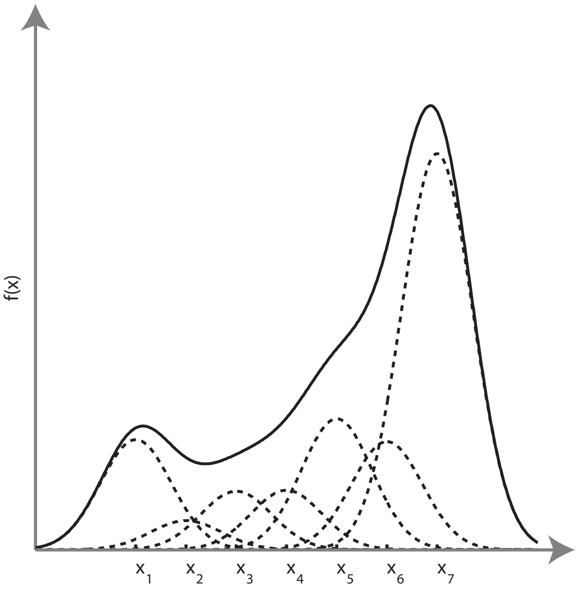

Figure 5. An example of a function from the linear function space defined in Eq. (14). Each function in the space is a linear combination of generating functions that are given by the kernel: $f(x) = \sum_i^N w_i k(x, x_i)$. Here we have depicted seven Gaussian basis functions as dotted lines. Their height is proportional to their weight w_i. The sum of these is shown with the solid line. All functions that can be expressed as such a linear combination of kernel functions are in the function space.

norm of a Gaussian function is one because

$$\|k(\cdot, x)\|^2 = \langle k(\cdot, x), k(\cdot, x)\rangle = k(x, x) = \exp^{-\|x-x\|^2} = 1 \tag{20}$$

which is only using the definition of the inner product (15) and the Gaussian kernel. Therefore, the cosine of the angle between two Gaussian functions is directly given by the inner product. As two Gaussian functions each have an expansion with only one weight that is set to one their inner product is $\langle k(\cdot, x), k(\cdot, x)\rangle = k(x, y)$, which is of course the reproducing property (18). Hence, $k(x, y)$ can be interpreted as the cosine of the angle between two Gaussian functions centered on x and y, respectively. This has interesting consequences. For two non-identical points x and y it holds that $1 > k(x, y) > 0$. Therefore, the angle between two Gaussian functions lies between 0 and 90 degrees. The further two Gaussians are apart the greater is their angle. Functions that are far apart are almost orthogonal. This makes sense because they span different parts of the function space. But as no two Gaussians are completely orthogonal it also means that the Gaussian functions do certainly not form an orthogonal basis of the function space. We have noted before that for the Gaussian kernel the kernel matrices (that collect all pairwise inner products of the Gaussians) are always full rank, hence any number of Gaussian functions are always linearly independent and therefore form a basis for the subspace that they span.

2.3.3. *Step 3: Constructing a linearization function.* The linearization function Φ maps points from the space in which x is defined to a space of functions over all possible x. In the example of the Gaussian kernel this means that we map a point x to the function $k(\cdot, x)$, a Gaussian function centered on x. In a psychological setting imagine x to be the representation of a stimulus in some perceptual space. Imagine further that k is interpreted as a similarity measure. The further two stimuli are apart the less similar they are and this relationship is captured in the Gaussian kernel. Mapping a stimulus x to the function $k(\cdot, x)$ means replacing a stimulus with its similarity to all other stimuli. Representation in this RKHS is literally representation of similarities (Edelman, 1998). In Chapter 3, on similarity, we discuss some consequences of this observation in more detail. Here, it suffices to say that calculating similarity by a Gaussian function is the same as taking an inner product in the corresponding RKHS that was constructed above.

2.4. Prototypes and exemplars. Let us stay with the example of the Gaussian kernel (19) for a while. Keep in mind that in a psychological setting the Gaussian kernel $k(x, y)$ is interpreted as the similarity between two stimuli x and y. To see the potential of the RKHS view of the Gaussian kernel for psychological theorizing, imagine we construct a prototype classifier in RKHS according to inequality (11). Let us assume that the bias parameter θ is set to zero. In this case, for a prototype classifier in the linearization space we decide that x belongs to class A and not to class B if $\Phi(x)$ is closer to the mean of the stimuli in A than in B. This can be done by checking that

$$\frac{1}{|A|}\sum_{a \epsilon A} k(a, x) > \frac{1}{|B|}\sum_{b \epsilon B} k(b, x), \tag{21}$$

that is the mean similarity of x to all exemplars of class A is bigger than the mean similarity to all exemplars of class B. The left side of the inequality (if appropriately normalized) can be interpreted as a kernel-density estimate for the class density of A (Ashby & Alfonso-Reese, 1995). It can also be interpreted as an estimate for the degree that a new x belongs to A. In any case, this is the most basic exemplar model of categorization, but it was derived from a prototype classifier.

2.5. Infinite dimensional perceptrons. Remember that the reason why we introduced the kernel trick was that a flexible preprocessing is needed so that a linear classifier can solve many non-linear categorization problems. In the previous section we examined the prototype classifier in an infinite dimensional function space, in this section we examine linear perceptrons in such a space. It will turn out that a linear classifier in an infinite-dimensional space can separate all possible stimuli in two classes.

Formally, a perceptron with a preprocessor Φ is given by an inner product of the pattern x mapped to a linearization space and a vector w in the same space: $\langle w, \Phi(x)\rangle$. Let us choose the infinite dimensional linearization space to be the RKHS that is associated with a suitable kernel, for example the Gaussian kernel. The perceptron in this RKHS is then defined by the inner product between two functions. The pattern x is preprocessed by the function $\Phi(x) := k(\cdot, x)$ which maps x to the function that describes its similarity to all other stimuli. The role of the weight vector w in the classical perceptron (10) is taken by the coefficients for a function in the RKHS. Let us denote this function with f. As before—see equation (17)—we denote the coefficients that f takes in the expansion given by the kernel functions with w. With this notation an infinite dimensional perceptron can be written as:

$$\langle f, \Phi(x)\rangle_{\mathcal{H}} = \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = \sum_{i=1}^{N} w_i k(x, x_i) = f(x). \tag{22}$$

As $k(\cdot, x)$ is the representer of evaluation (17), the inner product of f with the pattern x mapped to the linearization space is just f evaluated at x. If $f(x)$ is greater than a threshold the pattern x is categorized as one class otherwise as the other class.

The learning problem for an infinite-dimensional perceptron is then to find a suitable function f in the RKHS that can categorize the patterns under consideration correctly. It may seem that this is a difficult problem because we have to find a function in an infinite dimensional space rather than a weight vector in a finite dimensional space as for the perceptron. However, there is a simple solution for f. Say, a subject wants to learn to discriminate between two different categories of stimuli. The subject is given a training set of N exemplars x_1, ..., x_N. Each stimulus has a class label that we denote with y_1, ..., y_N. The class label y_i for pattern x_i can be either $+1$ or -1, depending on which category x_i belongs to. We will treat the categorization problem like a regression problem. The aim of the category learner is to find a function f defined on the perceptual space such that $f(x)$ is $+1$ whenever x belongs to one class and -1 when x is in the other class. Instead of searching for the best function in the whole RKHS we will only consider a subspace of all functions in the RKHS and show that in this subspace there is a function that can solve the regression problem perfectly. The subspace we consider is all linear combinations of the kernel functions on the exemplars:

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i). \tag{23}$$

The output of this function can be thought of as calculating a weighted similarity to all exemplars. This function is a linear combination of kernel functions. As all kernel functions are in the RKHS their linear combinations are also in the RKHS. Therefore, for a fixed set of N exemplars their linear combination spans a subspace of the RKHS that is at most N-dimensional. We refer to this subspace as the span of the exemplars. The span of the exemplars seems to be a only a small subset of all the functions in the infinite dimensional RKHS but it contains a function that solves the regression problem perfectly.

Finding a function f of the form (23) that solves the regression problem means finding weights w_1, ..., w_N for the exemplars such that for all j: $f(x_j) = y_j$. We introduce an N-dimensional vector y for the N labels. As we are only interested in the values that the function f takes on the exemplars x_j (with j from 1 to N) we only need to evaluate (23) at these values: $\sum_{i=1}^{N} w_i k(x_j, x_i)$. Let us use a vector w for the weights and using the same notation as before we write K for the matrix that collects all pairwise evaluations of k on the exemplars. Hence, the weights that we seek should solve $y = Kw$. If K has full rank—as the Gaussian kernel for example guarantees—then K is invertible and

$$w = K^{-1}y. \tag{24}$$

There is a unique vector of weights that solves the classification problem perfectly. If K has full rank this is always possible irrespective of the set of exemplars and their category labels[1]. As there is a solution in the span of the exemplars we do not need to work with the infinite dimensional RKHS to find a solution for the categorization problem in the RKHS.

2.6. Neural networks. The solution to the categorization problem that was discussed in the previous section can be understood as a weighted similarity to the exemplars. Exemplar models like this one can be implemented as a neural network. Imagine a cell that by means of learning has become sensitive to a particular stimulus y—its preferred stimulus. It will still fire if another stimulus x is sufficiently

[1] However, note that K might be close to singular in practice.

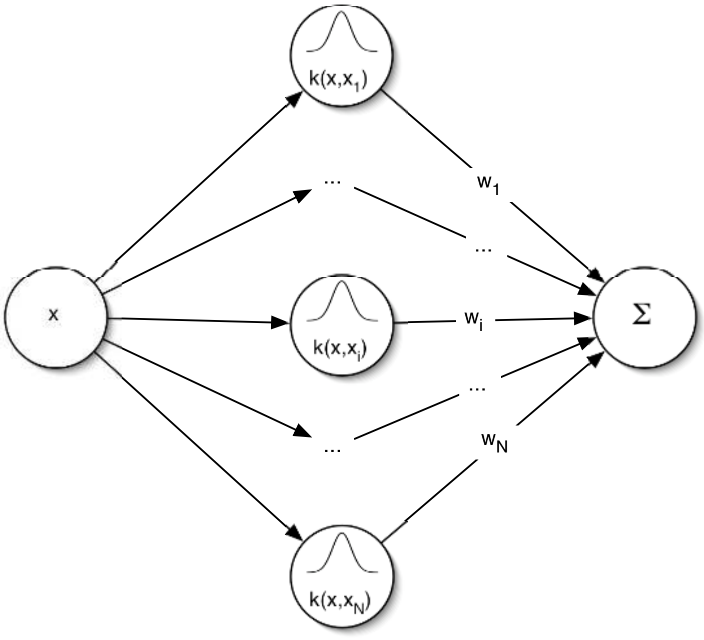

Figure 6. A RBF-network calculates a weighted sum over the responses of N cells with "tuning curves" given by k.

similar to the preferred stimulus. The way that the firing rate of the cell changes with changes in the stimulus is described by its tuning curve that can be modeled by a function like the Gaussian kernel[2]. A simple one layer neural network with several cells that are tuned to certain exemplars, x_1 to x_N, is depicted in Figure 6. Each cell responds to a stimulus x according to its tuning curve, given by k. The activity of all cells is collected by an output neuron that computes a weighted sum of its inputs. The function that this network implements was already given in equation (23).

In the neural network literature a function of the distance between two stimuli is called a radial basis function (RBF) kernel. The Gaussian kernel is the most prominent example for a radial basis function. As neural tuning curves are often found to have a shape like a radial basis function, RBF-networks have repeatedly been advocated as a model for brain function by Poggio and coworkers (Poggio & Girosi, 1989; Poggio, 1990; Poggio & Bizzi, 2004). They have also studied the link to reproducing kernel Hilbert spaces. From a mathematical view-point the problem they are addressing is the learning of an unknown function. Their approach motivates the use of kernels from a function approximation and regularization view.

3. Regularization

By using the exemplar network with as many free parameters as stimuli it is always possible to find weights such that the network can classify all training stimuli perfectly. The price for this flexibility is the danger of overfitting. A network may learn to categorize all training stimuli perfectly but only because it has learned the stimuli by heart. Any regularity in the data is overlooked in this way and therefore

[2] One important difference between tuning curves being modeled by a Gaussian and psychological similarity that is often also modeled by a Gaussian is of course that similarity is calculated between mental representations whereas tuning curves are usually based on physical measurements.

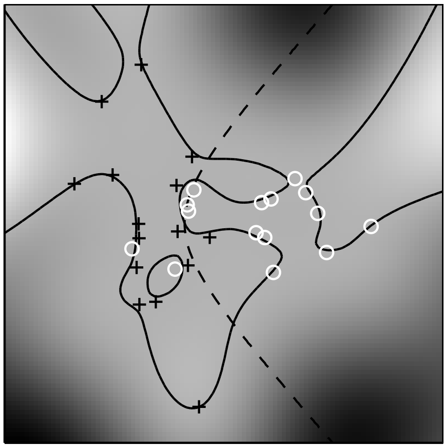

Figure 7. Stimuli from two categories are classified by an exemplar networks as given in Eq. (23) and Eq. (24). The solution is overfitted, that is the training stimuli can be categorized perfectly but generalization to new stimuli will be poor. The optimal decision bound for this category learning problem is shown as a dashed line.

the network will not be able to generalize. An example for overfitting is shown in Figure 7. Crosses and circles depict exemplars from two different categories. The two categories are defined by two overlapping Gaussian probability distributions. As we know the distributions that generate the stimuli we can calculate the optimal decision bound which for two Gaussians is a quadratic function (Ashby & Maddox, 1993). The dashed line is this optimal decision bound. The grayscale values depict the function f that was obtained by calculating exemplar weights with a Gaussian kernel as given in Eq. (24). The solid line is the contour line where $f(x) = 0$. On one side of this contour line the infinite dimensional perceptron would classify stimuli as belonging to one class, on the other side stimuli are classified into the other class. It can be seen that all training stimuli are categorized correctly. The resulting decision bound is obviously not very reasonable, and also very different from the optimal decision bound. Intuitively speaking, the decision bound that the exemplar network calculates is too complicated. The regression function f should not be allowed to vary so wildly and the decision bound should be smoother and less complicated.

One popular strategy to avoid overfitting is based on regularization (Bishop, 1995; Schölkopf & Smola, 2002; Poggio & Smale, 2003). In regularization there is an additional constraint on the function f that is sought: The function should not only fit the data it should also be smooth and not too "complex". To this end a penalty term is introduced that penalizes functions for being complex. Many modern model selection criteria can be seen as penalizing complexity (Pitt, Myung, & Zhang, 2002). Instead of only minimizing the error on the training exemplars, which can always be done perfectly, the error plus a penalty term is minimized. The penalty term is also called regularizer.

Let us denote the error that a function f makes on the training exemplars with $c(f)$. Possible examples for such a cost function are the number of misclassifications or the mean square error. If we denote the penalty term by Ω then the function that one seeks to minimize becomes

$$L(f) = c(f) + \Omega(\langle f, f \rangle_{\mathcal{H}}). \tag{25}$$

The penalty term is chosen as a strictly increasing function Ω of $\|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}$, the squared norm of the function in the RKHS that the kernel defines. The learning problem is now understood as trying to find a function f that minimizes L.

3.1. The representer theorem.

Ideally, one would like to find the function f that minimizes the regularized error L over all functions in the RKHS. Perhaps surprisingly, the optimal function can be represented as an expansion of the exemplars as given in Eq. (23). This result is called the *representer theorem*. It shows that for a large class of problems the optimal solution over all functions in the RKHS lies in the span of the exemplars. If the aim of a function learner (be it a brain or a machine) is to minimize a regularized loss then it makes sense to restrict the learning mechanism to an exemplar network of the form (23). The proof that is given by Schölkopf, Herbrich, and Smola (2001) is short, and it illustrates the power of the RKHS-view of kernels.

As f is in the RKHS we can split it up into a part $f^{\|}$ that lies in the span of the exemplars (23) and a part $f^{\perp}$ that is orthogonal to it. For the first term $c(f)$ in the regularized loss function $L(f)$ we need to evaluate $f = f^{\|} + f^{\perp}$ only on the exemplars x_1, ..., x_N. Remember that $k(x_i, \cdot)$ is the representer of evaluation for x_i and therefore (see 17)

$$f(x_i) = f^{\|}(x_i) + f^{\perp}(x_i) = f^{\|}(x_i) + \left\langle f^{\perp}, k(x_i, \cdot) \right\rangle_{\mathcal{H}}.$$

The second term is zero because by definition $f^{\perp}$ is orthogonal to $k(x_i, \cdot)$. Hence, the cost function $c(f)$ is independent of $f^{\perp}$. For the penalty term note that

$$\begin{aligned} \Omega(\langle f, f \rangle_{\mathcal{H}}) &= \Omega\left(\left\langle f^{\|}, f^{\|} \right\rangle_{\mathcal{H}} + 2\left\langle f^{\|}, f^{\perp} \right\rangle_{\mathcal{H}} + \langle f^{\perp}, f^{\perp} \rangle_{\mathcal{H}}\right) \\ &= \Omega\left(\left\langle f^{\|}, f^{\|} \right\rangle_{\mathcal{H}} + \langle f^{\perp}, f^{\perp} \rangle_{\mathcal{H}}\right). \end{aligned}$$

For a minimizer of L the orthogonal part $f^{\perp}$ has to be zero. To see this assume f is a minimizer but $f^{\perp}$ is not zero. As Ω is strictly increasing L can be decreased by choosing $f^{\perp}$ to be zero and hence f was not a minimizer—in contradiction to the assumption. Therefore, the best function in the whole RKHS is given by an expansion of the exemplars (23).

The importance of the representer theorem is that if the regularizer can be cast as a strictly increasing function Ω of $\langle f, f \rangle_{\mathcal{H}}$ then the optimal solution over the whole RKHS is a linear combination of kernel functions centered on the exemplars. Therefore, it is not necessary to work in the infinite dimensional function space to find the best function in it. To find the best function one only has to adjust the exemplar weights. Furthermore, the single best function can often be found analytically or with simple numerical procedures. All this makes exemplar networks so attractive for machine learning and perhaps this can also provide a theoretical justification for assuming a psychological mechanism that is based on storing exemplars. Exemplar models in psychology simply assume that all exemplars are stored without giving a rational explanation why this should be done. If the objective of a subject could be phrased in terms of a regularized loss of the above form (as it is often done is statistics and machine learning) then, as we know that the optimal solution lies in the span of the exemplars, we would have an argument for using exemplars in the first place.

3.2. Regularization example. To understand the rationale behind regularization better and to appreciate the representer theorem it is helpful to look at a more concrete example.

Let us assume the learning proceeds by trying to find a regression function f that minimizes a certain loss $L(f)$ that has two components: One component $c(f)$ that depends on the error on the training stimuli and a component $\Omega(\langle f, f\rangle)$ that penalizes the function's complexity (see Eq. 25, above). As a measure for the error that the learner makes on the training examples we will take the mean square error: $c(f) := \sum_{j=1}^{N}(f(x_j) - y_j)^2$. The squared loss is not the most reasonable loss function for a categorization problem. ALCOVE, a prominent exemplar model, for example, uses a different loss function (Kruschke, 1992). Intuitively, it seems better to minimize misclassifications directly and from a statistical view-point one would want to minimize the negative log likelihood. However, we have chosen to minimize the mean square error because it is conceptually easier. For the penalty term Ω we have chosen the simple case where it is linear with a positive parameter λ. The loss function (25) then becomes:

$$L(f) = \sum_{j=1}^{N}(f(x_j) - y_j)^2 + \lambda \langle f, f\rangle_{\mathcal{H}},$$

where y_j is the value that the function should output on exemplar x_j. The parameter λ can be thought of as controlling the trade-off between a good fit and the penalty.

Above we have—perhaps a bit hand-wavingly—referred to the effect of regularization as penalizing "complexity". We can understand what the regularization in Eq. (25) does by looking at the form of the penalty term which depends on the squared norm $\langle f, f\rangle_{\mathcal{H}}$ of the function f. Because of the representer theorem the optimal function is of the form (23) and we can rewrite the squared norm of the optimal function as

$$\begin{aligned}
\langle f, f\rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{N} w_i k(x_i, \cdot), \sum_{j=1}^{N} w_j k(x_j, \cdot) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j k(x_i, x_j) \\
&= w^T K w,
\end{aligned}$$

where we have used the linearity of the inner product and the reproducing property. In the regularizer the weights of the function f are multiplied by the similarity of the respective exemplars. In order to make two very similar exemplars output very different function values it is necessary to use very big exemplar weights. As large weights for similar exemplars can only be achieved at a high penalty the regularized function respects the similarity measure better than the non-regularized solution in Figure 7.

As we know by the representer theorem that the optimal function is of the form (23) minimizing $L(f)$ is equivalent to finding exemplar weights w such that

$$L(f) = \sum_{j=1}^{N}(\sum_{i=1}^{N} w_i k(x_i, x_j) - y_j)^2 + \lambda(w^T K w)$$

$$\begin{aligned} &= (Kw - y)^T(Kw - y) + \lambda(w^T Kw) \\ &= w^T(KK + K\lambda)w - 2y^T Kw + y^T Ky \end{aligned}$$

is minimal. The optimal weights can be found analytically by differentiating this quadratic loss function with respect to w. Setting to zero to find the optimum leads to the following solution for w:

$$\begin{aligned} 2(KK + K\lambda)w - 2Ky &= 0 \\ (KK + K\lambda)w &= Ky \\ (K + \lambda I)w &= y \\ w &= (K + \lambda I)^{-1}y. \end{aligned}$$

The Hessian of the quadratic function L is given by $(KK + K\lambda)$. As the Hessian is a sum of two positive definite matrices it is positive definite itself. Therefore, the solution for w is the unique minimum. A regularized solution for f for the same problem as in Figure 7 is shown in Figure 8. The regularized solution is less complicated and looks more reasonable than the non-regularized solution.

The regularized solution is also known as ridge regression. The non-regularized solution (24) can be recovered from ridge regression by setting the regularization parameter λ to zero. In this case the weights are simply calculated by taking the inverse of K. In the case where λ is large and $(K+\lambda I)$ is dominated by the diagonal matrix λI the weights all have the same absolute value and only vary in their sign. The decision bound in this case is identical to the kernel density estimators given in Eq. (21). In this extreme the decision bound is only determined by the similarity measure and not by the exemplar weights. Hence, the regularization parameter λ makes it possible to choose a solution that is in-between the two extremes: A weight based solution (24) that will always overfit and a similarity based solution (21) with all exemplar weights set to the same value. By allowing this extra flexibility it is often possible to achieve a better generalization performance than by relying on similarity alone. Of course, the value of the regularization parameter λ has to be chosen wisely. In machine learning this is considered as a model selection problem. One way to find a suitable regularization parameter is by using cross-validation and this is what we have done in Figure 8, too (Schölkopf & Smola, 2002; Pitt et al., 2002). Both, the chosen kernel and the regularization parameter, will determine how well the classifier will generalize to new patterns. It is in the construction of the kernel, however, that engineers can use their insights about a classification problem and their intuitions about the similarity of patterns.

4. Conclusions

We have introduced kernel methods as they are used in machine learning. The most important results here are the kernel trick and its link to reproducing kernel Hilbert spaces. On the way we have hinted to parallels with psychological theories. First, kernel methods can be implemented as a one-layer neural network. Second, the Gaussian kernel can be interpreted as a similarity measure and representation of the stimuli in a RKHS can be seen as representing the stimuli via their similarity to all other stimuli. Third, the most simple exemplar model of categorization is a prototype classifier in this RKHS. Fourth, regularization can be used to avoid overfitting. And fifth, the representer theorem shows that the best regularized function in RKHS can often be found in the span of the exemplars.

In the next chapter (Chapter 3), we describe how the RKHS framework arises naturally from Shepard's universal law of generalization. Shepard's law is closely related to geometric models of similarity and multidimensional scaling. Geometric

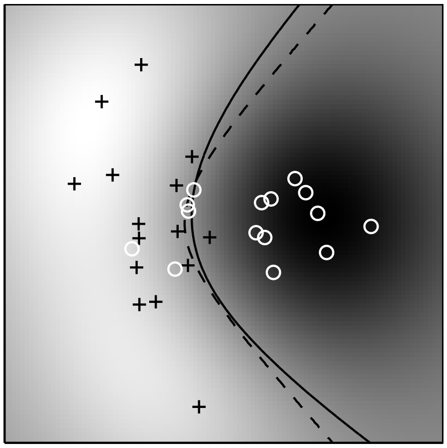

Figure 8. The same categorization problem as in Figure 7 but this time a regularized solution is shown. The regularized decision bound (solid line) is quite close to the optimal decision bound (dashed line).

models have been heavily criticized by Tversky and co-workers (Beals et al., 1968; Tversky, 1977; Tversky & Gati, 1982). One of their major criticisms concerns the assumption of the triangle inequality that is implicit in all geometric models. However, from the data that is available the triangle inequality is unproblematic as an assumption as long as it is not paired with a second assumption that is known as segmental additivity. The RKHS provides an elegant framework for metric models without segmental additivity.

From this chapter it should be obvious that exemplar models and kernel methods are based on the same ideas. Their relationship is discussed in greater detail in Chapter 4. Briefly, two very prominent exemplar models, the Generalized Context Model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992), both make use of kernels but in different ways. Only ALCOVE can be mapped directly to a machine learning method and even exhibits some regularization.

We imagine other potential uses for the mathematical tools we have presented. First of all, we hope that this part of the thesis opens up the recent machine learning literature for more psychologists. Many of the data analytic methods presented in machine learning could be used in psychology—irrespective whether they use the RKHS framework that was the focus here. For example, machine learning methods have been used in psychophysics (Wichmann, Graf, Simoncelli, Bülthoff, & Schölkopf, 2005; Graf, Wichmann, Bülthoff, & Schölkopf, 2006) and we believe that many more applications will follow. With regard to the Hilbert spaces one can be skeptic whether the infinite dimensional machinery is really necessary for psychological modeling. However, there are many cases where the data that is collected is in terms of functions and therefore naturally described with methods similar to the ones described here (Ramsay & Silverman, 1997). Furthermore, several other authors have also recently suggested to use infinite dimensional spaces in the theoretical analysis of behavior (Drösler, 1994; Townsend, Solomon, & Smith, 2001; Zhang, 2006). In the laboratory, stimuli are almost always defined by a small number of independent variables, and those are the stimuli that we used as examples in this paper. In these examples the approach was to map stimuli from a space with a small number of dimensions to an infinite dimensional space. More realistic stimuli will vary in a plethora of ways and not just along a small number of well-defined dimensions. Infinite dimensional spaces could be attractive for describing such natural stimuli—take for example the features of a face, the shape of a leaf, or the spectrum of a light source. Also the number of channels that humans use to analyze these stimuli might be very large—too large to enumerate them explicitly. The tools we presented in this chapter might also be useful for this enterprise.

Appendix: More RKHS examples

This section provides more examples for the construction of reproducing kernel Hilbert spaces.

Quadratic kernel. The quadratic kernel that we used as a motivating example was defined as $k(x, y) = \langle x, y \rangle^2$. Let us briefly check that it really satisfies the conditions for a positive definite kernel. First note that k is real-valued and symmetric. We then need to check that all kernel matrices are positive semi-definite. This can easily be done because the kernel was constructed to be the inner product in a three-dimensional vector space. This vector space was obtained by the map Φ as defined in Eq. (9). Let us consider N points x_1, ..., x_N in the original space. We map these N points to the three-dimensional linearization space where they have coordinates $\Phi(x_1)$, ..., $\Phi(x_N)$. The coordinates of all points can be collected in a $3 \times N$ matrix that we will just call Φ (the context should make clear whether we mean the function or the matrix Φ). Now, the matrix K that collects all pairwise inner products between all $\Phi(x_1)$, ..., $\Phi(x_N)$ is given by the $N \times N$ matrix $\Phi^T\Phi$. This matrix is positive semi-definite because for all w it holds that

$$w^T K w = w^T \Phi^T \Phi w = (\Phi w)^T (\Phi w)^T \geq 0.$$

The rank of K is at most 3 and therefore the kernel matrix K does not have full rank for $N > 3$. There are vectors other than the zero vector that make the quadratic form $w^T K w$ zero. Therefore, the kernel matrices of the quadratic kernel are only positive semi-definite and not positive definite.

Step 1: Constructing a vector space. We already know that the quadratic kernel is an inner product in a vector space—this is how we constructed it—but it is instructive to derive a linearization space directly from the kernel. The purpose is to illustrate what the construction in the main text does, and why the construction also works for the quadratic kernel. Remember that the quadratic kernel was defined by

$$k(x, y) = \langle x, y \rangle^2 = x_1^2 y_1^2 + \sqrt{2} x_1 x_2 \sqrt{2} y_1 y_2 + x_2^2 y_2^2.$$

The function space that we will construct is given by linear combinations of different generating functions (Eq. 14). Each function in the space is of the form $f(x) = \sum_i^N w_i k(x, x_i)$. The space that is spanned by this generating system of kernel functions is the three-dimensional space of all functions that are given by

$$f(x) = x_1^2 u_1 + \sqrt{2} x_1 x_2 u_2 + x_2^2 u_3. \tag{26}$$

There are many bases in which this space could be expressed. We have chosen this one because it is particularly convenient as will be seen below. To show that the space that is spanned by this basis is the same space that is spanned by the generating system $f(x) = \sum_i^N w_i k(x, x_i)$ we need to demonstrate that each function that can be expressed as a linear combination of generating functions is also expressible in the suggested basis and vice versa. First, note that every possible generating function $k(\cdot, y)$ can be expressed in the suggested basis with the following coefficients: $u = (y_1^2, \sqrt{2} y_1 y_2, y_2^2)^T$, because the function $f(x) = k(x, y)$ has these coefficients when interpreted as a function of x with fixed parameters y. Hence, every function that is a linear combination of some generating functions $k(\cdot, x_i)$ can also be expressed in this basis. Second, every function in the suggested basis can also be expressed as an expansion of the form $f(x) = \sum_i^N w_i k(x, x_i)$. In fact, each function in the suggested basis can be expressed by a set of three generating functions that also form a basis of the space. We take $k(\cdot, a)$, $k(\cdot, b)$ and $k(\cdot, c)$ with $a = (1, 0)^T$, $b = (1, 1)^T$ and $c = (0, 1)^T$. With this choice

$$\begin{aligned}
f(x) &= w_1 k(x, a) + w_2 k(x, b) + w_3 k(x, c) \\
&= w_1 \left(x_1^2 a_1^2 + 2x_1 x_2 a_1 a_2 + x_2^2 a_2^2\right) + \\
&\quad\; w_2 \left(x_1^2 b_1^2 + 2x_1 x_2 b_1 b_2 + x_2^2 b_2^2\right) + \\
&\quad\; w_3 \left(x_1^2 c_1^2 + 2x_1 x_2 c_1 c_2 + x_2^2 c_2^2\right) \\
&= x_1^2(w_1 + w_2) + \sqrt{2} x_1 x_2 (\sqrt{2} w_2) + x_2^2 (w_2 + w_3)
\end{aligned}$$

and therefore each function that is expressed with a vector u in the basis suggested above (Eq. 26) is also expressible as a linear combination of the three functions $k(\cdot, a)$, $k(\cdot, b)$ and $k(\cdot, c)$—with

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 1 & 1 \end{pmatrix} \text{ and } A^{-1} = \begin{pmatrix} 1 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 1 \end{pmatrix} \tag{27}$$

we can change the basis by using $u = Aw$ and $w = A^{-1}u$.

Note that the the generating system of functions is very big. In fact, there are uncountably many generating functions that could be used in an expansion of the form $f(x) = \sum_i^N w_i k(x, x_i)$. Recall that the arguments of k are from $\mathbb{R}^2$ and hence there are as many potential $k(\cdot, x_i)$ as there are points in the plane. However, this big generating system of functions only spans a three-dimensional space.

Step 2: Constructing an inner product. Now that we know what the function space that is constructed by a linear combination of kernel functions looks like, consider the inner product that is defined by Eq. (15). Take two functions f and g from this space. Both have an expansion in terms of the generating functions. Let us take expansions given by the three basis functions $k(\cdot, a)$, $k(\cdot, b)$ and $k(\cdot, c)$ and denote the coefficients for f and g by v and w, respectively. For the inner product we need the kernel matrix K that is given by all pairwise evaluations of $k(x, y) = \langle x, y\rangle^2$ on $a = (1, 0)^T$, $b = (1, 1)^T$ and $c = (0, 1)^T$. The inner product is then

$$\langle f, g\rangle_{\mathcal{H}} = w^T K v = w^T \begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix} v. \tag{28}$$

The inner product is the same as the standard inner product in the basis used in Eq. (26). The matrix A, shown in Eq. (27), transforms the vectors w and v to the basis of Eq. (26). As $K = A^T A$, which can easily be checked, we immediately see that $w^T K v = (Aw)^T (Av)$, and therefore K corresponds to the standard inner product in this basis.

Step 3: Constructing a linearization function. The linearization function that we seek is $\Phi(x) = k(\cdot, x)$. Each point x is mapped to a function. In the basis given by Eq. (26) this function is described by three coefficients: $\Phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)^T$. The standard inner product in this basis implements the kernel condition (Eq. 12). To see this take a second point y and map it to the linearization space: $\Phi(y) = (y_1^2, \sqrt{2} y_1 y_2, y_2^2)^T$. With this the inner product $\langle \Phi(x), \Phi(y)\rangle =$

$\Phi(x)^T\Phi(y) = k(x,y)$ directly gives the definition of the quadratic kernel. Our example that motivated the introduction of kernels used exactly this linearization function (Eq. 9). However, it has to be emphasized that even though the linearization function and the linearization space are exactly the same as before a new interpretation is gained. The points in the linearization space are interpreted as coefficients for a linear combination of basis functions—each point represents a function. The beauty of the construction of the RKHS as given in the main text is that it does not depend on a basis for the function space. Hence, even if we did not know that all functions can be expressed using Eq. (26) we would still be able to map a point into the function space by using $\Phi(x) = k(\cdot, x) = \langle \cdot, x\rangle^2$, and we would also know that $k(x,y)$ calculates the inner product in this space. In this example, we have just made the space and the inner product that are induced by the quadratic kernel explicit.

A finite dimensional RKHS. The variables x and y in the construction of the RKHS that is given in the main text are usually thought of as taken from a real vector space. Therefore, there are uncountably many different possible values for x. In practice there are only going to be a finite number of patterns anyway and it is instructive to see what the RKHS looks like in this case. Consider a finite set of patterns and number them from 1 to N. These objects need not be taken from a vector space. Each real-valued function on these N objects can be described by a real vector f that collects the function values f_1, ..., f_N. As the RKHS that we will construct is a space of functions over the finite number of stimuli a finite-dimensional RKHS is an N-dimensional vector space. In a slight abuse of notation, but in order to make the parallels to the derivation in the main text more transparent, we will sometimes use x and y as indices in place of i and j. Hence, if we want to evaluate the the function f on a pattern x we write this as f_x. Similarly, if we evaluate f on another pattern y this is denoted as f_y.

A positive definite kernel function that is defined over a finite set is completely described by a symmetric and positive semi-definite $N \times N$ matrix, that we denote with K. Here, we make the stronger assumption that K has full rank and is therefore positive definite. This makes the discussion a bit simpler and also makes sure that we can describe all possible functions on the patterns. The task is to find a function Φ that maps each pattern to the N-dimensional space of functions on the patterns. The map should be such that the kernel implements the inner product in this space (12). Let us denote the vector that is assigned to the pattern x with Φ_x. For a given positive definite matrix K and all x and y it should hold that

$$k_{xy} = \langle \Phi_x, \Phi_y \rangle_{\mathcal{H}} \,. \tag{29}$$

This should be considered as the finite analogue of equation (12). Thus, we are looking for an inner product and N-dimensional vectors Φ_1, ..., Φ_N such this holds true. There is a straightforward solution to this. Let Φ be the $N \times N$ matrix that is constructed by concatenation of the column vectors Φ_1, ..., Φ_N. If we take the inner product to be simply the standard inner product the kernel condition can be written as $K = \Phi^T\Phi$. As K is positive definite, by assumption, it can always be decomposed in such a way. We can, for example, use the eigenvalue decomposition to construct a Φ. However, in this example we want to follow the construction as it is given in the main text. For this we need to follow three steps, constructing a vector space, an inner product and a linearization function.

Step 1: Constructing a vector space. Remember that we denoted the matrix of all pairwise evaluations of k on N stimuli with K. First, let us make sure we define the vector space properly. If we take as a basis the rows of K we can express any vector f representing any function values f_1, ..., f_N on the patterns as

$$f_x = \sum_{i=1}^{N} w_i k_{ix},$$

that is an expansion with a vector of coefficients w. Note the relation of this expansion to the expansion in (14). We can write this compactly as $f = Kw$. You can see what would happen if we had not assumed that K is full rank. In this case K could not span the full N-dimensional space that we need to describe all possible function values for N patterns.

Step 2: Constructing an inner product. We define the following inner product between a vector f with coefficients w and a vector g with coefficients v in the basis given by the rows of K:

$$\langle f, g \rangle_{\mathcal{H}} = \langle w, v \rangle_K = w^T K v = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i v_j k_{ij},$$

which clearly defines an inner product because K is, by assumption, positive definite. In the finite case the inner product defined in Eq. (15) that is defined between functions becomes an ordinary inner product (6).

Step 3: Constructing a linearization function. Let us denote the x^{th} column of the identity matrix by δ_x (this notation should remind the reader of Dirac's delta distribution); it is one for the x^{th} dimension and zero otherwise. In a further stretch of notation let $k_{\cdot x}$ denote the function that corresponds to the x^{th} column of the (symmetric) kernel matrix K. Hence, $k_{\cdot x}$ is the representer of evaluation:

$$\langle k_{\cdot x}, f \rangle_{\mathcal{H}} = \langle \delta_x, w \rangle_K = \delta_x^T K w = \delta_x^T f = f_x.$$

The reproducing property is also fulfilled in the basis given by K because

$$\langle k_{\cdot x}, k_{\cdot y} \rangle_{\mathcal{H}} = \langle \delta_x, \delta_y \rangle_K = \delta_x^T K \delta_y = k_{xy}.$$

In fact, it is trivially fulfilled because the inner product is directly given by the positive definite matrix K.

The representation that we seek for the x^{th} stimulus is, as before, given by the kernel function $\Phi_x := k_{\cdot x}$. With this choice the kernel condition (29) is satisfied. In the basis given by K the representation for the x^{th} stimulus is the x^{th} column of the identity matrix. This is, however, not the only solution. It is very instructive to see what the inner product we defined looks like in some other bases.

4.0.1. *Different bases.* We can change the basis and get a different set of vectors with a different inner product that still satisfy the kernel condition (29). Of particular interest is the basis which turns the inner product as given by K into the standard inner product. This basis can be found by the eigenvalue decomposition of $K = \Psi^T \Lambda \Psi$, as given in Eq. (9):

$$\langle f, g \rangle_{\mathcal{H}} = \langle w, v \rangle_K = \left\langle \sqrt{\Lambda} \Psi w, \sqrt{\Lambda} \Psi v \right\rangle.$$

Hence, if we choose Φ_x and Φ_y to be the x^{th} and y^{th} column of $\sqrt{\Lambda}\Psi$, and if we further take the standard inner product the kernel condition $k_{xy} = \langle \Phi_x, \Phi_y \rangle$ is also satisfied.

To understand the RKHS better consider yet another basis for the functions f on N patterns. Apart from the two bases that we have discussed so far (the basis given by the rows of K and the one we obtained from the eigenvalue decomposition of K) there is a third interesting basis for the function space defined on the patterns 1, ..., N. This basis is given by the standard basis where the coefficients are simply the function values f_1, ..., f_N themselves. A natural question is how the inner product as given by k acts on the actual function values. Take again the vector f

specified in the basis given by K. Let w be the weights of f such that the vector of function values is $f = Kw$ and let v be the weights of another vector g such that $g = Kv$. Then the inner product of f with g is

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}} &= \langle w, v \rangle_K \\ &= w^T K v = (K^{-1} f)^T K (K^{-1} g) \\ &= f K^{-1} g = \langle f, g \rangle_{K^{-1}} . \end{aligned}$$

Hence, in the standard basis the inner product that is defined by k corresponds to the matrix K^{-1} which is positive definite because it is the inverse of a positive definite matrix. It is very important to realize that we could define many different inner products directly on the basis in which f and g are specified, for example one could just take the standard inner product $f^T g$ or one could use $f^T K g$. The inner product that we constructed above calculates the inner product $f^T K^{-1} g$. It is special in the sense that in the basis given by K the inner product is also K, and this results in the reproducing property.

CHAPTER 3

Similarity

If we understood the processes underlying the perception of similarity we would have a handle on many psychological phenomena. In perceptual organization, the Gestalt-law of similarity underlies perceptual grouping. In categorization, similar objects form concepts. In learning, transfer depends on the similarity of training and test items. Appealing to similarity is, of course, not enough to explain these phenomena. We need to have a clear conception of what similarity is in order to use it as an explanatory construct—and despite the fact that it seems intuitively clear what is meant by similarity, some researchers have argued that the concept is almost too flexible to be of any use. Nevertheless, similarity continues to be a central concept in many psychological theories (Medin et al., 1993).

The most influential approach to similarity has been geometrical. The central idea in this approach is that stimuli are represented in a perceptual space and the distance between stimuli in this space determines their similarity. In the simplest case the space is assumed to be Euclidean and the similarity of stimuli decreases with their distance in the space. Multidimensional scaling (MDS) is a class of algorithms that makes it possible to reconstruct the coordinates in the putative perceptual space from similarity data, for example similarity ratings or confusion probabilities. Shepard (1987) argued that the best experimental measure for similarity are generalization gradients. He further presented several datasets that indicated that generalization gradients are an exponential function of the distance in perceptual space. This relationship between generalization gradients and perceptual spaces is usually referred to as Shepard's *universal law of generalization*.

In a well-known series of papers Tversky and colleagues have challenged the idea of a geometric representation (Beals et al., 1968; Tversky, 1977; Tversky & Gati, 1982). They provided convincing evidence that geometric representations cannot account for many human similarity judgments. Even though the criticism has been substantial, MDS has been used in practice with considerable success (Townsend & Thomas, 1993). Categorization models in particular have relied heavily on geometric representations—seemingly unfazed by Tversky's criticism (Nosofsky, 1986). In this cahpter we will reconcile Tversky's critique with Shepard's universal law of generalization. Read carefully, Tversky's most fundamental critique does not exclude the possibility of a metric perceptual space, it only attacks the commonly used metrics with additive segments (Tversky & Gati, 1982). This class includes Euclidean spaces, spaces with a Minkowski p-norm and curved Riemannian geometries. We will introduce a representation of the perceptual space that arises naturally from Shepard's law and that is not affected by Tversky's criticism. This representation has several psychologically interesting properties: It does not have additive segments, it is bounded and it represents stimuli by its similarity to all other stimuli (Edelman, 1998). It is based on the mathematical theory of reproducing kernel Hilbert spaces that can be used to model the similarity of stimuli as inner products (as presented in the previous chapter).

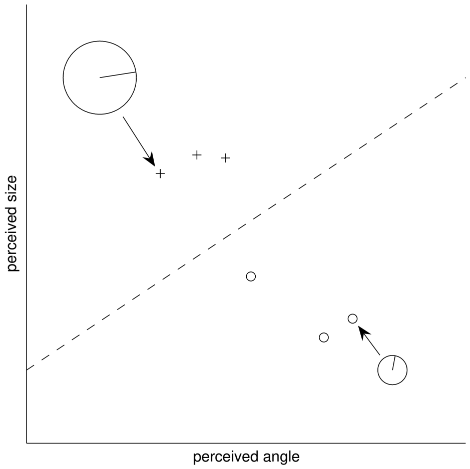

FIGURE 1. An illustration of a perceptual space. Stimuli are circles with a spoke and can vary on two dimensions. Two artificial categories are depicted, separated by a linear decision bound.

1. Perceptual spaces

1.1. Multidimensional scaling and categorization. There have been early attempts to model similarity judgments as inner products. Ekman's group in Stockholm made the assumption that stimuli are represented in the mind as vectors in a multidimensional Euclidean space and that the similarity of the points is given by their inner product. There are several variants that were carefully distinguished. For example, one could take the inner product directly or, alternatively, the angle between the vectors, or yet another possibility, the projection of one vector onto the other. Experimentally, the latter was not considered a similarity judgment but a containment judgment which has the interesting property that it is not symmetric (for an overview on this approach see Gregson, 1975; Borg & Groenen, 1997).

A little earlier, Torgerson (1952) presented a method that is now widely known as classical multidimensional scaling. Instead of requiring a direct measurement of similarity this method indirectly determined the dissimilarity between stimuli by using the method of triads. Under the assumption that the dissimilarity is linear with the distance in a Euclidean space it is possible to reconstruct the coordinates of the stimuli in the perceptual space using a procedure that was suggested by Young and Householder (1938).

The idea that stimuli can be represented as points in a perceptual space has had a major impact on categorization models. All the early work on MDS has assumed that the perceptual space behaves like Euclidean space. This assumption has great intuitive appeal because it allows researchers to visualize subjects' presumed representation of the stimuli. Consider, as an example, the following popular stimuli: Circles with a spoke (Shepard, 1964). Two examples are shown as inlets in Figure 1. These stimuli can vary on two obvious dimensions. By varying the radius of the circle stimuli of different perceived sizes are produced. By varying the angle of the spoke stimuli of different perceived angles are generated. Figure 1 shows the perceptual space that is defined in this way. Each point in the plane represents the perception of one of the stimuli. The x-axis depicts the perceived angle of a stimulus and the y-axis depicts the perceived size of a stimulus. The perceived size and angle are of course different from the physically specified size and angle.

To illustrate categorization in perceptual spaces we have plotted two clusters of three stimuli each. The first cluster consists of large stimuli with spokes pointing to the right (crosses), whereas stimuli in the second cluster are smaller with spokes pointing upward (circles). It is very tempting to draw a line (not necessarily straight) that separates the two clusters in order to explain a subject's categorization behavior (Ashby & Gott, 1988). The perceptron, for example, implements a linear decision rule (Rosenblatt, 1958). By comparing the similarity to the mean of the stimuli in each cluster the prototype classifier also leads to a linear decision bound (Posner & Keele, 1968; Reed, 1972). Exemplar theories postulate a perceptual space, too, but can explain more complicated decision rules based on the similarity to all the stimuli that were shown to the subject (Nosofsky, 1986; Kruschke, 1992).

1.2. Dimensions and metrics. While the idea of a perceptual space seems intuitively plausible its theoretical foundations are far from compelling (Townsend & Thomas, 1993). The problems already start with the definition of a dimension. Often the dimensions are thought to be fixed by sensory processing, as if they were direct readings of the sensors. In the case of the circles with spokes there is a tacit agreement between experimenter and participants that size and angle are the dimensions of interest, and not for example color. Even if the stimuli are simple it is often not clear along which dimensions subjects perceive the stimuli. For example, for rectangles there has been a debate whether the perceptual dimensions follow the length of the sides or area and aspect ratio (Krantz & Tversky, 1975). An alternative to the fixed dimensions approach would be that the features that participants use for categorization are formed as a process of learning (Schyns, Goldstone, & Thibaut, 1998).

Even if the dimensions on which a subject perceives some stimuli are assumed to be known we will not know how the dimensions are combined to form an overall percept of similarity—but this combination of dimensions is necessary to form a decision of category membership in all the categorization models mentioned above. A form of metric is needed that measures the closeness of stimuli in perceptual space. Without such a metric the term perceptual space would seem to be rendered vacuous. In Figure 1 we have depicted the perceptual space with two dimensions in the plane. This makes it tempting to simply use Euclidean distance in the plane as a metric. However, this choice is not justified psychologically, it is only a choice of convenience that is made because Euclidean geometry is so intuitive.

As the two dimensions are drawn orthogonally it is also tempting to think that the two dimensions are perceptually independent. There are many (mostly operational) definitions of independence in the literature. Ashby and Townsend (1986), for example, have characterized the notion of perceptual independence based on signal detection theory and perceptual noise. Other definitions are based on the intuition that it should be possible to change the perception on one dimension without changing the perception on the other dimension. Two related concepts in the literature on similarity are integral and separable dimensions. The circles with a spoke are a typical example of separable dimensions. It is possible to attend to each dimension alone independent of the other. For integral dimensions stimuli form a unitary percept and the underlying dimensions cannot be attended to without interference from variations in the other dimension. Hue and saturation in color space are often cited as an example for integral dimensions. Incidentally, separable dimensions have been associated with the city-block metric and integral dimension with the Euclidean metric. While the city-block metric and the Euclidean metric are the most commonly used metrics they are by no means the only possible choices (Shepard, 1964; Garner, 1974).

With the assumptions of a perceptual space in place MDS can be used to recover underlying dimensions and the configuration of stimuli. To this end data on the similarity of different stimuli is collected. MDS then tries to embed the stimuli in a metric space such that the similarity relationship in the data is preserved as well as possible: Stimuli which are measured to be highly similar should be very close together in space, and stimuli which are measured to be very dissimilar should have a large distance between them. In practice MDS with either the Euclidean or the city-block metric in a low-dimensional space works surprisingly well and often leads to interpretable results.

1.3. Experimental measures of similarity. Before an MDS analysis can be undertaken an appropriate experimental measure of similarity needs to be found. Direct measures can be used for this. For example, subjects can be asked to judge the dissimilarity of stimuli on a rating scale from one to seven. Alternatively one could use the method of triads, where participants have to choose which of two comparison stimuli is more similar to a reference stimulus. If one has reason to believe that one's measure of dissimilarity is on an interval scale one can try to directly use the dissimilarity measure (plus an additive constantd) as a distance in a space, as it is done in classical MDS (Torgerson, 1952).

Following the lead of Shepard the categorization literature has often relied on indirect measures of similarity. Shepard (1987) argued that generalization gradients should be used to measure similarity and this is the stance that is taken in almost all exemplar models. In classical conditioning generalization gradients are obtained by conditioning on a certain stimulus and measuring the response to related, but different, stimuli (Ghirlanda & Enquist, 2003; Mostofsky, 1965). For example, a dog could be conditioned to salivate in response to a 1000 Hz tone. The generalization gradient is obtained by measuring the salivation of the dog in response to neighboring frequencies. Not surprisingly the generalization to new stimuli is higher the more similar the new stimuli are to the conditioned stimulus. Intuitively one would like to explain generalization in terms of psychological similarity and indeed researchers have tried to obtain measures of similarity that are independent of any generalization behavior (e.g. by integrating just-noticeable differences). In animal studies, however, this proved to be hard and led Bush and Mosteller (1951, p. 413) to conclude: "Although there are several intuitive notions as to what is meant by 'similarity', one usually means the properties which give rise to generalization. We see no alternative to using the amount of generalization as an operational definition of degree of 'similarity'."

If generalization gradients were the best measure to assess similarity, Shepard reasoned, then generalization gradients should be used in the construction of a perceptual space. In psychophysics researchers tried to measure perceptual spaces in units of just-noticeable differences, Shepard suggested to use generalization gradients instead[1]. With overlapping measurements of generalization gradients it is indeed possible to construct a perceptual space that is uniform with respect to the generalization gradients (Shepard, 1965). The distance in perceptual space is related to the similarity of stimuli by the amount of generalization exhibited. The same distance means the same amount of generalization. Ordinal MDS procedures, as developed by Shepard (1962) and Kruskal (1964), not only recover the coordinates of the stimuli in the putative perceptual space but also the shape of the generalization gradient. Applying ordinal MDS to many data sets Shepard (1987)

[1] In this context consider the following quote by Stevens (1965, p. 25): "The generalization gradient of the animal trainers is the psychometric function of the weight lifters".

found a pattern that is now called Shepard's *universal law of generalization*: The amount of generalization decreases (approximately) exponentially with the distance in perceptual space.

Shepard (1957, 1958) had used the exponential relationship already earlier to explain identification data in humans. In an identification task participants have to associate each stimulus in a set with a unique name. In a classical paired-associate experiment participants are shown a stimulus on each trial and have to respond with its name. If they do not respond with the right name they are corrected. This procedure is repeated until a performance criterion is reached. During learning participants confuse very similar stimuli more often than very dissimilar stimuli. Hence, these confusions are an indirect measure of stimulus similarity. Shepard seems to have thought that the confusions that arise during learning are reflections of generalization gradients in humans (witness the references to the animal learning literature in his early work). This is the reason why Shepard's law is not referred to as the "universal law of confusability" (Chater & Vitanyi, 2003) even though it might be that this is what it is. Experimentally, it is often hard to tell whether generalization gradients really reflect generalization in the literal meaning of the word or whether they are merely a reflection of confusion in memory or perceptual indiscriminability. Some animal learning theorists have argued that the concept of generalization is superfluous and discrimination is the only concept that is needed (Brown, 1965). Generalization might be only failure to discriminate. As a theoretical construct generalization refers to a covert process that leads a subject to respond to a new stimulus in the same way as to a previously learned stimulus despite the ability of the subject to tell the stimuli apart. This is the meaning that is intended by Shepard and it is also how generalization gradients are meant to be used in categorization research. For categorization studies the case where participants group two stimuli together into one category just because they fail to discriminate between them seems uninteresting. However, in some categorization studies it was not always clear how much of the generalization gradient is really due to generalization and how much is due to the fact that subjects cannot tell the stimuli apart (Nosofsky, 1986; Shepard, 1986; Ennis, Palen, & Mullen, 1988). In any case the confusions can be used as a measure of similarity.

2. Universal law of generalization

2.1. **l_p spaces.** It is tempting to assume that the representation of a stimulus is given as a point in a vector space. The dimension of the vector space ought to describe the perceptual dimensions along which stimuli can vary. With respect to the norm in this space it has become customary to use a weighted l_p norm for the length of a n-dimensional vector x:

$$\|x\|_p = \left(\sum_{i=1}^{n} \alpha_i \, |x_i|^p\right)^{\frac{1}{p}}. \tag{30}$$

with positive weights α_i. The weights are needed to allow for systematic variations of the norm over tasks or over individuals. This norm induces a metric on the space (which is also known as the Minkowski p-metric or power model):

$$d_p(x,y) = \|x-y\|_p = \left(\sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p\right)^{\frac{1}{p}}. \tag{31}$$

On a first glance, this metric seems to be an ad-hoc choice but it is implied by a set of desirable axioms that include the metric axioms, segmental additivity and conditions on the dimensions and the combination of dimensions (Tversky & Krantz, 1970).

The l_p formula is a norm and induces a metric only for $p \geq 1$. A metric has to fulfill three axioms: First, it is zero only for the distance from a point to itself ($d(x, y) = 0$ iff $x = y$). Second, it is symmetric ($d(x, y) = d(y, x)$) and third, it fulfills the triangle inequality, that is the direct connection from one point to another is at least as short as a detour over a third point ($d(x, z) \leq d(x, y) + d(y, z)$). For $p < 1$ equation (31) does not fulfill the triangle inequality—an issue that is crucial for psychology and that will be discussed below. Irrespective of whether d_p is a metric or not we will call it a distance (Blumenthal, 1953). We will only call it a metric if it also satisfies the triangle inequality (i.e. for $p \geq 1$).

It has been argued that the two psychologically most interesting cases of d_p are the city-block distance ($p = 1$) for separable stimulus dimensions and the Euclidean distance ($p = 2$) for integral stimulus dimensions (Shepard, 1964; Garner, 1974). Separable stimuli are stimuli which can be analyzed into their dimensional parts. It is possible to attend to just one of the dimensions without interference from the other dimensions. For integral stimuli this is not possible. A long list of studies used the l_p norm either directly or in the form of the Euclidean or city-block metric (e.g. Attneave, 1950; Shepard, 1964; Garner, 1974; Nosofsky, 1986; Kruschke, 1992).

2.2. Generalization gradients. If one is willing to commit oneself to a vectorial representation of stimuli and the distance d_p on this space there is still the question of how the distance in this space relates to the measured (dis)similarity of the stimuli. Intuitively, similarity should decrease and dissimilarity increase with distance.

As mentioned before, Shepard (1987) argued that the best measure for similarity are generalization gradients. He analyzed several data sets with his ordinal multidimensional scaling method and found that the non-linear relationship between the distance in the psychological l_p space and the measured similarity is generally monotonic and, in Shepard's terms, concave upward. In its stronger version Shepard's claim is that the relationship is exponentially decreasing. We refer to this exponential relationship as the universal law of generalization. Shepard's finding was in accordance with his much earlier suggestion of the exponential as a link between confusion probabilities and psychological distance (Shepard, 1957) and his diffusion model of similarity (Shepard, 1958). Furthermore, he tried to deduce the exponential from assumptions on optimal classification performance (Shepard, 1987). His assumptions were extremely restricted but a recent rational analysis of categorization gives a more compelling account (Tenenbaum & Griffiths, 2001). Shepard's work has been extremely influential and has led others to use the exponential as a similarity measure (e.g. Nosofsky, 1986; Kruschke, 1992; Love et al., 2004). A very general formulation for the similarity between two representations x and y is:

$$k(x, y) = \exp(-d_p(x, y)^\gamma), \tag{32}$$

an exponential of the distance d_p (31) raised to the power of γ. Shepard's original formulation did not have the exponent γ but many other authors make use of this extra parameter. Note again that the space that x and y are defined in is a psychological space. The coordinates of the stimuli in the psychological space is what multidimensional scaling is trying to reconstruct from the data.

Under certain circumstances the similarity measure as given by (32) leads to a so-called positive definite kernel and therefore opens up the rich theory of Hilbert spaces for the analysis of similarity. Discussing the preconditions and consequences of this observation is the purpose of this section. Finding that a symmetric, real

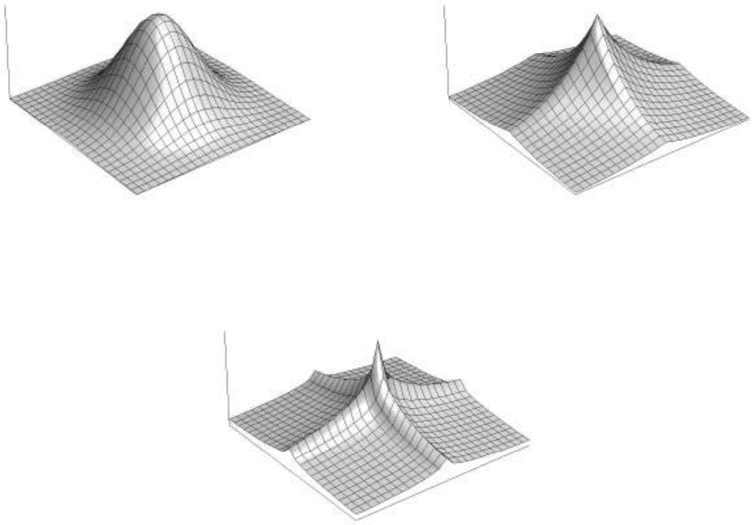

FIGURE 2. The similarity kernel for different values of p. From top left to right to bottom: For $p = 2$ a Gaussian is obtained, for $p = 1$ a Laplacian is obtained, and for $p = \frac{1}{2}$ the axes are very prominent.

function $k(x, y)$ is positive definite is extremely interesting from a mathematical perspective because it means that k can be represented as an inner product in a vector space. positive definite kernels are the infinite dimensional analogue of positive definite matrices in finite dimensional vector spaces. In linear algebra, each symmetric, real, and positive definite matrix can define an inner product on a finite dimensional vector space and the corresponding statement is true for positive definite kernels and infinite dimensional vector spaces. A complete and possibly infinite dimensional vector space with an inner product is called a Hilbert space. Using the techniques presented in Chapter 2 we will describe such a Hilbert space that is associated with Shepard's universal law of generalization. Let us briefly remind the reader of the results of Chapter 2 that are needed here.

2.3. The similarity kernel. A real and symmetric function $k(\cdot, \cdot)$ is called a positive definite kernel if for all choices of N points $x_1, ..., x_N$ from the domain of k, the following holds:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j k(x_i, x_j) \geq 0 \tag{33}$$

for all possible real coefficients w_i. If k is a psychological similarity function and the x_i are stimuli this means that for all possible stimuli the matrix of pairwise similarities is always positive semi-definite (compare Eq. 13).

The real and symmetric function $k(x, y)$ as given in (32) is such a positive definite kernel only for certain choices of γ and p. The exact conditions are complicated but known results are summarized by Koldobsky and Koenig (2001). For the current discussion it is enough to restrict attention to the simpler case $\gamma = p$

as the most widely used versions of the similarity kernel are of this form:

$$k(x, y) = \exp(-d_p(x, y)^p) = \exp(-\sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p). \tag{34}$$

Nosofsky (1990) calls this case with $\gamma = p$ "interdimensional multiplicative" because similarities are calculated for each dimension and then multiplied. With p chosen to be two the similarity measure has the form of a Gaussian kernel. With p chosen to be one the function is sometimes called Laplacian. These two cases correspond to the Euclidean and the city-block metric, respectively. Figure 2 shows the similarity kernel for $p = 2$, $p = 1$ and $p = \frac{1}{2}$. Contrary to the Gaussian kernel the other two kernels have clearly defined axes.

While the Minkowski p-metric (31) only defines a metric for $p \geq 1$ the similarity measure in (34) is a positive definite kernel for $0 < p \leq 2$. This is a classic result on positive definite kernels (Schoenberg, 1938). To the best of our knowledge there is no paper in psychology that claims a value for p bigger than two. Thus, the fact that (34) is not positive definite for $p > 2$ appears to be no serious restriction. However, there are several reports for a p smaller than one (Shepard, 1964; Tversky & Gati, 1982; Indow, 1994). In these cases trying to model similarity with a Minkowski metric is problematic because (31) is not a metric if $p < 1$—but the axioms of a metric space have been essential in the development of MDS. The similarity measure is, however, still a positive definite kernel for $0 < p < 1$ and therefore the kernel framework might provide us with an alternative model.

2.4. Reproducing kernel Hilbert space, revisited. We have observed that the above measure for similarity is a positive definite kernel. We will now introduce a vector space using this positive definite kernel as an inner product, following the same procedure as in Section 2.2. Let us assume, for simplicity, that the perceptual space is $\mathbb{R}^n$. The vector space $\mathcal{H}$ that will be constructed below is a space of real functions defined on the perceptual space, that is a function f in the vector space $\mathcal{H}$ is of the form $f : \mathbb{R}^n \to \mathbb{R}$.

The crucial idea is that we identify each stimulus with its similarity to all other stimuli (Edelman, 1998). For each stimulus x in the perceptual space there is a function from $\mathbb{R}^n$ to $\mathbb{R}$ that captures the similarity of x to all other stimuli in the perceptual space. This function is simply $k(\cdot, x)$ with a fixed x and interpreted as a function of its first argument. This function lies in the vector space $\mathcal{H}$ that we will construct. In this way, we identify each stimulus x in the perceptual space with a function, its similarity function, in $\mathcal{H}$. Instead of examining the perceptual space directly we will analyze the space of functions t̋hat is defined on the perceptual space and that contains all the similarity functions associated with each stimulus. It will turn out that this space has psychologically interesting properties. We will denote the function that maps each stimulus to its similarity function in $\mathcal{H}$ with $\Phi : \mathbb{R}^n \to \mathcal{H}$ and define it to be

$$\Phi(x) = k(\cdot, x). \tag{35}$$

The vector space $\mathcal{H}$ is now defined to be the set of functions that can be described as a finite linear combination of similarity functions. Each function f in $\mathcal{H}$, by definition, can be written as

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i) \tag{36}$$

for some N and a choice of points $x_1, ..., x_N$ with real coefficients $w_1, ..., w_N$. As noted in Section 2.6, it is no coincidence that this equation looks like a one-layer neural network, and because it is a linear combination of kernel functions these functions form a vector space.

There is a natural way to equip this vector space with an inner product. Let $g(x) = \sum_{i=1}^{M} v_i k(x, y_i)$ be another function from the vector space. An inner product between these functions can be defined as

$$\langle f, g \rangle = \sum_{i=1}^{N} \sum_{j=1}^{M} w_i v_j k(x_i, y_j). \tag{37}$$

This can be shown to be well-defined and it is symmetric due to the symmetry of k. It is linear in its arguments, too, due to the linearity of the sum. To show that it is an inner product we need to make sure that it is also positive definite, that is $\langle f, f \rangle \geq 0$ and equality only holds for $f = 0$. Positivity is guaranteed by the defining property of a positive definite kernel k (33). Definiteness follows automatically for positive definite kernels but is a bit more difficult to see (Schölkopf & Smola, 2002).

The vector space with the inner product that we introduced is almost a Hilbert space. Hilbert spaces can be thought of as a generalization of Euclidean spaces with a dimension that may be infinite. In order to be a Hilbert space the space needs to be complete, and the space we constructed can be completed by including certain limit points (Schölkopf & Smola, 2002). This completed space is then called a *reproducing kernel Hilbert space* (RKHS). It is called reproducing because of the following property,

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{N} w_i k(x, x_i) = f(x), \tag{38}$$

stating that the inner product between a function f and one of the basis functions $k(\cdot, x)$ evaluates the function at x. Hence, when we take the inner product of two similarity functions

$$\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y), \tag{39}$$

the function $k(\cdot, y)$ is evaluated at x.

Remember that in Eq. (35) we decided to map each stimulus x to the vector space by applying the function $\Phi(x) = k(\cdot, x)$, thereby identifying each stimulus with its similarity function. Because of the reproducing property (39) the inner product of two stimuli x and y in the RKHS is simply given by their similarity:

$$\langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y). \tag{40}$$

Calculating the similarity between two stimuli using a positive definite kernel k as given in (34) is therefore the same as taking the inner product in the Hilbert space that we constructed above. The similarity is given by an inner product as in the early work of Ekman (Gregson, 1975; Borg & Groenen, 1997). Shepard's suggestion to use the exponential as a link between distance and similarity has brought us back to the roots of MDS, the use of inner products. This is ironic because Shepard introduced ordinal scaling methods in order to go beyond the Euclidean case with its positive definite matrices. The introduction of the exponential as a link between distance and similarity has reintroduced the constrained of positive definite matrices.

2.5. The kernel metric. Like Euclidean space Hilbert space is a very rich structure with an inner product, a norm that is induced by the inner product and a metric that is induced by the norm. The norm of a function f in the Hilbert space is naturally defined as the square root of the inner product with itself $\|f\|^2 = \langle f, f \rangle$.

In particular, for the similarity kernel (34) all stimuli are mapped to the unit sphere in Hilbert space:

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = k(x,x) = \exp(0) = 1. \tag{41}$$

As all the points of the input space lie on the unit sphere in the Hilbert space the inner product is the cosine of the angle between the vectors in the Hilbert space.

Given a norm a natural definition of a metric is given by the norm of the difference vector. In the Hilbert space the distance between two functions f and g would then be given by the metric d'_p defined as $d'_p = \|f - g\|$. Hence, the inner product in Hilbert space naturally induces a metric on the space via the norm. Recall how we have arrived at this metric:

$$l_p \rightarrow d_p \rightarrow d_p^p \rightarrow k \rightarrow d'_p.$$

We started off with the l_p formula (30), this naturally defined d_p (31). This was taken to the power of p, giving rise to d_p^p. Taking the exponential gave rise to the similarity kernel k (34). As this is an inner product for $0 < p \leq 2$ we could naturally define a new metric d'_p. For the similarity kernel as introduced above we have the peculiar situation that the dissimilarity d_p that is used in the definition of the similarity kernel k is not induced by it. Instead, the inner product k induces another metric d'_p on the space. The new metric d'_p is different from d_p and is given by the distance of the stimuli in Hilbert space:

$$\begin{aligned} d'_p(x,y)^2 &= \|\Phi(x) - \Phi(y)\|^2 \\ &= \langle \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) \rangle \\ &= \langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(y) \rangle + \langle \Phi(y), \Phi(y) \rangle \\ &= 2 - 2k(x,y) \qquad (42) \end{aligned}$$

where we have used that the similarity kernel $k(x,x) = 1$ for all x. It is very interesting to note that this new metric is a monotone transform of d_p. Recall that in the Shepard-Kruskal multidimensional scaling procedure only ordinal properties of the data are used and therefore this new metric space is as good a representation for ordinal data as the l_p space. The similarities stay the same, only the metric of the space is changed. Shepard's analysis demands that the similarities and the metric are in a monotonically decreasing relationship—they still are for the metric d'_p. Even the coordinates of x and y stay the same. In the kernel framework, the inner product is an inner product in a Hilbert space. In the Hilbert space the similarity and the distance measure d'_p are in a natural relationship. The distance measure is derived from the similarity measure by using Euclidean intuitions about angles and distances.

This new metric has another noteworthy property: It is bounded from above. Points far apart in the space are separated by a distance which is at most $\sqrt{2}$ (the square of the distance as given in (42) approaches 2 if the similarity $k(x,y)$ approaches zero—making $\Phi(x)$ and $\Phi(y)$ orthogonal). Psychologically this is a highly desirable property because it means that a stimulus that is already very different from another stimulus cannot become much more different. In fact, very often the notions of perceptual difference and similarity are only meaningful locally and measurements of large perceptual distances are not available. An example would be color space where it is easy to obtain local measurements of similarity, for instance by looking at discrimination thresholds. However, global measures are not available. If directly asked for a judgment of the dissimilarity of colors far apart in color space, typically subjects find themselves unable to express a more precise answer than "totally different" (Indow, 1994). This important aspect of psychological similarity is captured naturally by the similarity kernel (34).

3. Triangle inequality

It may seem this new metric inherits all problems of the distance d_p on which it is based—but this is not so. Of course, it has to fulfill the metric axioms. In some cases the distance from a point to itself may not be zero and the symmetry of the empirical distances is not warranted. These violations may not always be explained by measurement noise and response biases. Considerable criticism of metric approaches has focused on symmetry and on constant self-similarity (Tversky, 1977). Symmetry, for example, can be violated if the comparison has a direction and one of the stimuli is more prototypical than the other, or receives more attention. Checking for violations of symmetry is relatively easy and even if an experimental measure is not completely symmetric, in practice it is often simply forced to be symmetric. Similarly, constant self-similarity is simply assumed in practice. Obvious violations of constant self-similarity can however be observed in confusion data or same-different experiments. In any case, whether one's data shows symmetry and constant self-similarity, at least approximately, can easily be checked. There are many situations where both assumptions will approximately hold. For cases where they do not hold, Dzhafarov and Colonius (2006) have recently described a principled procedure that can convert data from same-different experiments into a metric.

The most fundamental property of any metric model is perhaps the triangle inequality, which states that the direct path between two points is at most as long as any detour via a third point (Tversky & Gati, 1982). We will give a detailed explanation why the criticism of the triangle inequality as it applies to the d_p metric does not apply to the similarity kernel approach. Briefly, the reason is that the triangle inequality is usually paired with a second assumption, called segmental additivity (Beals et al., 1968), that does not hold for the kernel metric.

3.1. Concave iso-similarity contours. In an early paper, Shepard noted that concave (i.e., indented) iso-similarity contours lead to a violation of the triangle inequality for the l_p norm (Shepard, 1964). Figure 3 shows the unit "balls" for the l_p norm for different values of p assuming equal weights for both dimensions. All points on the curves have distance one to the center (in their respective norms). Figure 4 shows why the triangle inequality is not fulfilled for values of $p < 1$. For $p < 1$ the unit ball becomes concave. In this case, the distance from x to y is one, the distance from y to z is also one. Therefore, traveling from x to z via y takes two units but traveling directly, that is on a straight line, from x to z takes more than two units (the distance from x to w is greater than one and the distance from z to w is also greater than one). In his paper Shepard found violations of the triangle inequality but he could attribute them to pooling subjects with different response strategies. He further argued that the triangle inequality should be assumed and that violations can be explained by shifts in attention. Nevertheless, it seems psychologically plausible, or at least possible, that two stimuli that match on one dimension (like x and y in Figure 4) are more similar to each other than stimuli that do not match on any dimension (like x and w in Figure 4). The intuition is that stimuli that have matching dimensions have more in common with each other than stimuli that do not match on any dimension. In such a case Shepard would have to assume that dimensions that match receive greater attention than dimensions that do not match.

3.2. Triangle inequality or segmental additivity. Tversky (1977) was guided by the intuition of matching to develop his famous contrast model of similarity. His model does not require any of the metric axioms and it works without

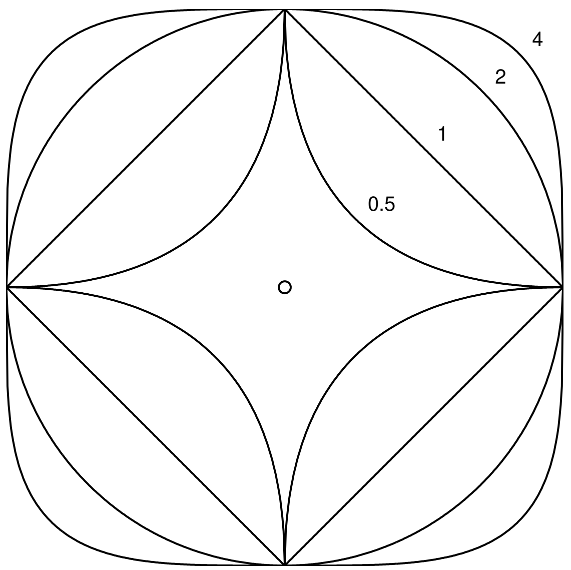

Figure 3. Unit balls of the l_p norm for different values of p.

positing shifts in attention. To reinforce his position Tversky and Gati (1982) conducted a study that tested the triangle inequality more directly than Shepard (1964) did. For this they sought measurements of stimulus dissimilarity for a wide variety of stimuli. Contrary to Shepard who used similarity choices they used mostly dissimilarity judgments. The dissimilarity measurements are usually only on an ordinal level, at most on an interval scale. With a finite set of points the triangle inequality can always be trivially satisfied by adding a big enough constant to the dissimilarity measures. However, together with a second assumption, called segmental additivity (Beals et al., 1968), testable predictions for metric models can be made (the so-called corner inequality which is explained below has to be fulfilled). Tversky and Gati could show that these predictions are not met by most of their data. This strongly suggests that either the metric axioms or segmental additivity does not hold for most of their stimuli.

By segmental additivity they meant the following: A segment is the shortest path between two points and a path is a sequence of points. As an example for a path think of a morph sequence between visual stimuli. The shortest morph sequence is called a segment. Let the points x and z be joined by the shortest path that connects them (e.g., a straight line if the space is Euclidean) and let w be one of the points on the way. It is tempting to make the following assumption: The psychological distance from x to z is exactly the sum of the distances from x to w and from w to z, $d(x, z) = d(x, w) + d(w, z)$. This is called segmental additivity. Implicitly we made this assumption above when we demonstrated that the distance d_p (31) does not fulfill the triangle inequality if $p < 1$. The assumption of segmental additivity is so intuitive that if it were to be given up the whole idea of representing similarity by geometric relations in a psychological space would seem to lose its intuitive appeal. Metrics with segmental additivity are a rather wide

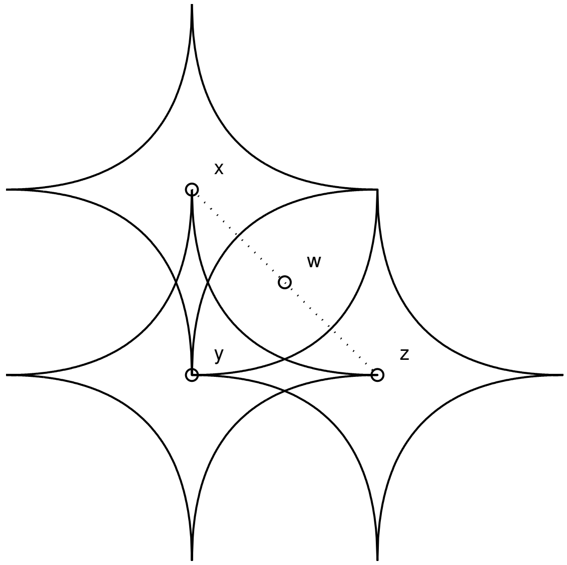

FIGURE 4. Violation of the triangle inequality for concave unit balls. The distances from x to y and from y to z are 1. Hence, traveling from x to z via y takes 2 units. Traveling from x to z directly via w takes more than 2 units as w is outside the unit balls of x and z.

class of metrics. They include all Minkowski metrics (d_p with $p \geq 1$) but also Riemannian curved geometries. Several recent scaling and embedding methods make use of segmental additivity (Dzhafarov & Colonius, 2006; Roweis & Saul, 2000; Tenenbaum, Silva, & Langford, 2000). Tversky and Gati found however that metrics with additive segments cannot account for their data.

But Tversky and Gati go one step further. They consider segmental additivity as essential for the enterprise of modeling similarity in perceptual spaces with geometric relations. As their data indicate that we cannot have both, the metric axioms and segmental additivity, they abandoned the metric approach and favored a non-metric approach. Already earlier, Tversky (1977) had suggested a non-metric model, implementing a matching mechanism, that can account for their data: The contrast model. Nevertheless, they also point out that the metric approach could be saved by sacrificing segmental additivity, only to add that this solution would not be "compatible with lay geometric intuitions" (Tversky & Gati, 1982, p. 151). Our new metric d'_p on the psychological space does not have the property of segmental additivity but is theoretically well motivated. We can also provide some geometric intuitions why segmental additivity is not as crucial as it may seem. Hence, we believe that the kernel metric gives an interesting solution that is different from the contrast model and more in the spirit of MDS. But before we explain how the kernel metric can account for the data of Tversky and Gati we examine their reasoning in detail.

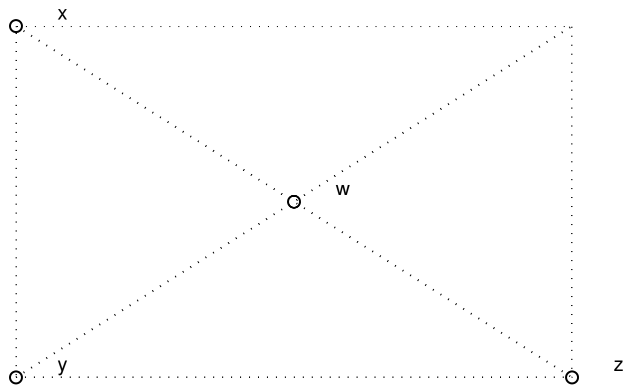

FIGURE 5. Illustration of the corner inequality. The center path is from x to z via w. The corner path is from x to z via y. If the center path exceeds the corner path the corner inequality is violated.

3.3. Corner inequality or coincidence hypothesis. Consider the perceptual space depicted in Figure 5. It is a two-dimensional space with stimuli z and y agreeing on the first dimension and stimuli y and x agreeing on the second dimension. In an experiment these dimensions could be the size and angle of a circle with a spoke, or the two sides of a rectangle, or hue and saturation of a color chip. In many experiments the participants are aware of the dimensions that the experimenter manipulates, especially when the dimensions are of the separable kind (Shepard, 1964; Garner, 1974) and the stimuli are constructed by a factorial design. If subjects actively look for features that two objects have in common in order to make similarity judgments then we would expect that z and y are similar because they have the same value one of the dimensions, and for the same reason y and x should be similar, too. As x and z have nothing in common they should be very dissimilar. However, as x is close to y and y is close to z, x and z should not be too far apart because of the triangle inequality. Intuitively speaking, if we assume the triangle inequality then there is not much room for x and z to be very different from each other. Hence, the triangle inequality, so it seems, may be inconsistent with the idea that subjects match features in their assessment of similarity.

The above hand-waving reasoning is hard to put to test without putting concrete numbers on the distances between the stimuli. Unfortunately, measurements of perceptual similarity are usually only on an ordinal scale level (at most interval scale) and therefore the numbers that one would like to have are not easily available. Tversky and Gati realized that pairing the triangle inequality with segmental additivity can give ordinal predictions. The triangle inequality states that $d(x, z) \leq d(x, y) + d(y, z)$, assuming that segmental additivity also holds, and that we have a fourth stimulus w that lies on the straight line that connects x and z, we get the following necessary condition:

$$d(x, w) + d(w, z) \leq d(x, y) + d(y, z). \tag{43}$$

Tversky and Gati call this condition the *corner inequality*. The right-hand side is called the corner path and the left-hand side the center path, with respect to the rectangle that is defined by the levels that the stimuli take on the dimensions (see Figure 5). All Minkowski metrics (d_p with $p \geq 1$, see Eq. 31) satisfy the corner inequality. Note that if $p = 1$, for the city-block metric, equality holds. As we have seen in Figure 4 for $p < 1$ the corner inequality is violated. The corner inequality is violated if the center path exceeds the corner path. The center path clearly exceeds the corner path in an ordinal sense if

$$\begin{gathered} d(x,w) > d(x,y) \quad \text{and} \quad d(w,z) > d(y,z) \\ \text{or} \\ d(x,w) > d(y,z) \quad \text{and} \quad d(w,z) > d(x,y). \end{gathered} \tag{44}$$

If subjects are matching the dimensions of the stimuli in their assessment of similarity it could very well be that, for example, $d(x, w) > d(x, y)$ because x and y have the same value on one dimension and x and w are difficult to compare because they have to be compared across two dimensions. Subjects' sensitivity to matching dimensions is called the *coincidence hypothesis* by Tversky and Gati. The coincidence hypothesis predicts that the corner inequality is violated. As an ordinal test for this violation we can use condition (44). Note that this ordinal test cannot detect all but the most grave violations of the corner inequality. Nevertheless, Tversky and Gati found serious violations for many similarity judgments with many different stimulus sets.

In the derivation of the ordinal test it was assumed that a rectangle is constructed with stimuli agreeing on some dimensions and with some stimuli being placed on the way between other stimuli. With some stimuli, for example colors, one cannot be sure that one really has chosen the right dimensions and it will be hard to construct stimulus sets with triples of stimuli located on a segment in perceptual space. It might not even be clear that the perceptual space is really two-dimensional. However, Tversky and Gati also analyzed all their data using ordinal multidimensional scaling procedures that estimate the coordinates of the stimuli in perceptual space. In these procedures one usually uses the l_p norm that poses more constraints on the distances than just the metric axioms and segmental additivity. With these methods one can obtain an estimate of the p parameter in the l_p norm. They found, if p is not restricted to be greater than one (i.e., the triangle inequality is given up), the best fit for most of their data sets can be achieved by a $p < 1$. Increasing the number of dimensions by one or two did not improve the fit as much as allowing a $p < 1$. However, model selection for the dimensions of a MDS solution is notoriously difficult and it is not clear how the flexibility of the model changes when changing p at the same time. In any case, Tversky and Gati took their analysis as an indication that similarity is boosted if stimuli agree on a dimension.

Tversky and Gati conclude that models that assume the triangle inequality and segmental additivity cannot account for many human similarity judgments. The major exception was color space for which they could not reject the hypothesis that it fulfills both, the metric axioms as well as segmental additivity. The other stimuli were of the separable kind and interestingly the degree with which the assumptions were violated seemed to vary with how transparent the dimensional structure of the stimuli was to the participants. The more transparent the dimensional structure was to the subjects the more they were inclined to base their assessments of similarity on matching dimensions.

3.4. Non-metric or metric without segmental additivity. As mentioned before, these results have led Tversky and Gati, and many researchers after them, to prefer non-metric models of similarity. The contrast model (Tversky, 1977) is the most prominent non-metric model and is explicitly built on the intuition of matching features. Nevertheless, Tversky and Gati do acknowledge that the triangle inequality on its own is not constraining the class of similarity models very much. There are metric models that can be reconciled with the coincidence hypothesis. Those models do not have the property of additive segments that is characteristic of basically all the intuitive geometries that have been used to model and, perhaps more telling, to visualize similarity. There are many metrics that do not have additive segments. For example, the so-called "metric for bounded response scales" is even discussed in a standard reference on MDS and was introduced to deal with exactly those issues raised by Tversky and co-workers (Borg & Groenen, 1997). Another such metric that is of particular interest here was suggested by Tversky and Gati themselves and is given by the l_p formula (31) taken to the power of p:

$$d_p(x,y)^p = \|x-y\|_p^p = \sum_{i=1}^{n} |x_i - y_i|^p \,. \tag{45}$$

This definition results in a metric for $0 < p \leq 1$, as also noted by Carroll and Wish (1974). The triangle inequality is satisfied under these circumstances because x^p is a concave function for $0 < p \leq 1$ and positive real x:

$$\begin{aligned} d_p(x,y)^p &= \sum_{i=1}^{n} |x_i - y_i|^p \\ &\leq \sum_{i=1}^{n} |x_i - z_i|^p + |z_i - y_i|^p \\ &= d_p(x,z)^p + d_p(z,y)^p. \end{aligned} \tag{46}$$

Along each dimension distances do not add up unless $p = 1$. Let us assume $p < 1$. Take three stimuli a, b and c that are identical on all but the i^{th} dimension and with b located between a and c, that is $a_i < b_i < c_i$. For all such stimuli it holds that $d(a,c) < d(a,b) + d(b,c)$. We call such a metric an *intradimensionally subadditive* metric. Large differences are down-weighted relative to small differences along each dimension.

Let us re-examine the example in Figure 4. If we choose the p^{th} power of d_p with $p < 1$ as a metric then the triangle inequality holds, as just shown. The direct connection from x to z is actually equal in length to the detour via y. The distance from x to z is $d(x,z) = 1^p + 1^p = 2$ and the distance from x to y is $d(x,y) = 1^p = 1$ and so is the distance from y to z, $d(y,z) = 1$. While w appears to be on the way from x to z it turns out that the sum of $d(x,w) = \left(\frac{1}{2}\right)^p + \left(\frac{1}{2}\right)^p = 2 \cdot \left(\frac{1}{2}\right)^p$ and $d(w,z) = 2 \cdot \left(\frac{1}{2}\right)^p$ is greater than $d(x,z) = 2$. The center path exceeds the corner path. We have an example of a metric that is consistent with the coincidence hypothesis

The unit ball of this metric is the same as for a l_p norm with a $p < 1$ as depicted in Figure 3. It is in fact possible to have an indented iso-similarity curve if the metric does not satisfy segmental additivity (Carroll & Wish, 1974). The argument that led to a break-down of the triangle inequality for the l_p norm for $p < 1$, as illustrated in Figure 4, implicitly assumed segmental additivity. In particular, it assumed that w lies on a segment between x and z. The metric in (45) does not satisfy segmental additivity. Recall that for two points x and z that are joined by an additive segment it holds that for all points w that lie on the segment $d(x,z) = d(x,w) + d(w,z)$. Along a segment the triangle inequality becomes an equality. If there are pairs of points in a metric space that cannot be joined by an additive segment then we say that the metric space does not fulfill segmental additivity. Hence, a necessary condition for segmental additivity is that there is at least one point w that lies between any two points x and z in the sense that $d(x,z) = d(x,w) + d(w,z)$

(This condition is called metric convexity by Blumenthal, 1953). For the metric in (45) with $p < 1$ there are points that do not have this property. Not even pairs of points that lie on one stimulus axis have this property because the metric is intradimensionally subadditive.

For a metric with indented unit balls it is exceedingly difficult to interpret a configuration of stimuli as depicted in Figure 5 as a map or any other intuitive geometry, "despite the natural tendency to do so" (Tversky & Gati, 1982, p. 151). The problem is that there is no natural notion of "on the way" between two points. There is no obvious way one can construct a path, a sequence of points between two points, in the space such that the sum of the partial distances equals the full distance. The full distance can be shorter than the sum of the partial distances.

3.5. Kernel metric and segmental additivity. Tversky and Gati (1982, p. 151) conclude that the choice between non-metric models and metrics without segmental additivity is "more likely to be made on the basis of theoretical rather than empirical considerations". According to Tversky and Gati, the appeal of metric models is considerably reduced if there are stimuli for which there is no easily interpretable metric with additive segments to account for their similarity. There is a lot to be said in favor of non-metric models that explicitly try to capture the psychological processes underlying similarity judgments without any concern for the metric axioms. However, the metric axioms are among the most fundamental notions in mathematics. Giving up the metric axioms means that one cannot make use of the sophisticated apparatus that is built upon them. The kernel metric d'_p (42) that we introduced above is, on the other hand, theoretically well-motivated by Shepard's law and does not have additive segments. This can be deduced from the fact that the kernel metric is bounded by $\sqrt{2}$. Even though a path may become longer and longer the maximum distance between any two points is $\sqrt{2}$. Thus, the kernel metric may provide a theoretically well-founded metric alternative to the non-metric models that Tversky and Gati favor.

Like the p^{th} power of d_p (45) the kernel metric is derived from the l_p formula (30). Note that the exponent in the definition of the kernel (34) is the p^{th} power of d_p. The iso-similarity curves of the kernel metric are identical to the iso-similarity curves of the l_p norm (Figure 3) because the same value for the l_p formula implies the same distance in the kernel metric. Contrary to the l_p norm, the kernel metric can also have indented iso-similarity curves. It can show matching behavior in accordance with the coincidence hypothesis with values of $p < 1$. In fact, it is a metric for $0 < p \leq 2$. Furthermore, it is a metric that is derived from an inner product, and therefore we may use some of our Euclidean intuitions in its analysis.

The metric d'_p leads to segmental additivity in a higher dimensional space. The points in the psychological space are mapped to the unit sphere in the infinite dimensional Hilbert space (41). In Figure 6 we have depicted a three dimensional subspace that contains the points x, w and z from Figure 5 mapped into the Hilbert space using the mapping $\Phi(x) = k(\cdot, x)$. As any finite dimensional subspace of a Hilbert space is simply a Euclidean space, we can interpret the distances in Figure 6 in the normal intuitive sense. The metric d'_p is the metric of the Hilbert space in which the original psychological space is embedded and therefore the distance between two points is given by the chord that joins them. The locations of the three points depicted in the figure are calculated from their inner products that are given by the similarity kernel. The similarity kernel gives the inner product between two points and therefore determines the angle between them. As the similarity kernel can only give values between zero and one, stimuli that are less similar to each other are more orthogonal to each other. Hence, the maximum distance between two stimuli is $\sqrt{2}$ because all stimuli lie on the unit sphere.

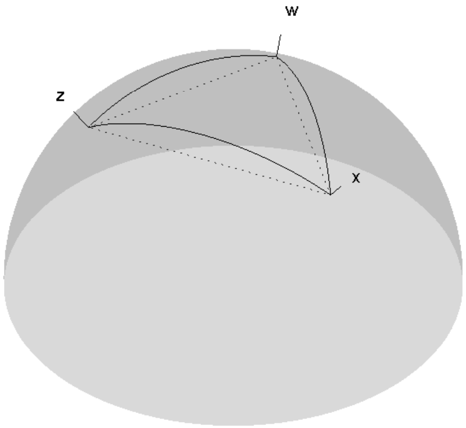

FIGURE 6. The similarity kernel maps the stimuli x, w and z from Figure 5 to the unit sphere in a Hilbert space. The distance from x to z is smaller than the sum of the distances from x to w and from w to z. The metric does not have additive segments in the original space because the distances are computed by the shortest connection in Hilbert space (the dotted lines).

The distance between x and z is given by the straight, dotted line that connects the two points in the Hilbert space. Note that in the embedding space w is not on the way from x to z. In fact, none of the points that lie on the chord that joins x and z is a potential stimulus because we know that all stimuli from the original space lie on the unit sphere when mapped to the Hilbert space (41). And even worse, w that seems to lie on the way between x and z in Figure 5, lies in a different dimension in Figure 6. As the kernel matrix of our similarity measure always has full rank (if no two points are identical) the vectors that represent each stimulus are always independent. The shortest path in the Hilbert space goes through points that are no stimuli. Hence, segmental additivity does not hold in the original psychological space which is the one depicted in Figure 5 and also the one that Tversky and Gati examined. It is as if you can tunnel through the sphere in order to get to another stimulus. You do not have to visit any other stimuli on the way. Any visit to another stimulus implies a detour (even if the stimuli only differ in one dimension). Hence, no two distinct points x and z in the original space can be joined by an additive segment in the original space because for all other points w in the original space it holds that $d(x,z) < d(x,w) + d(w,z)$.

A value of p smaller than one means that either the metric axioms do not hold or one has to look for a monotonous transform of the l_p formula such that the metric axioms hold but segmental additivity is violated. The metric d'_p provides one possible solution. But one that is well-motivated by Shepard's law. The iso-similarity curves can be indented because they arise from an inner product and not from a metric with additive segments. The similarity kernel defines an inner product for $0 < p \leq 2$ and for $p < 1$ in particular. Thus, the kernel metric is immune to the strong criticism put forward by Tversky and Gati. The argument that showed for the l_p norm that the triangle inequality is violated for $p < 1$ (Figure 4) does not work because in Hilbert space w does not lie on the way between x and z.

In order to construct a metric that is consistent with the coincidence hypothesis we have embedded the psychological space into an infinite dimensional Hilbert space. The embedding into an infinite dimensional space seems to be a drastic step. A finite dimensional Euclidean space with a higher dimension could also solve the problem. In fact, this is what is usually done in multidimensional scaling: The dimensions of the embedding space are increased until a satisfactory fit to the data is achieved. In the paper by Tversky and Gati this possibility is rejected because increasing the dimensions of the embedding space beyond the dimensions of the stimulus space did not increase the goodness of the fit as much as allowing for a p smaller than one in the l_p formula (31).

In normal MDS, if the stimulus space has a smaller dimension than the perceptual space then the stimuli that are presented to a subject will fall onto a (nonlinear) submanifold in the embedding space. The manifold has the dimension of the stimulus space. A simple example for this is the color circle (Shepard, 1980). If an experimenter chooses the one-dimensional set of stimuli that is comprised of only monochromatic lights then these stimuli will have to be embedded on a circle in two dimensions. The distance is given by the direct connection in the embedding space and not by the shortest path on the stimulus manifold (that only consists of monochromatic lights). In such a case it is no surprise that the metric does not have additive segments. This chordal metric is in fact the standard example for a metric without additive segments (Beals et al., 1968). The kernel metric we have presented here is very similar to this example, it represents each stimulus on the unit sphere in a Hilbert space.

3.6. Similarity choice and categorization. As mentioned before, even if the stimuli are only of dimension two interpreting the kernel metric as a 2d map is problematic—especially with $p < 1$. However, it is possible to get a feeling for how such a metric behaves by looking at a simple example. Figure 7 shows two stimuli in a perceptual space. Imagine a third stimulus that is varied in an experiment. The purpose of the experiment is to assess the similarity of the third stimulus to the other two stimuli and to check for subadditivity. This might be done in several ways. We can use generalization gradients. We can count how often the third stimulus is confused with the other two in a suitable task. We can ask participants to use a rating scale. We could also use the method of triads and ask participants directly which of the two exemplars depicted in Figure 7 is more similar to the third stimulus. Probably subjects' similarity choices will not be deterministic and therefore we can only record the relative frequency of a choice. It is not at all clear how, for example, similarity ratings relate to generalization gradients unless one tries to explicitly model the psychological processes that give rise to the subjects' responses. In ordinal MDS one only hopes that in all tasks subjects use the same psychological space to generate their responses and that all these measures of similarity are merely monotone transforms of the underlying metric. Which for choice probabilities is clearly problematic (Krantz, 1967).

Nevertheless, let us consider the scenario where a participant has to choose to which of the two exemplars depicted in Figure 7 the third stimulus is more similar. The gray-level in Figure 7 represents the choice probabilities of the MDS-choice model for two stimuli and a d_p with $p = \frac{1}{2}$. In the MDS-choice model the choice

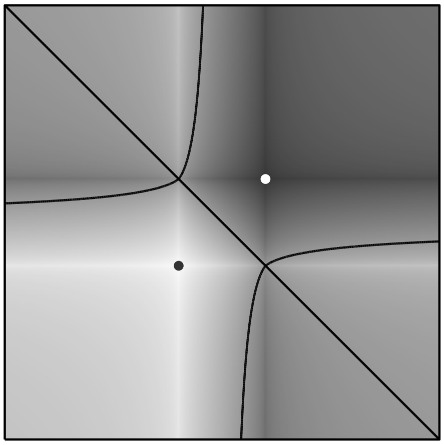

Figure 7. The MDS-choice model with a similarity kernel that has an exponent smaller than one. The two axes are two perceptual dimensions, the two points are two exemplars and the gray-level denotes the probability that a new stimulus will be assigned to one of the two exemplars. The solid black lines is where the probability is one half. With an exponent smaller than one the MDS-choice model can be made to show matching behavior. Stimuli that match an exemplar on the perceptual axes show a high similarity.

probabilities are calculated by plugging the similarity kernel into Luce's choice rule (Nosofsky, 1986; Shepard, 1957). White means a high probability that the subject assigns this point to one stimulus, black means a high probability for the other stimulus. The two axes represent two perceptual dimensions. The similarity kernel clearly follows these axes. The solid black lines mark the equivocality contour where a stimulus is equally likely to be assigned to one of the two exemplars. With an exponent smaller than one the similarity kernel shows matching behavior. If a test stimulus matches one of the exemplars on one of the dimensions it is more likely to be assigned to this exemplar. Even if it seems that the stimulus is closer the other exemplar by just looking at the picture. Remember, that the similarity kernel can be interpreted as an inner product in a Hilbert space. The two-dimensional stimulus plane depicted in Figure 7 is mapped onto the unit-sphere in the Hilbert space with the distance between the points given by the chord-metric (see Figure 6) and is therefore without additive segments. What seems to be a long path to travel in the two-dimensional plane is not the direct connection taken by the assessment of similarity.

The MDS-choice model is the basis for the Generalized Context Model (GCM), a prominent exemplar model of categorization that was introduced by Nosofsky (1986). One should expect that matching effects also play a role in categorization (Verguts, Ameel, & Storms, 2004). and therefore categorization models that are based on metric models should have similar problems than the underlying models for similarity. Nosofsky was well aware of Tversky's criticism of the use of the l_p norm but he used it nevertheless. He fitted the MDS-choice model to confusion data in an identification task. As he used stimuli with separable dimensions it was to be expected that for his data the exponent p should equal one or even be smaller than one. However, this is not what he found. He found that a Gaussian similarity kernel with a Euclidean metric could fit his identification data very well and smaller exponents led to dramatically worse fits. The reason for this may lie in the fact that Nosofsky's stimuli were difficult to discriminate perceptually (Nosofsky, 1986; Shepard, 1986; Ennis et al., 1988). His generalization gradients are partly explained by perceptual noise. Nosofsky had to use very confusable stimuli for the identification experiment. If the participants had been able to learn the identification task perfectly the MDS-choice model could not have been fitted to the data. The purpose of the study was to link confusion probabilities in an identification task with choice probabilities in a categorization task—and confusion probabilities can only be obtained if there is confusion to start with. Similarity was only invoked indirectly in both tasks. Tversky, in contrast used dissimilarity judgments of highly discriminable stimuli.

Nosofsky (1986) speculated that despite Tversky's results the psychological space could still be Euclidean: Participants in a more cognitive task, like dissimilarity judgment, actively look for matching dimensions and hence might change their attention weights from trial to trial. Be this as it may, our discussion of Tversky's criticism in the context of the similarity kernel clearly shows that categorization models that use the similarity kernel do not need to be overly concerned by Tversky's criticism. It may turn out that an exponent smaller than one is needed for the l_p norm for some categorization data using easy to discriminate, cognitive rather than perceptual stimuli. In any case, subadditivity can easily be incorporated in categorization models by choosing a p smaller than one. If p is simply seen as a free model parameter then this is perhaps not surprising. However, it is reassuring to know that this can be done without having to give up the metric axioms. Subadditive behavior in categorization can be dealt with without explicitly incorporating feature matching mechanisms as for example in (as for example in Verguts et al., 2004).

4. Conclusions

We have demonstrated how the serious concerns about the triangle inequality that accompany all metric models of similarity can be addressed in a principled manner. It was important to realize that the experimental tests of the triangle inequality were always in conjunction with a second assumption: segmental additivity. Hence, the data that seemed to contradict the triangle inequality can be explained by a metric without segmental additivity. Shepard's law of generalization can be used to induce an inner product in a Hilbert space which in turn induces a metric with several psychologically appealing properties. It is does not have additive segments and avoids the serious criticism of the triangle inequality. It is also bounded from above and therefore captures the intuition that similarity makes the most sense locally with only small changes in the stimulus. Stimuli far apart in perceptual space are merely completely different and more precise judgments of similarity are difficult. Table 1 summarizes the properties of the kernel metric d'_p and compares it to the other distance functions described in this paper.

Remember that the embedding into Hilbert space—drastic as it may seem—is just a new view on an old model of similarity and the embedding into this Hilbert space follows naturally from the definition of the similarity measure. All we have done is to reinterpret Shepard's similarity measure as an inner product in an implicitly given space. The Hilbert space does not show up explicitly in any of the equations, neither for the inner product nor for the metric. Everything operates in the dimensions of the original psychological space for which the l_p norm is defined. A useful way to think about these issues is that we have replaced the

	$d_p(x, y)$ (31)	$d_p(x, y)^p$ (45)	$d'_p(x, y)$ (42)
metric	$p \geq 1$	$p \leq 1$	$p \leq 2$
inner product	$p = 2$		$p \leq 2$
segmental add.	$p \geq 1$	$p = 1$	—
intra. subadd.	—	$p < 1$	$p \leq 2$
center > corner	$p < 1$	$p < 1$	$p < 1$
bounded	no	no	yes

Table 1. A comparison of the properties of different distances discussed in this paper. We always implicitly assume $p > 0$ for all entries in the table and d'_p is the distance that is induced by the kernel $k(x, y) = \exp(d_p(x, y)^p)$. The first row shows the conditions under which the distances are also metrics. The second row gives the conditions that allow a distance to be expressed as a metric that is derived from an inner product. For d_p^p we haven't said anything about whether it can be induced by an inner product. The third row notes when the distance is a metric that also fulfills segmental additivity. For the metric d'_p there are no additive segments. The fourth row is concerned with metrics that are intradimensionally subadditive. The fifth row shows under what conditions the center path exceeds the corner path. This is the case if the iso-dissimilarity contours are indented (Figure 4 and Figure 5). In the last row we note whether the metric is bounded.

task of choosing a metric for the psychological space (d_p) and a transform between similarity and distance (e.g. the exponential) with the task of choosing a similarity kernel on the psychological space (the exponential of d_p^p). The similarity kernel induces a natural metric that is bounded and that circumvents the problems of the l_p norm for values of $p < 1$. The metric however has lost its pivotal role in the theory. Similarity and inner product become the core concepts from which the concepts of dissimilarity and metric are derived.

As the similarity kernel (34) is a positive definite kernel a data matrix with similarity measures (bias corrected confusion probabilities in Shepard's case) would have to be positive definite up to noise, too. Thus, ironically, Shepard's proposal of the exponential as a transformation between measured similarity and distance takes us back to the roots of multidimensional scaling: To the use of inner products in a Euclidean space just like Ekman and Torgerson have pioneered it.

CHAPTER 4

Categorization

Intuitive definitions of categorization tend to invoke similarity. Objects that are similar are grouped together in categories. Within a category similarity is very high whereas between categories similarity is low. Similarity is at the heart of many categorization models. Prototype theories postulate that categorization depends on the similarity of stimuli to an abstracted idea (Posner & Keele, 1968; Reed, 1972) and exemplar theories calculate the similarity to memory representations of previously encountered stimuli (Medin & Schaffer, 1978; Nosofsky, 1986; Kruschke, 1992). A potential problem for these models is that they put the burden of explanation onto the intuitive concept of similarity. Despite serious problems in defining similarity (Medin et al., 1993) models of categorization continue to rely on similarity.

The appeal of invoking similarity in categorization models stems from the need to generalize. Given a stimulus that was never encountered before, how can it be categorized correctly based on limited experience with previous stimuli? An easy answer seems to be that a new stimulus is simply categorized in the same way as similar stimuli before. Correct generalization to new stimuli thus depends crucially on choosing the right similarity measure. Shepard (1987) famously has turned this reasoning around and used generalization to measure similarity. He also tried to deduce a similarity measure such that the generalization performance is likely to be good (Shepard, 1987; Tenenbaum & Griffiths, 2001; Chater & Vitanyi, 2003).

In Shepard's work the idea of a perceptual space has played a major role. The similarity measure he suggested, Shepard's *universal law of generalization*, operates on a mental representation assumed to be a metric space. Shepard's work on generalization and similarity (Shepard, 1957, 1987) cannot be separated from his work on categorization (Shepard et al., 1961; Shepard & Chang, 1963) and multidimensional scaling (Shepard, 1962). Since the work of Shepard it has become common for perceptual categorization models to assume a perceptual space and use Shepard's law as a similarity measure on this space. Exemplar models in particular strongly rely on Shepard's work (Nosofsky, 1986; Kruschke, 1992). These models are very similar to a class of popular tools in machine learning and statistics: kernel methods. This observation has first been made by Ashby and Alfonso-Reese (1995). Here, we draw parallels between recent progress in kernel methods and exemplar theories of categorization.

1. Kernel methods

In the past psychological theories of learning and categorization were a major influence for engineers to build machines that are capable of intelligent behavior. This is signified by the vast engineering literature that has been published on artificial neural networks and reinforcement learning. More recently, in machine learning there has been an increased interest in kernel methods. Even though these methods can be implemented in simple neural networks they are usually not psychologically or biologically motivated but instead are seen to be grounded in statistics and functional analysis. However, as will be shown here, researchers in kernel methods are often guided by the same intuitions about similarity and generalization that also guide psychologists in their theories on categorization. Hence, we will argue that theoretical progress in machine learning can also lead to new insights in psychology.

Methods that are based on kernel ideas are often found to have cutting-edge performance in real-world applications. For example, benchmark data sets for digit recognition are often used to compare the performance of different learning algorithms. The task for a learning algorithm in this setting is to correctly classify handwritten digits it has never seen before based on experience with a limited number of examples. For a long time a hand-tuned neural network held the world record on digit recognition benchmarks until a much simpler kernel method, called support vector machine (SVM), was shown to achieve better performance with much less effort on the side of the engineer. Today, support vector machines are found in applications ranging from bioinformatics to machine vision (Christianini & Schölkopf, 2002).

The successful application of kernel methods to real-world classification problems has led to an explosion of theoretical work in the field of machine learning. While there was already a considerable amount of theory on artificial neural networks progress has been hindered by the complexity of the neural networks that were used in practice. Kernel methods are built on linear methods and are therefore a lot easier to analyze than the non-linear neural networks (Schölkopf & Smola, 2002).

In this chapter we demonstrate that the Generalized Context Model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992), two well-known exemplar models, are very closely related to a machine learning method called kernel logistic regression (Hastie et al., 2001). The link between the psychological models and the machine learning method is their use of a radial-basis-function (RBF) neural network (Poggio & Girosi, 1989; Poggio, 1990). To this end we first recaptiulate the ideas behind kernel methods and RBF-networks that were introduced in Chapter 2. This is followed by a section on exemplar models where we discuss their history, explain their connection to kernel methods, and point to important differences in their response rules that affect how Shepard's law enters the models. The kernel-view makes the differences between the models more transparent and it also allows an easy comparison with methods from machine learning and statistics.

Like in psychology there is a tight relationship between similarity and generalization in machine learning. However, insights from machine learning show that while it is very important to choose the right similarity measure, this is not always enough to be guaranteed to have a good generalization performance. More specifically, if used naively kernel methods will be prone to overfitting. The section on Generalization discusses the consequences of these insights for exemplar theories of categorization. Exemplar theories have thus far exclusively relied on similarity for explaining generalization. In fact, a major criticism of exemplar theories has always been that they do not show any form of abstraction and therefore are often thought not to be able to generalize at all. We show how related kernel methods in machine learning assure good generalization performance by a mechanism called regularization. We argue that similar mechanisms need to be implemented in psychological models if they ought to exhibit a good generalization performance. To demonstrate how this could be done in principle, we analyze ALCOVE's learning algorithm from a regularization perspective. We find that ALCOVE's behavior can be justified on theoretical grounds, thus providing, for the first time, an analysis of the generalization abilities of exemplar theories.

1.1. The similarity kernel, revisited. In order to model the similarity, that is the generalization gradient, between two stimuli x and y we first interpret x and y as coordinates in a n-dimensional perceptual space. The perceptual distance in

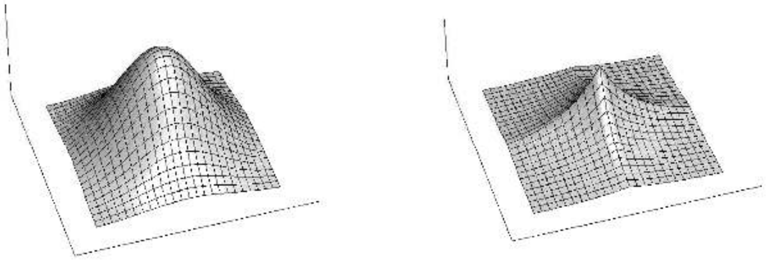

FIGURE 1. The similarity kernel for different values of p. For $p = 2$ a Gaussian is obtained (left panel), for $p = 1$ a Laplacian is obtained (right panel).

this space is usually modeled as (see Eq. 31, above)

$$d_p(x, y) = \left(\sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p \right)^{\frac{1}{p}} . \tag{47}$$

Most often the distance d_p takes the specific form of either the city-block or the Euclidean distance, with p chosen to be one or two, respectively. The α_i are positive weights that are needed to model the relative importance of the stimulus dimensions that possibly change with the experimental context. Using Shepard's law the generalization gradient between x and y is modeled as

$$k(x, y) = \exp\left(-d_p(x, y)^\gamma\right) . \tag{48}$$

We refer to the function k that models the generalization gradient as the *similarity kernel*. It is an exponential function of the distance d_p between the two stimuli in perceptual space. Deviating from Shepard's original formulation, the distance is nowadays often modified by taking it to the power of γ to give the model more flexibility. In models of categorization one often finds that γ is chosen to be equal to p (Nosofsky, 1990). As in Eq. (34) in Chapter 3 the similarity kernel is then given as

$$k(x, y) = \exp(-d_p(x, y)^p) = \exp\left(- \sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p \right) . \tag{49}$$

With p chosen to be two the similarity kernel is called a Gaussian kernel. The Gaussian kernel is extremely popular in machine learning where it is used to model the similarity of all sorts of things. The left panel of Figure 1 shows again a Gaussian kernel in two dimensions. Imagine a two-dimensional perceptual space, for example perceived size and angle of circles with a spoke. Stimulus y is fixed at the center of the Gaussian and the height of the plot depicts the similarity of all other stimuli in the plane to stimulus y. The generalization gradient is rotation invariant, it falls off in the same way in all directions of space. With p chosen to be one the similarity kernel is sometimes called a Laplacian kernel (in analogy to the Laplacian distribution). This case is depicted in the right panel of Figure 1. The Laplacian kernel is not rotation invariant. The generalization gradients fall off differently in different directions of space. In particular, along the stimulus axes similarity fades away more slowly.

The similarity kernel as defined in equation (49) has several psychologically and mathematically interesting properties that we explored in detail Chapters 2 and 3. Let us briefly remind the reader of the most important results: For values of p that lie between zero and two the function k is a so-called positive definite kernel (Schoenberg, 1938). This insight opens up a large box of mathematical tools from functional analysis that can be used to gain a better understanding of psychological models of similarity. In fact, the same tools have greatly deepened the understanding of machine learning methods that also use positive definite kernels.

Instead of representing each stimulus y by its coordinates in perceptual space we can represent each stimulus by its similarity to all other stimuli. Formally this can be done by representing each stimulus y by a function on the perceptual space. This function is the similarity of y to other points in the perceptual space: $k(\cdot, y)$ with a fixed y and interpreted as a function of its first argument. In this representation, representation is literally representation of similarities (Edelman, 1998). Each stimulus is represented by a function. Therefore, the distance between stimuli can be defined as a distance between functions. It turns out that the similarity kernel has a natural distance between functions associated with it. This metric is for example bounded from above. The distance cannot become greater than a certain value. Psychologically this is a very desirable property because at some point stimuli are just completely different and they cannot be made more different than that. Furthermore, the distance that we discussed in Chapter 3 does not have additive segments. In a series of papers Tversky and colleagues have heavily criticized geometric representations of similarity (Beals et al., 1968; Tversky, 1977; Tversky & Gati, 1982). It often goes unnoticed that the representations that they criticize are all based on metrics with additive segments. Metrics without this property escape most of their criticism. The similarity kernel can be used to define such a metric without additive segments in an elegant way.

1.2. Neural networks, revisited. The similarity kernel forms the basis of many categorization models. As noted several times now, exemplar theories (in particular, but other methods, too) make heavy use of the similarity kernel (Nosofsky, 1986; Kruschke, 1992). The idea that underlies all exemplar models is that stimuli are stored in memory and new stimuli are categorized based on the similarity to the stored exemplars. This idea can be formalized in a neural network. In fact, the ALCOVE model for categorization that will be discussed in more detail in below is such a neural network model.

Imagine a cell that after learning is tuned to an exemplar x_i. It will also respond to other stimuli x if they are sufficiently similar to x_i. To model the similarity we use of course the similarity kernel as given in equation (32). In exemplar models the similarity to several exemplars $x_1, ..., x_N$ is usually a weighted sum of the similarity to each exemplar:

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i). \tag{50}$$

The function that this equation computes can be represented graphically as a one-layer neural network. Figure 6 shows such a network. In the neural network literature nets with “tuning functions” similar to the similarity kernel are called radial basis function (RBF) nets. These nets have repeatedly been advocated as a plausible model for brain function by Poggio and coworkers (Poggio, 1990; Poggio & Bizzi, 2004).

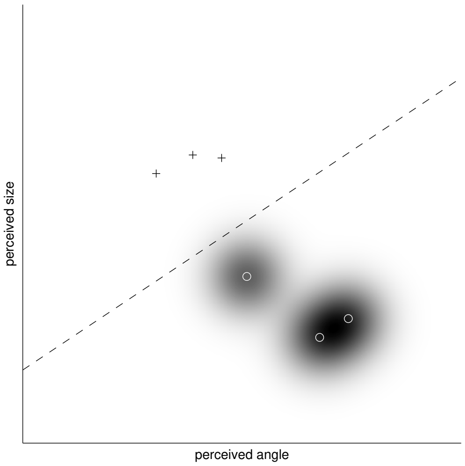

FIGURE 2. The summed similarity to the exemplars of one category is depicted with gray levels. The summed similarity is akin to kernel density estimation in statistics.

From a theoretical viewpoint it is useful to see the function f as a linear combination of basis functions $k(\cdot, x_i)$. The basis functions are the similarity kernels centered on the exemplars. Each exemplar is represented by its similarity to all other stimuli in the perceptual space. The function f is composed out of these similarity functions. Mathematically speaking the functions that can be implemented in this way form a vector space. If k is a positive definite kernel—as are many of the kernels that are used in psychology—this vector space can be given an inner product and a corresponding metric. This is the kernel metric that was introduced in Chapter 3.

For now it is enough to imagine that all weights are set to one. Figure 2 shows again the two categories of circles with spokes that were already depicted in Figure 1 of Chapter 3 (p. 35). The summed similarities to all exemplars of one of the categories (circles) is shown by gray levels. The more black a region of perceptual space is the more similar stimuli in this region are to the exemplars of the category. We have used a Gaussian kernel for illustration even though the circles with spokes have separable dimensions. The generalization gradient of the Gaussian can be seen very clearly for the one single stimulus close to the dashed category boundary. The kernel is simply put on top of the exemplar. For the other two stimuli in this category the generalization gradients overlap quite a bit and form a bigger hump. Regions with a high density of exemplars will therefore lead to a high output of the exemplar network (50), if all the weights are set to one. Hence, the output of the network can be interpreted as a measure for category membership or, if appropriately normalized, as an estimate for the probability that a stimulus from the category lies in a certain region of space. In statistics the same idea is used in so-called kernel density estimators (Ashby & Alfonso-Reese, 1995).

1.3. Conclusions. We have briefly recapitulated the idea of a perceptual space and similarity measures that are based on Shepard's universal law of generalization. We noted again that Shepard's law is akin to what is called a kernel in machine learning and statistics. Ashby and Alfonso-Reese (1995) have already compared exemplar theories of categorization to kernel density estimators. However, recently, methods based on kernels have attracted a lot of attention in machine learning. In what follows we will first systematically compare two psychological exemplar theories (GCM and ALCOVE) to a method from machine learning: kernel logistic regression. We will then go on to address the issue of generalization from a machine learning point of view.

2. Exemplar models

Historically, the first use of the similarity kernel was in an identification task (Shepard, 1957). This identification task is also the theoretical backbone of one of the most prominent exemplar models, the Generalized Context Model (GCM, Nosofsky, 1986). ALCOVE (attention learning covering map, Kruschke, 1992), a connectionist variant of the GCM, also makes heavy use of the similarity kernel. In the following we will trace the development of the GCM from the identification task and give a detailed comparison of the GCM and ALCOVE, highlighting the differences in the use of the similarity kernel.

Taking a kernel-view onto exemplar models also reveals their relationship to RBF-networks and machine learning methods, especially a method called kernel logistic regression. We believe the connections between categorization models and their heritage become clearer if they are discussed in the context of the mapping hypothesis, and this is what we will do first.

2.1. The mapping hypothesis. In two seminal studies Shepard et al. (1961) and Shepard and Chang (1963) examined the relationship between identification and categorization. In identification tasks participants learn to call each of a set of stimuli by a unique name. This may be achieved in a paired-associate paradigm where the experimenter shows the stimuli repeatedly to the participant and asks her for the corresponding name. If the participant calls the wrong name she is corrected. During this process of learning stimuli that are more similar to each other are confused more often. This is not necessarily due to their perceptual indiscriminability. The original idea in these studies relates back to the idea of generalization gradients: Stimuli are confused because their generalization gradients overlap and not because they cannot be discriminated. But of course stimuli might also be confused due to their insufficient representation in memory. Over time the participant will have built a better representation of the stimuli and will have associated each stimulus with its unique label—at least as far as this is possible given memory constraints and the discriminability of the stimuli.

A very simple hypothesis about categorization suggests that categorization might work similar to this rote-learning mechanism for identification. For each stimulus in the set the participant has to learn a label, the only difference being that in the categorization task labels are not uniquely identified with a stimulus. If there are two categories then there are only two labels but many more stimuli. In the cited studies it was hypothesized that the participants have the same pattern of confusions as in the identification task: More similar stimuli are confused more often. Therefore, it should have been possible to predict the errors in categorization from the errors in identification. Confusions within a class do not lead to mistakes but when stimuli from different classes are confused then an error is made. This was later called the mapping hypothesis (Nosofsky, 1986).

It turned out that the mapping hypothesis is not very good at predicting categorization performance, at least not for separable dimensions (Shepard et al., 1961). It provides a better account for integral dimensions (Shepard & Chang, 1963). One explanation could be that categorization is more than just rote-learning and some sort of abstraction, like formation of a prototype (Posner & Keele, 1968), is happening. Another explanation was suggested by Shepard et al.: Even if the underlying representation is the same in both tasks a participant's attention might be directed to different dimensions of the stimuli in the two tasks. For example, if one of the dimensions of the stimuli is more diagnostic for the categories than the other dimensions participants could put more attention on it. This could reduce the confusions along this dimension. One way to achieve this is to assume that similarity is not a unitary concept that is invariant under the different tasks. Similarity could be dependent on whether the subject is performing an identification or a categorization task. This idea is formalized in Nosofsky's GCM (Nosofsky, 1986). In Nosofsky's experiments it proved to provide a better account of categorization performance than the simple mapping hypothesis.

2.2. The MDS-choice model. As an identification model is the starting point for the GCM, it is natural to describe the model for the identification task first. In each trial a participant has to choose a response from a set of possible responses. A very simple and widely-used model for choice behavior in general was investigated by Luce (1959, 1963, 1977). The model has close connections to the method of paired comparisons, as well as the modeling of tournaments, and, finally, to logistic regression (Bradley, 1976; David, 1988). It is still widely used in in psychology and economics (McFadden, 2003; Train, 2003; Kuss et al., 2005; Wichmann & Hill, 2001; Jäkel & Wichmann, 2006) even though it is known to be problematic in several respects (Tversky, 1972; Luce, 1977; Görür et al., 2006). For an identification task a model in the same spirit was first discussed by Shepard (1957). The probability of answering with response r_i when the stimulus was x_j is given by Luce's well-known choice rule

$$P(r_i|x_j) = \frac{\pi_{ij}}{\sum_{k=1}^{N} \pi_{kj}}, \tag{51}$$

where the number of stimuli and responses is N. In Shepard's identification model π_{ij} is interpreted as the similarity between the stimuli x_i and x_j (with π_{ij} being positive). This basic model is usually supplemented with response bias terms that we will ignore for simplicity.

If no additional structure is assumed for the π_{ij} nothing is gained from this formulation. Shepard (1957) assumed that the π_{ij} are a monotonically decreasing function of the distance between the stimuli x_i and x_j in a psychological space. To make the model feasible he also assumed that the psychological space is Euclidean and that the relationship between similarity and distance is exponential. Shepard's suggestion was essentially to use equation (48), $\pi_{ij} = k(x_i, x_j)$.[1] This model has become to be known as the MDS-choice model (Nosofsky, 1986).

For example, imagine the psychological space to be two-dimensional. Instead of having to estimate the N^2 probabilities of confusion only the $2N$ coordinates of the stimuli have to be estimated. As Shepard (1957) assumed the distances in the similarity kernel (48) to be Euclidean he could use classical multidimensional scaling to recover the coordinates. Later he used his ordinal scaling method to get independent support for the shape of the similarity kernel (Shepard, 1965, 1987). Today, the similarity kernel is usually assumed to be known and the coordinates in

[1] He used $p = 2$ together with $\gamma = 1$. Furthermore, he constrained the π_{kj} in the denominator to add to one so that his similarity measure is directly given by the bias corrected confusion probabilities.

the multidimensional space are routinely estimated by using maximum likelihood (Nosofsky, 1986)[2].

By a simple reparameterization $\pi'_{ij} = \log \pi_{ij}$ it is easy to see that Luce's choice rule (51) is identical to the multinomial logit model (Train, 2003)

$$P(r_i|x_j) = \frac{\exp(\pi'_{ij})}{\sum_{k=1}^{N} \exp(\pi'_{kj})}. \tag{52}$$

It is instructive to note that if π'_{ij} is a linear function of some observed variables standard logistic regression is recovered. Consider the simple case where in each trial of an experiment there are just two possible responses and just two stimuli, e.g. a subject having to decide which of two possible weights is given into his hand. One might want to try whether the mass m_1 of stimulus x_1 and the mass m_2 of stimulus x_2 can be used as predictors of choice probability. With a scale parameter β that has to be estimated (and again ignoring response biases) the probability for the first response when the first stimulus is presented is

$$\begin{aligned} P(r_1|x_1) &= \frac{\exp(\beta m_1)}{\exp(\beta m_1) + \exp(\beta m_2)} \\ &= \frac{1}{1 + \exp(-\beta(m_1 - m_2))} \\ &= \text{logistic}(\beta(m_1 - m_2)) \end{aligned}$$

the logistic function of the difference of the masses. As weighing follows the Weber-Fechner law it seems more appropriate to regress on the logarithm of the mass rather than the mass. However, even if one knew on which transformation of a variable to regress, it raises the issue whether it should be identified with π_{ij} in Luce's choice rule or with π'_{ij} in the logit model.

In the identification model we want the similarity of the stimuli to be the variable that controls response probabilities. In the MDS-choice model the similarity is identified with π_{ij} and therefore the multinomial logistic is calculated over the logarithmic similarity. If (49) is used as a similarity measure the similarity is an exponential of the p^{th} power of d_p, the distance in psychological space. Hence, the logistic is calculated on the p^{th} power of d_p and not on the similarity:

$$P(r_i|s_j) = \frac{k(x_i, x_j)}{\sum_{k=1}^{N} k(x_k, x_j)} = \frac{\exp(-d_p(x_i, x_j)^p)}{\sum_{k=1}^{N} \exp(-d_p(x_i x_j)^p)}.$$

Hence, when interpreted as a logit model the MDS-choice model does not make use of the similarity kernel. It is a logit model on the p^{th} power of d_p. Interestingly, the p^{th} power of d_p is a metric for $0 < p < 1$ even though d_p is not a metric for $0 < p < 1$. We have discussed this metric as one of the metric alternatives to Tversky's famous contrast model (Tversky, 1977) above, in Chapter 3.

2.3. The Generalized Context Model. Using the mapping hypothesis it is straightforward to work out the probabilities for the category responses once the identification model is specified. A number of categories $C_1, \ldots, C_M$ with associated responses $R_1, \ldots, R_M$ are defined in a way that each possible stimulus $x_1, \ldots, x_N$ belongs to exactly one of the categories. The probability of observing response R_m given the stimulus was x_j is then

$$P(R_m|x_j) = \sum_{x_i \epsilon C_m} P(r_i|x_j) = \frac{\sum_{x_i \epsilon C_m} \pi_{ij}}{\sum_{k=1}^{N} \pi_{kj}} = \frac{\sum_{x_i \epsilon C_m} \pi_{ij}}{\sum_{m=1}^{M} \sum_{x_i \epsilon C_m} \pi_{ij}}. \tag{53}$$

[2] A very similar procedure for dimensionality reduction has been suggested in machine learning recently (Hinton & Roweis, 2003)

For simplicity we have again ignored response biases. Following the MDS-choice model, Nosofsky identified the similarity measure (32) with $\pi_{ij} = k(x_i, x_j)$ (Nosofsky, 1986, 1987). He called this model the Generalized Context Model because with a certain choice of similarity kernel it can be seen as the continuous generalization of an earlier exemplar model with binary features that was called Context Model (Medin & Schaffer, 1978). Note that the identification model is recovered if every stimulus has a unique label, that is there is a different category for each stimulus.

From a statistical viewpoint the GCM is a multinomial logit model, too. Let us introduce the shorthand

$$f_m(x) = \sum_{x_i \epsilon C_m} k(x, x_i)$$

for the sum of the similarities. This is a special RBF-network (50) with the weights for the exemplars in a category set to one and the other weights set to zero—but note that a later formulation of the GCM explicitly includes weights for exemplars (Nosofsky, 1992). Consider the case with only two categories. The GCM (53) then simplifies to

$$\begin{aligned} P(R_1|x_j) &= \frac{f_1(x_j)}{f_1(x_j) + f_2(x_j)} \\ &= \frac{1}{1 + \frac{f_2(x_j)}{f_1(x_j)}} \\ &= \text{logistic}(\log f_1(x_j) - \log f_2(x_j)). \end{aligned} \tag{54}$$

Nosofsky (1986) first fitted the MDS-choice model to identification data. This resulted in a map of the stimuli in the psychological space. Ideally, from this it would have been possible to predict the performance of participants in a categorization task directly. However, a naive application of the mapping hypothesis does not give accurate predictions. As mentioned before, one explanation for this failure posits a change in the similarity measure due to attention. Nosofsky allowed the model some extra flexibility by allowing the weights α_i in the distance d_p (47) to be different in the identification and the categorization task. A higher weight for a dimension may be interpreted as allocating more attention on this dimension. The categorization model that is constrained in this way, with all free parameters except the weights determined by the identification task, was able to account for his categorization data.

2.4. ALCOVE. Inspired by the success of the GCM, and probably also by the general excitement about neural network models at the time, Kruschke (1992) developed a connectionist variant of the GCM. As crucial ingredients for his model he identified the similarity measure that can be given a tuning curve interpretation and the attention weights that Nosofsky used. He formulated the model as a network and added a backpropagation learning algorithm to account for the adjustment of the attention weights—hence the name ALCOVE (attention learning covering map).

There are input nodes for each psychological dimension and they can be scaled with the attention weights. In one version of ALCOVE, there is a neuron in the hidden layer for each exemplar that occurs in an experiment. The activation of the hidden layer neurons is determined by the similarity measure, that is the similarity of the input to the stimulus they are tuned to. This version of ALCOVE is most similar to the GCM in that both assume that there is an explicit representation of the exemplars and only the exemplars. The "covering map" in the acronym ALCOVE actually refers to another variant of ALCOVE where it is assumed that the hidden layer neurons cover the whole input space. Even before the network sees any exemplars there are neurons tuned to parts of the input space. In both cases, the response of a hidden neuron that is tuned to a stimulus x_j to a presentation of stimulus x_i is $k(x_i, x_j)$.

There are important differences between ALCOVE and the GCM. In the GCM, as given in equation (53), the similarities to all exemplars are simply added up. In ALCOVE the output neurons collect a weighted sum of all hidden neurons. Assume again there are M categories $C_1, \ldots, C_M$ and one output neuron for each category. The activation f_m of the neuron that is responsible for category C_m is defined as a weighted sum of the activation of the hidden layer neurons:

$$f_m(x) = \sum_{i=1}^{N} w_{mi} k(x, x_i).$$

Each output neuron m has its own weights that are collected in a vector w_m. Each output neuron is an RBF-network with a kernel given by the similarity measure (see equation (50) and Figure 6). Instead of using Luce's choice rule and generating the probability for the category responses with the mapping hypothesis, ALCOVE uses the logit response rule (52) directly on the weighted similarities to the exemplars without recourse to an identification task

$$P(R_m|x_j) = \frac{\exp(f_m(x_j))}{\sum_{m=1}^{M} \exp(f_m(x_j))}.$$

An identification task can of course be set up by having as many categories as stimuli but the identification and the categorization task cannot be linked by the mapping hypothesis. This important conceptual difference between the GCM and ALCOVE should not be overlooked because the mapping hypothesis provided the main motivation for the GCM.

For obvious reasons ALCOVE is called kernel logistic regression in machine learning and statistics (Hastie et al., 2001). It is an RBF-network combined with a logit model. In the important case where there are just two categories ALCOVE reduces to

$$\begin{aligned} P(R_1|x_j) &= \frac{\exp(f_1(x_j))}{\exp(f_1(x_j)) + \exp(f_2(x_j))} \\ &= \text{logistic}(f_1(x_j) - f_2(x_j)). \qquad (55) \end{aligned}$$

The first term in the logistic function is a non-parametric measure for the degree that the stimulus belongs to the first category. The second term does the same for the second category. The logistic function is simply applied to the difference of the two category scales[3].

2.5. Comparison of GCM and ALCOVE. Figure 3 shows a comparison between ALCOVE and the GCM for a simple two category classification task. For both models the attention and the exemplar weights are set to one. Both models are depicted with the Euclidean and the city-block metric. On the equivocality contour the summed similarity to all exemplars of one class equals the summed similarity to the exemplars of the other class (Ashby & Maddox, 1993). As we assume subjects are unbiased the probability for the subject to respond with one class is one half. The equivocality contour is shown as a dashed line. First note that the equivocality contour is the same for the GCM and ALCOVE. ALCOVE

[3] In the two-category case ALCOVE is heavily overparameterized. There is a full RBF-network f_1 with as many weights as exemplars for category one and a full network f_2 for category two. One RBF-network $f = f_1 - f_2$ with the weights set to the difference, $w_{1i} - w_{2i}$, would be enough.

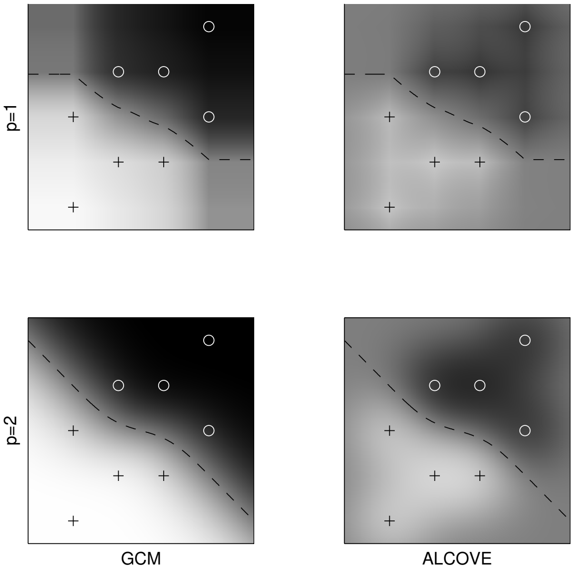

FIGURE 3. A comparison of GCM and ALCOVE for the city-block metric ($p = 1$) and the Euclidean metric ($p = 2$). Circles and crosses depict exemplars from two classes. For both categorization models the attention and the exemplar weights are set to one. The grayscale shows the response probabilities that would be obtained for a new stimulus at each position. White areas are classified as 'cross' with probability one. Black areas belong to the other class. The dashed line depicts the equal-probability contour which is the same for the GCM and ALCOVE. Outside the generalization gradients of the exemplars the models make very different predictions.

performs logistic regression on the difference of the summed similarities (55) and therefore the choice between Euclidean and city-block only makes a difference close to the exemplars. Beyond the generalization gradients of the exemplars categorization performance drops to chance level because the summed similarities go to zero. ALCOVE and the GCM make very different predictions on stimuli that are outside the generalization gradients for the exemplars. The GCM operates on the log of the summed similarities (54). Therefore, points that are clearly on one side of the decision bound are categorized more easily in the GCM.

Showing no generalization beyond its generalization gradients one could therefore say that ALCOVE behaves like Spence's classic model for discrimination learning (Spence, 1937) and thus shows no "true" categorization behavior. To illustrate this in the simplest situation possible imagine that there are just two categories with only one stimulus each. Further assume that the perceptual space is only one-dimensional. Such a one-dimensional space is depicted in Figure 4. The dotted lines depict the generalization gradients for the two stimuli from the two categories. One stimulus (with positive values on the right y-axis) is located at $x_1 = +1$ the other stimulus is located at $x_2 = -1$ (with negative values on the right y-axis). The generalization gradients of the stimuli overlap even though they belong to different categories. Let us calculate the probability that ALCOVE will categorize a new stimulus x as belonging to the same category as x_1 by using equation (55) and the similarity kernel (49):

$$\begin{aligned} P(R_1|x) &= \text{logistic}(f_1(x) - f_2(x)) \\ &= \text{logistic}(\exp(-d_p(x,x_1)^p) - \exp(-d_p(x,x_2)^p)) \\ &= \text{logistic}(\exp(-|x-x_1|^p) - \exp(-|x-x_2|^p)). \end{aligned}$$

As before, we have simply assumed that the weights for each of the exemplars is set to one. Therefore, the response of category scales f_1 and f_2 is directly given by the generalization gradient for the two stimuli x_1 and x_2, respectively. The probability that ALCOVE responds with category one (R_1) is shown as a dashed curve in Figure 4 (the scale is on the left y-axis). The Gaussian case where $p = 2$ is given in the lower panel and the exponential with $p = 1$ is given in the top panel. As the distance from the two categories increases towards the right or left end of the x-axis the similarity to both stimuli goes to zero. For a new stimulus x that lies outside the generalization gradients of both stimuli the probability of responding with either category is close to one half.

In contrast to ALCOVE, the GCM is capable of categorization beyond its generalization gradients. Intuitively, if a subject has really learned to categorize the two stimuli—as opposed to only discriminate them—one would expect that stimuli that are more extreme than the training exemplars are categorized easily. This criterion is used in animal studies to define categorization (e.g. Ohl, Scheich, & Freeman, 2001). In the example in Figure 4 stimuli that have a larger x-value than x_1 should be categorized easily because they are further away from zero. While this may sound like subjects have to implement an explicit rule in order to behave accordingly the GCM shows this behavior without representing a decision bound explicitly. This can be illustrated in the simple one-dimensional case with only two stimuli. Using the same assumptions as for ALCOVE in this example the GCM (54) becomes:

$$\begin{aligned} P(R_1|x) &= \text{logistic}(\log f_1(x) - \log f_2(x)) \\ &= \text{logistic}(-d_p(x,x_1)^p + d_p(x,x_2)^p) \\ &= \text{logistic}(-|x-x_1|^p + |x-x_2|^p). \end{aligned}$$

The logistic function is calculated on the distance and not the similarity kernel (because there is only one stimulus in this simplified example the logistic and the exponential annihilate each other). The exact behavior of the GCM depends on the exponent p. In our example, if $p = 2$ the logistic will depend on x. The bigger x the higher the probability that a stimulus is categorized as category one. For $p = 1$ the response probability does not depend on x if x is bigger (smaller) than x_1 and x_2.

2.6. Conclusions. We have traced the history of exemplar models and the similarity kernel back to the work of Shepard (1957, 1958) on generalization gradients and identification tasks. A bit later the idea to link identification and categorization via the mapping hypothesis was first tested experimentally (Shepard et al., 1961; Shepard & Chang, 1963). In parallel Shepard (1962) developed his ideas on ordinal MDS. Using the concept of attention weights Nosofsky (1986, 1987) was able to assemble all the parts into a working model of categorization and link it to the existing literature on exemplar based categorization (Medin & Schaffer, 1978). A bit later still, Kruschke (1992) suggested a connectionist variant of the GCM that is closely related to RBF-networks (Poggio, 1990) and kernel logistic regression (Hastie et al., 2001). We have seen that both, the GCM and ALCOVE, are based on the logit rule and the use of the similarity kernel but with important

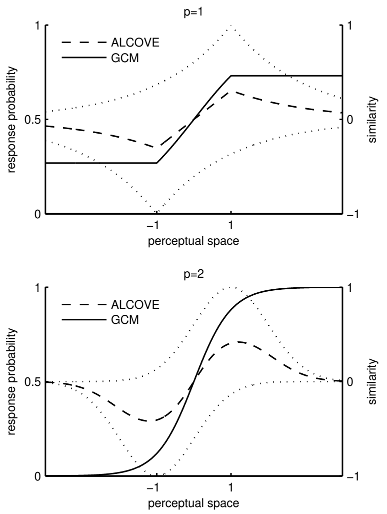

FIGURE 4. A comparison of GCM and ALCOVE for the simplest case with just one dimension and only one stimulus per category. The dotted lines depict the generalization gradients (scale on the right axis) for two stimuli. The stimulus for category one is centered at $+1$ and the stimulus for category two is centered at -1 with negative similarity values. The probability that ALCOVE would respond to a new stimulus with category one is given as the dashed line (scale on the left axis). The response probability of the GCM is given as a solid line.

differences. In contrast to the GCM, ALCOVE does not use the mapping hypothesis. Also, ALCOVE does not show much categorization beyond its generalization gradients. This demonstrates that the way the generalization gradients enter the response rule in a categorization model has an influence on how new stimuli will be classified. Whether this classification is likely to be correct is, however, also influenced by other factors that we will discuss in the next section.

3. Generalization

There is an intriguing duality in fitting a categorization model to human categorization responses. A subject performing a categorization task produces categorical data that is then analyzed by fitting a categorization model, not to the category labels that the subject learned but to the categorical responses of the subject. A categorization model, a classifier in machine learning terms, can be applied to any data with categorical responses—be it a problem of medical diagnosis, bioinformatics, machine vision or participants pressing buttons in a psychological experiment. A human trying to learn new categories in a categorization experiment has to solve the same statistical problem as a machine classifier that is used to explain some categorical data. Both, humans and machines, have to try to find a regularity in the data they observe. Hence, the issue of generalization arises twice in psychological categorization models. First, there is the issue whether a chosen model can explain the category responses of subjects in an experiment and whether the model will be able to generalize to new experimental situations. Second, there is also the issue of whether a human subject who has learned a category according to the model would be able to generalize to new stimuli—that is whether the category is learned well.

A very flexible model like kernel logistic regression will give a good statistical description of many data-sets and will show a good generalization performance in many applications, including psychological modeling, if it is used properly. Recently, the issue of whether psychological models, and categorization models in particular, provide a good statistical description of experimental data has received increased attention (Pitt et al., 2002; Pitt, Kim, Navarro, & Myung, 2005). The crucial question always is how well will the model predict future data, that is will it generalize.

When applied to human responses in a categorization task a statistical model like kernel logistic regression also suggests how humans perform the categorization task. Providing a good statistical description of the experimental data alone cannot be the hallmark of model selection in this case. It is certainly a necessary prerequisite but the process model that is suggested by the categorization model also has to be psychologically plausible. The development of the GCM was guided by the mapping hypothesis. It was not just a blind application of the logit model. Similarly, ALCOVE is not just an application of kernel logistic regression but in addition to accounting for the response probabilities it also tries to account for the time-course of learning. All three—the GCM, ALCOVE, and kernel logistic regression—assume the underlying representation that the subject uses is based on storing exemplars in memory.

While the experimental evidence for exemplar theories has been much debated, the kernel-view gives theoretical justifications for using exemplar theories in the first place. This section deals with the important problem of how subjects can generalize in categorization tasks. How is it possible to assign the right category label to a new stimulus that has never been encountered before based on the limited experience with other stimuli of the same kind? If categorization is said to be a useful behavior then mostly because of its role in induction and prediction (Anderson, 1991). Correct categorization allows organisms to apply knowledge about a category to an individual object. Hidden properties can be inferred and interaction with the object can be planned. In the exemplar models that we have presented generalization is explained by appealing to similarity. Psychologically, this makes a lot of sense because it links categorization behavior with classic work in stimulus generalization for classical conditioning and discrimination learning. Based on Shepard's work exemplar theorists have basically completely identified similarity with generalization. However, we will argue that exclusive reliance on similarity will not necessarily lead to good generalization performance. There are additional statistical considerations that need to be taken into account. This will not come as a surprise to the critics of exemplar theories who have always doubted that merely remembering exemplars can lead to proper categorization. This does, however, not mean that exemplar models cannot generalize. Quite to the contrary: In machine learning, kernel methods are among the most successful tools precisely because they are known to generalize well—if they are used wisely. We will show how exemplar theories can be made to reliably extract the structure underlying a category. To this end, we will discuss how kernel methods in machine learning and statistics deal with the problem of generalization.

3.1. Kernel density estimation. The category learning problem is often phrased as a density estimation problem (Aizerman, Braverman, & Rozonoer, 1964a; Fried & Holyoak, 1984; Nosofsky, 1990; Ashby & Alfonso-Reese, 1995). Imagine two classes, exemplars from each category are drawn from a probability density function that completely determines the distribution of features within each category. If a learner knew the distribution of features within a category she could examine the features of a new stimulus and assign it to the category with the

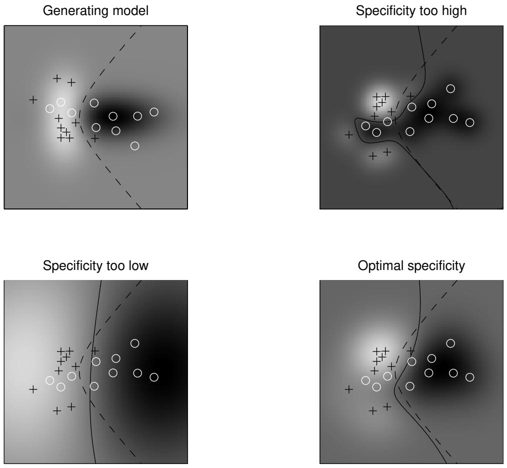

Figure 5. A two class problem. Samples from two classes (crosses and circles) are generated from two overlapping Normal distributions. In the upper left panel the difference of their densities is shown as gray-levels and the optimal decision bound is shown as a dashed line. The other panels show kernel density estimates of the two classes with varying specificity of a Gaussian kernel and the corresponding decision bounds.

highest likelihood of having generated this pattern of features. Hence, learning to categorize could mean learning the distribution of features[4].

The upper left panel in Figure 5 gives an example. It shows a two dimensional stimulus space. The features of each stimulus are represented by coordinates in the stimulus space. This space could either be a physically specified space or a perceptual space. The difference of the densities of two overlapping Normal distributions is indicated by the gray levels. The darker regions correspond to high density regions of one of the classes whereas the lighter regions correspond to the other class. Probabilistic category structures similar to this one are frequently used in experiments (Fried & Holyoak, 1984; Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995, 1996). We have drawn ten exemplars from each of the distributions for illustration (circles and crosses).

As we know the distribution of the two classes we can calculate the optimal decision bound between the two classes which for two Normal distributions is generally quadratic (Ashby & Maddox, 1993). The optimal decision bound is shown as a dashed line. A subject that tries to maximize performance, that is correct responses, should place the decision criterion along the optimal decision bound. On one side of the decision bound the subject should always choose one category label and on the other side she should always choose the other label. This sharp bound without probabilistic responding will give the best generalization performance. However, subjects may not respond deterministically and different models make different assumptions about how probabilistic decision are (Ashby & Maddox, 1993). In the following we will ignore this additional complication and only

[4] It is, however, potentially easier to directly learn the decision function rather than trying to solve the difficult problem of density estimation first (Vapnik, 2000).

talk about generalization performance under the assumption of (almost) deterministic responding. But we want to mention that there is evidence that subjects do respond deterministically under certain circumstances (Ashby & Gott, 1988) and that exemplar models can be adapted to account for this (Ashby & Maddox, 1993).

Even though we can calculate the optimal decision bound for the example in Figure 5, the subject cannot know the true distributions because all the subject has observed is a limited number of exemplars from these two categories. Therefore, the subject cannot respond optimally. One possible strategy in this case is to try to estimate the two category distributions from the observed exemplars and choose a decision bound that would be optimal for the estimated category distributions. This can be done by assuming a particular parametric family for the category distributions and trying to estimate their parameters. For example, a category learner may assume that the distributions are Normal in which case she has to estimate means and covariances (Fried & Holyoak, 1984). This strategy will work well if the underlying category structure that she tries to learn is approximately Normal. A more flexible category learner would, however, try to avoid making very specific assumptions about the unknown distributions.

Exemplar models have been compared to the more flexible (non-parametric) kernel density estimators (Ashby & Alfonso-Reese, 1995). In the simplest exemplar model each data point is replaced by a kernel function, e.g. a Gaussian kernel. As explained above the summed similarity to all exemplars from one class can be seen as a density estimate. The upper right panel of Figure 5 shows such a kernel density estimate. Black areas have a high similarity to the exemplars of one of the classes (circles) and white regions have a high similarity to exemplars of the other class (crosses). For the density estimator a high similarity to exemplars from one class translates into a high likelihood that a new stimulus that falls into this region belongs to the corresponding category. The black solid line indicates the equivocality contour where the similarity to the exemplars from one class equals the similarity to the exemplars from the other class. This equivocality contour could be used as a decision bound.

3.2. Finding the right kernel. In the example in the upper right panel of Figure 5 the specificity, that is the width, of the similarity kernel is chosen to be too narrow. New stimuli are essentially categorized in the same way as the most similar past exemplar. If this exemplar happens to lie on the wrong side of the optimal decision bound it is very likely that a wrong decision will be made. The similarity kernels of different exemplars hardly overlap and therefore generalization to new stimuli is poor. The decision bound that the category learner chooses is able to categorize all past exemplars perfectly but only because she has learned the idiosyncrasies of this particular set of exemplars. This is called overfitting. The learner has not learned anything about the structure of the categories but instead has only learned the labels and the exemplars by heart. The bottom left panel shows the opposite case where the specificity of the kernel is chosen to be too low. A wide similarity kernel means that exemplars far away from a new stimulus can influence the guess to which category it belongs. Also in this case the resulting decision bound will be very different from the optimal decision bound. Hence, it is important to choose the width of the similarity kernel to be appropriate for the problem and the sample size at hand in order to assure good generalization performance. The lower right panel of Figure 5 shows the decision bound that results from a well-chosen kernel width.

Sometimes it may be possible for a subject to choose a reasonable kernel width before seeing the first exemplars but in general the specificity and the relative contribution of the attention weights have to be adapted by learning as well. In

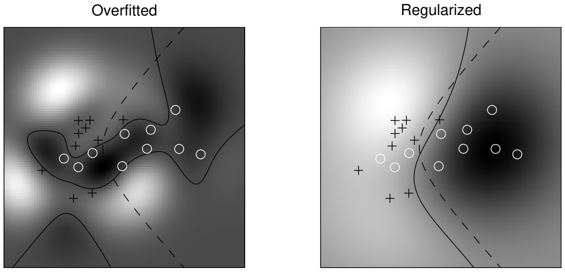

Figure 6. Unless regularized an exemplar model with exemplar weights will overfit.

machine learning choosing a kernel and setting the parameters of the kernel is considered to be a model selection problem. In psychological models the form of the kernel is given but its parameters may be adapted during learning. ALCOVE adapts its attention weights during learning but Kruschke (1992) does not directly address generalization performance. In machine learning a common way to choose the best parameters for the kernel is by using cross-validation procedures (see Pitt et al., 2002, for an overview on model selection and cross-validation). Instead of trying to minimize the error on all the known exemplars—which can always be driven to zero by choosing a narrow enough kernel as seen in Figure 5—one tries to obtain an estimate of the generalization error by repeatedly splitting the data into a training and a test set. For a certain setting of the specificity one asks how well the model uses the exemplars in the training set to predict the category membership of the exemplars in the test set. The parameter value that gives the lowest estimated generalization error is the one that will be used. This procedure was applied to obtain the specificity value for the lower right panel of Figure 5. We are not suggesting that human subjects use a procedure akin to cross-validation but we want to point out that from a normative point of view the choice of similarity kernel is crucial. If the similarity kernel is adaptable then subjects should pay close attention to their generalization performance while changing it.

3.3. Overfitting with exemplar weights. The problem of overfitting becomes even more pressing with the introduction of exemplar weights into categorization models—as already seen in Chapter 2. Both ALCOVE and a later version of the GCM have such weights (Kruschke, 1992; Nosofsky, 1992). It is desirable to introduce these weights for several reasons. It is unlikely that subjects are able to remember all exemplars and be able to attach the same weight to each of them. Probably there will be frequency and recency effects as well as forgetting. Some of the exemplars are more representative for a category than others and may get a greater weight. Furthermore, from a statistical point of view the exemplar weights introduce a greater flexibility which makes it possible to learn more complicated decision bounds. However, if these exemplar weights can be modified by learning it follows that each exemplar has its own free parameter—an almost sure recipe for overfitting (Pitt et al., 2002).

Recall that ALCOVE is built on an RBF-network. The RBF-network implements a function by expressing it as a weighted sum of kernel functions centered on the exemplars:

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i). \tag{56}$$

As noted before, Poggio suggests RBF-networks as a biologically plausible model for brain function (Poggio, 1990; Poggio & Bizzi, 2004). It is a common view to see the brain as a supervised learning machine. The network gets some input, calculates a function and receives feedback on the error it has made. This basic set-up is used in most artificial neural network approaches and underlies the backpropagation algorithm (Rumelhart et al., 1986). Hence, learning means to adapt the weights in equation (56) such that the error is minimized. The function that ALCOVE's backpropagation learning algorithm is trying to learn outputs a plus one for one of the categories and a minus one for the other category.

It is possible to give the optimal weights for this function without running a backpropagation algorithm. Let f be a vector of the function values $f_i = f(x_i)$ that we want the function to take on the exemplars. Let K be a matrix with entries $k(x_i, x_j)$ in the i^{th} row and j^{th} column. This matrix is called the kernel matrix. Let w be the vector of weights that we seek to implement the function. With this notation we can rewrite the neural network (56) in matrix notation as $f = Kw$. As shown in Chapter 2 K is positive semi-definite, if k is a positive definite kernel. For many kernels, such as the Gaussian, K is even positive definite and therefore invertible. Hence, we can find unique weights such that the function f makes no error at all on the exemplars: $w = K^{-1}f$. The resulting function f for the exemplars from Figure 5 is shown in the left panel of Figure 6. This function outputs a plus one for all exemplars from one of the categories and a minus one for all exemplars of the other category.

The fitted function does not capture the underlying regularity well. The reason for this is that by freely allowing the weights to be adapted we can override the similarity based categorization. The exemplar weights defeat the purpose of introducing a similarity measure for the stimuli. The similarity measure is introduced because similar stimuli should be treated similarly. Very similar stimuli are very likely to belong to the same category. However, the exemplar weights can be adjusted in a way that even very similar stimuli belong to different categories without interfering with each other. Imagine the case where we only have two very similar stimuli x_1 and x_2 that have very different function values $f(x_1) = 1$ and $f(x_2) = -1$. Say, their similarity is .99 and self-similarity is 1. In order to make the network (56) output the the right values, the small difference of .01 between their similarity and their self-similarity needs to be compensated by large weights of 100 and -100.

3.4. Regularization, revisited. One way to deal with overfitting in neural networks is regularization (Bishop, 1995; Orr & Müller, 1998). This is the approach discussed in Chapter 2 above and it is also used for kernel logistic regression (Hastie et al., 2001). The basic idea in regularization is that weights are not allowed to become too big. As large weights can override the similarity based categorization the weights should be as small as possible. This is achieved by trading-off the error that the classifier makes with the size of the weights. Recall that learning in the neural network setup means finding weights w such that a loss function $L(f)$ is minimized (where f is the function that an exemplar net with weights w

implements). Let us call the error that the classifier makes on the training exemplars $c(f)$. The penalty for large weights is called a regularizer and we denote it with $\Omega(\langle f, f\rangle)$ here. With this notation the loss function that a regularized RBF-network minimizes becomes (see Eq. 25 above):

$$L(f) = c(f) + \Omega(\langle f, f\rangle). \tag{57}$$

The regularizer reflects a "complexity" constraint on the function that the network implements. It is good if the the available data is fitted well but this should not be done at all costs. The fitted function should not be too complicated because complicated functions are more likely to overfit. Most model selection criteria trade goodness of fit versus model complexity (Pitt et al., 2002).

The right panel of Figure 6 shows the same categorization problem as before but this time regularization techniques were used. The gray levels code a function f of the form (56) that minimizes $L(f)$ in equation (25) with c chosen to be squared error and Ω chosen to be linear in the squared length of the vector w. For this loss function and several other interesting loss functions the optimal weights w are unique and can be found easily (see the discussion of the representer theorem and the regularization example in Chapter 2). Because of the regularization the category learner did not try to fit the available exemplars perfectly but instead traded off goodness of fit with the penalty term. Clearly, the model is closer to the optimal decision bound than without regularization. Intuitively speaking the regularizer penalizes large exemplar weights that are necessary to make two similar stimuli have different category labels.

It should be emphasized again that the exemplar network by itself does not guarantee a good generalization performance. After all, the exemplar network can always implement a function that can fit all exemplars perfectly—no matter what they look like. It is the joint choice of the kernel and the regularizer that determines the generalization performance of the network. The kernel captures some assumptions about the category structure. The regularizer penalizes greedy optimization of goodness of fit. Different problems require different kernels and different regularizers. In machine learning the kernel and the regularization parameters are usually chosen by cross-validation.

3.5. Learning a category with ALCOVE. The learning algorithm of ALCOVE greedily tries to minimize the classification error on the exemplars. In ALCOVE the error can be minimized in two ways: Firstly, by adjusting the attention parameters and therefore the generalization gradients (47) and secondly by adjusting the exemplar weights (56). We have shown above that for such models there is a danger of overfitting. If ALCOVE is shown the same exemplars over and over again its backpropagation algorithm can find a solution that categorizes these exemplars perfectly—no matter what the category structure is. In fact, even if ALCOVE's attention weights were non-adaptable there would always be a set of exemplars weights that allows perfect classification. As ALCOVE has been quite successful in describing subjects' learning curves in various categorization tasks this raises the question whether human subjects do also overfit. Considering that humans seem to categorize new stimuli reliably in every-day life this seems, however, unlikely. But perhaps humans do overfit in the experiments that they perform in the laboratory, and laboratory experiments are what exemplar theories try to model.

In most of the earlier experiments in favor of exemplar theories participants were shown a small number of exemplars over and over again. Remember that in the classic work of Shepard et al. (1961) and Shepard and Chang (1963) the original motivation was to see whether categorization can be described as mere rote-learning of labels—this was called the mapping hypothesis. Only eight stimuli were presented to the subjects, there was no noise and there were no transfer items. The GCM, too, was set up in order to link categorization with a rote-learning identification task and the accompanying experiments used only a small number of stimuli (Nosofsky, 1986). Also the experiments by Medin and Schaffer (1978), which are widely seen to provide good evidence for exemplar theories, have recently been criticized on the grounds that they used only few stimuli and poorly differentiated categories (Smith & Minda, 1998, 2000). Hence, subjects are perhaps encouraged to adopt an exemplar-memorization strategy in experiments even though they may not do so in every-day categorization. Some of the categories used in psychological experiments have so little structure that rote-learning of exemplars is in fact the only strategy that will make it possible to solve the task (Shepard et al., 1961; Feldman, 2000)[5]. If there are transfer items in these experiments they are only used to assess the predictions of the model (e.g. Medin & Schaffer, 1978; Nosofsky, 1986). There is usually no right or wrong answer for the subjects. Therefore, there is no rational strategy to which a participant's behavior could be compared in order to asses her generalization performance.

Other experiments have explicitly compared human performance with the performance of an ideal observer (Fried & Holyoak, 1984; Ashby & Gott, 1988; Ashby & Maddox, 1992; McKinley & Nosofsky, 1995, 1996). Those studies used overlapping probabilistic categories like the one shown in Figure 5. This scenario is perhaps more akin to natural category learning. Contrary to many categories in psychological experiments, natural categories have a structure. Presumably it is this structure that humans learn when they learn a category. Rosch and colleagues (Rosch & Mervis, 1975; Rosch et al., 1976) have argued that on the basic level the stimuli within a natural category share perceptual properties and that the distribution of the properties of a category are not completely random—but also not deterministically defined by necessary and sufficient conditions. As very little is known about the actual structure of natural categories we may choose to use categories like the one shown in Figure 5 as a proxy. This has the advantage that the number of possible exemplars is infinite and subjects never encounter the same exemplar again. Furthermore, there is an objective way to assess a subject's generalization performance. Clearly, under these conditions a strategy that simply remembers all encountered exemplars seems unreasonable. Some of the above studies have nevertheless successfully fitted exemplar models to human responses (McKinley & Nosofsky, 1995, 1996).

Interestingly, in this scenario with overlapping probabilistic categories, ALCOVE will not overfit as easily and its behavior results in exemplar networks that are regularized. A subject encounters a new stimulus but does not know its category label. She predicts the category of the stimulus based on previous exemplars. Then she receives feedback about the true category label. It is reasonable to set the exemplar weights to zero before an exemplar has been encountered. After ALCOVE is given the true category label of a new exemplar it may be necessary to assign a large weight to this exemplar in order to output the correct label. How much the weights are allowed to change is determined by the learning rate parameter in ALCOVE. If the learning rate does not allow big changes in the weights this is akin to regularization that also penalizes large weights in order to avoid overfitting. Limiting the influence of individual points has a regularizing effect by increasing the *stability* of the solution. Indeed, solutions that are stable in the sense of not depending too strongly on any individual training point can be shown to generalize

[5] Unless the subject redefines the perceptual dimensions as discussed by Shepard et al. (1961).

well with high probability (Bousquet & Elisseeff, 2002; Poggio, Rifkin, Mukherjee, & Niyogi, 2004).

Note also that the feedback that the subject receives is a direct measure of the generalization error—similar to cross-validation. The prediction error is a direct measure of her generalization performance because each stimulus is a new stimulus that has never been encountered before. This is in contrast to experimental procedures where the same stimuli are shown over and over again. Therefore, in the case where each stimulus is a new stimulus ALCOVE does not try to minimize the error on past exemplars but the prediction error on new exemplars. Early stopping in artificial neural networks is used for the same reason (Orr & Müller, 1998). Hence, for ALCOVE the learning rate parameter is crucial for the models generalization performance.

3.6. Prototype vs. exemplar models. ALCOVE is prone to overfitting when shown the same small number of stimuli over and over again. This does not necessarily constitute an argument against ALCOVE because also human subjects may simply memorize stimuli under such artificial conditions. Under more difficult—and realistic—conditions where there are plenty of stimuli and perfect categorization performance is not possible, ALCOVE behaves more reasonable: It will not overfit easily because it has a built-in regularization mechanism and directly minimizes prediction error. However, critics of exemplar theories may still object to the idea that all exemplars have to be stored in memory. Some prototype models view all stimuli as (random) distortions of the average stimulus of a category. In experiments, artificial category structures have been set up that can be described completely by the average stimulus, sometimes together with the covariance structure of the categories (Posner & Keele, 1968; Reed, 1972; Fried & Holyoak, 1984). In such experiments the task of the subject is naturally described as trying to *abstract the idea* that underlies the category. It is one thing to say that the category structure that the subject is supposed to learn is well described by a prototype but it is another thing to claim that subjects do extract the prototype when they learn such a category. However, if subjects in a task where the category is indeed defined by a prototype only memorize exemplars one would doubt that they understood the gist of the category. It seems they would miss the underlying regularity that defines the category if they only memorized the exemplars. But if real-world categories are more complicated than the prototype view suggests subjects should really adopt a more flexible strategy.

The prototype vs. exemplar debate can be framed in terms of mental representations. Subjects may store a summary representation of a category or they may store exemplars of the category. The debate can also be seen as being about which assumptions a category learner makes about the category she is learning (Ashby & Maddox, 1993; Ashby & Alfonso-Reese, 1995; Briscoe & Feldman, 2006). Prototype theories make very strong assumptions about the category structure. The whole category structure can be summarized by the prototype. This will lead to a good generalization performance even with only a few trials of learning if the category structure to be learned is really so simple. Exemplar theories with exemplar weights, like ALCOVE, are at the other extreme. They are very flexible category learners and can learn more complicated category structures. However, it is not true that they do not make any assumptions about the category structure. The assumptions are only given implicitly by the choice of kernel and the way that the learning algorithm sets the weights. Therefore, it is a lot harder to say, what it is that these models learn from the exemplars. Nevertheless, even if they do not abstract anything from the data they are able to learn something about the structure of the category that enables them to generalize to new stimuli.

Unless all the evidence in favor of exemplar theories is completely misleading because of small and ill-defined categories in the experiments (Smith & Minda, 1998, 2000), one would hope that exemplar theories scale up to real-world categorization behavior. Kernel methods in machine learning have already proven to be successful in real-world applications. And as we have shown, these methods build on similar intuitions as exemplar theories. In fact, kernel methods outperform other methods with more restrictive assumptions, like prototype classifiers, on real-world data sets (Schölkopf & Smola, 2002). This could suggest that the restrictive assumptions of prototype theories are not met for natural categories and more flexible mechanisms, as implemented in exemplar models, are needed to deal with real-world categories.

There remains the problem that seemingly all exemplars that are encountered need to be stored. However, the exemplar idea might scale up to a realistic number of stimuli if not all exemplars are remembered but only certain crucial ones. This problem has also been addressed in machine learning where it is also desirable to store only as few exemplars as necessary in memory. Intuitively, exemplars with small weights can be forgotten without changing the overall performance of the classifier. One of the reasons for the success of support vector machines in machine learning is that most of the coefficients in the kernel expansion (56) are indeed zero (Vapnik, 2000; Schölkopf & Smola, 2002). Hence, the corresponding exemplars need not be remembered. Solutions that only require few exemplars to be remembered are called *sparse* in machine learning. There are variants of several kernel classifiers, including kernel logistic regression, that try to achieve the same categorization performance with remembering fewer exemplars (Hastie et al., 2001; Schölkopf & Smola, 2002). The idea that a few representatives may be enough has been suggested in the object recognition literature (Poggio & Edelman, 1990) and is emphasized in several recent categorization models (Rosseel, 2002; Verguts et al., 2004; Love et al., 2004). The interesting psychological question is of course which exemplars are remembered and which are not. It could be a mere question of primacy, recency and frequency but there could also be representational considerations. On the one hand, some exemplars are simply better representatives for a category than others. On the other hand, some exemplars are more important to determine the decision bound between categories. Kernel methods in machine learning could inspire new psychological models that do not have to remember all exemplars but still achieve a good generalization performance, like the support vector machine.

3.7. Conclusions. Generalization is central to theoretical approaches to the statistical learning problem (Vapnik, 2000). In psychological categorization research the problem of generalization is often hidden behind the prototype vs. exemplars debate. Prototype theorists assume very restricted category structures and can therefore generalize well even with very few exemplars (Smith & Minda, 1998)—if their assumptions are true. Exemplar theories can deal with very complicated category structures but are prone to overfitting if not regularized properly. Our contribution here is to directly address concerns about generalization performance of exemplar theories by demonstrating how good generalization may be achieved. Our discussion has mainly been guided by regularization techniques as they are used in machine learning. We demonstrated that ALCOVE has mechanisms that are akin to regularization already built-in. The question whether humans regularize in a similar way, and if they do, what their regularization looks like, opens new directions for empirical research. There is evidence that humans cannot learn arbitrary category structures and that some categories are harder to learn than others (McKinley & Nosofsky, 1995; Feldman, 2000; Minda & Smith, 2001; Ashby, 2001; Alfonso-Reese, Ashby, & Brainard, 2002; Briscoe & Feldman, 2006). Such results potentially inform us about the restrictions within which category learning is possible and may give hints to the assumptions (e.g., small exemplar weights are to be preferred) on which humans base their category learning. Machine learning methods also suggest a middle-ground between prototype and exemplar theorists by showing that flexible categorization models are possible that do not need to remember all exemplars but still generalize well.

CHAPTER 5

Discussion

1. Exemplar models and object recognition

It has been noted in a recent review that there are parallels between the object recognition and the categorization literature (Palmeri & Gauthier, 2004). In both fields there are models that assume that the memorization of exemplars underlies human performance in the respective tasks. In most object recognition tasks the same object has to be recognized irrespective of view and lighting conditions. Often, however, object recognition also refers to basic-level categorization. To recognize an object is then the same as assigning it to the correct basic-level category. For categorization not only variations in view and lighting conditions need to be discounted but also variations in shape. A major aspect of the object recognition literature is that object recognition directly works on the images as an input to the recognition system—and therefore variations in view or lighting are as important as clutter in the visual scene. This is realistic because also the brain needs to work with the input from the retina. And ultimately, of course, visual categorization needs to start from the visual input. The models considered in this thesis, however, started from a perceptual space that is closer to the way the experimenter conceptualizes the stimuli. For example, if a participant categorizes rectangles the perceptual space will only consist of rectangles parametrized, say, by perceived width and height. As the perceptual space is a primitive in the theory nothing is said about how the features that a subject perceives are extracted from the image. Without being too specific about which features are extracted from an image there is evidence that object recognition might be achieved by storing different views of objects and is therefore similar to ideas presented in the exemplar model literature (Liter & Bülthoff, 1998; Tarr & Bülthoff, 1998).

Like exemplar models, some models in object recognition explicitly make use of RBF-networks (Poggio & Edelman, 1990; Bülthoff & Edelman, 1992; Riesenhuber & Poggio, 1999). The similarity kernel in exemplar models is certainly suggestive of a neural tuning curve in the same way as the RBF-kernels in object recognition models are interpreted as neural tuning curves. However, the similarity kernel resulted from work on multidimensional scaling and whether similarity in a perceptual space is really the same as a neural tuning curve is at least debatable. Nevertheless, it is certainly a very attractive idea that object recognition and categorization might be explained with the same mechanisms. Especially as these mechanisms take the form RBF-networks that are theoretically elegant and not implausible from a biological point of view.

2. Exemplar models and the brain

Given that object recognition and visual categorization are probably closely related a neural model of visual categorization is likely to involve the same visual areas as a model for object recognition. Current neural models of object recognition assume that the visual signal is processed along the ventral stream from V1 through V2 and V4 up to inferotemporal cortex (IT) (Riesenhuber & Poggio, 2000).

Somewhere in this hierarchy invariance with respect to position, lighting and view is assumed to be computed. Also the features that are used for categorization are likely to be extracted in this hierarchy. These features might be linked to the perceptual dimensions that form the basis of the categorization models discussed in this thesis. To avoid confusion, the models discussed in this thesis are all purely psychological. There is no need for a psychological theory to be reduced to a neural theory. However, RBF-networks are suggestive of a possible link between psychological theories and their neural correlates—and some work along these lines has been done already.

In object recognition, RBF-networks with their tuning curves have inspired neurophysiological work in monkeys that has found some evidence for view-tuned neurons (Logothetis, Pauls, Bülthoff, & Poggio, 1994; Logothetis, Pauls, & Poggio, 1995). Some neurons in IT do indeed show tuned responses to different views of the same object and could be the neural correlate of RBF-cells in the models. On the categorization side, ALCOVE has been used to model the behavior of neurons in IT during a shape categorization task (Beeck, Wagemans, & Vogels, 2001, 2004). This work is remarkable because it has tried to link multidimensional scaling results in humans and monkeys with IT cell activity. And indeed the activity of IT neurons seems to be related to the perceptual spaces derived from MDS. Also the GCM, that does not lend itself easily to a neural interpretation, has inspired single cell recordings in monkey IT (Sigala, Gabbiani, & Logothetis, 2002; Sigala & Logothetis, 2002). This work suggested that the representations in IT can be reshaped based on the diagnosticity of the object features—perhaps implementing the changes in similarity that are an integral part of the GCM. In general, IT seems to be a likely candidate to look for the neural basis of visual similarity. It is known that IT responds to "moderately complex object features" and that there is a continuous organization with regard to some features (Tanaka, 1996, p. 109).

Neurophysiology suggests that extraction of visual features and computation of visual similarity are processed along the ventral stream with IT at the top of the hierarchy. However, even if exemplar similarity was reflected in IT there would be more to a full neural description of categorization. Categorization behavior also involves decision making, learning and memory components—visual similarity is only one aspect of a full description of categorization. Hence, we would expect that several non-visual areas of the brain are also involved in categorization, such as prefrontal cortex (PFC), hippocampus and the basal ganglia. Especially PFC seems to be a likely candidate to implement the actual decision rule and has indeed been implicated in visual categorization (Freedman, Riesenhuber, Poggio, & Miller, 2001, 2003). While similarity and decision making were a major focus of this thesis we have not discussed the role of memory in great depth—we have just assumed that either prototypes or exemplars can be stored somehow. Without doubt memory plays a crucial role in categorization. In analogy to implicit and explicit memory systems, and based on neuropsychological evidence, Ashby and colleagues have argued that there are at least two different categorization systems (Ashby & Waldron, 1999; Ashby, Ell, & Waldron, 2003; Ashby & O'Brien, 2005). One system, implemented by the PFC, mediates explicit categorization rules. This system is highly flexible, categorization rules can be changed quickly, and at least in humans these rules can be verbalized. The other system is mediated by the basal ganglia and shows characteristics of procedural learning: Categorization is automatic and unconscious. This system is not necessarily exemplar based. However, we would expect that if categorization mechanisms are indeed exemplar based then the retrieval of exemplars is more likely to be based on implicit memory mechanisms than on explicit memory.

3. Exemplar models and natural categorization

In as much as the brain is probably adapted to the environment that we live in, the category learning mechanisms that it implements are probably also adapted to the environment. We have emphasized throughout this thesis that a useful category learning algorithm needs to be able to generalize. Furthermore, we have suggested that regularization mechanisms might be used to improve the generalization performance of exemplar theories as these mechanisms have proved to be successful in real-world applications in machine learning. However, it is incredibly difficult to specify what constitutes a natural categorization task. But without an understanding of natural categorization behavior it is impossible to judge whether the models that are suggested would actually work in the real-world—especially as most studies in the laboratory work with very reduced, artificial stimuli.

As a first step towards studying more realistic categorization processes in the laboratory we have recently collected a database of images of leaves from different trees. This database exhibits the complexity of natural stimuli with up to thirty different categories while at the same time being suitable for psychophysical investigations in the laboratory. This database of natural categories should allow us to tackle several crucial questions. First, we can get an idea about the actual generalization abilities of human subjects. Of course we assume that generalization is quite good under natural conditions but this needs to be checked. Second, we are able to address some of the parameters that mediate learning. How many exemplars are needed until a participant reliably generalizes? Some exemplars are more typical for a category, do they facilitate learning? How fast is categorization? All these questions can provide constraints on potential models. Third, as we do not have easily parametrized, artificial stimuli it is not clear what are the dimensions and features that the subjects use. While this is problematic for modeling there are ways to find out what features subjects use. We can of course use multidimensional scaling methods. We can also try to find out which aspects of the shapes of leaves are correlated with category decisions—for this we can even use some of the machine learning techniques that were discussed in this thesis. Recently, this approach has proved to be successful in face perception and should also be applicable to leaves (Graf & Wichmann, 2004; Wichmann et al., 2005; Graf et al., 2006).

4. Conclusions

While stronger connections with neuroscience and the object recognition literature may be desirable for work in categorization, this thesis is mainly concerned with the connections to machine learning. As in psychology it is common in machine learning to consider the problem of categorization in connection with similarity and generalization. In psychology dissimilarity has traditionally been modeled as a distance in a multidimensional space and the same is true for machine learning. This insight about similarity is of interest irrespective of whether one takes a prototype or an exemplar view of categorization. Both of them rely on some sort of similarity measure or distance in a multidimensional space. As exemplar theories often use Shepard's law we showed that they can be seen as kernel methods. In particular, the model underlying ALCOVE is the same as kernel logistic regression. The exemplar weights give the model too much flexibility and therefore the generalization ability of ALCOVE needed to be assessed carefully. We have suggested that regularization techniques could be employed to assure good generalization.

In the early days of machine learning, psychology and neuroscience were a major inspiration that drove research in machine learning. Today, mainstream machine learning is far removed from psychological modeling but instead tries to build systems that work for real-world problems. As this thesis has demonstrated, there are still important parallels between machine learning and psychology. Machine learning has made great progress in recent years and results from machine learning should feed back into psychology. Apart from the insights that we have presented in this thesis machine learning methods can provide standards to which human performance and model performance can be compared and they can suggest new experiments (Graf & Wichmann, 2004; Wichmann et al., 2005; Graf et al., 2006). More importantly, theoretical work in machine learning may offer a better understanding of the core problems of learning and categorization. For example, what is the role of the complexity of the category that is to be learned (Feldman, 2000; Alfonso-Reese et al., 2002; Fass & Feldman, 2003)? And under what circumstances does a category learner generalize well? After all, human categorizers and machine classifiers have to solve the same problem.

Bibliography

Aizerman, M. A., Braverman, E. M., & Rozonoer, L. I. (1964a). The probability problem of pattern recognition learning and the method of potential functions. *Automation and Remote Control*, *25*(9), 1175.

Aizerman, M. A., Braverman, E. M., & Rozonoer, L. I. (1964b). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, *25*(6), 821-837.

Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult. *Perception & Psychophysics*, *64*(4), 570-583.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409-429.

Ashby, F. G. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77-96.

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

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114-1125.

Ashby, F. G., & Gott, R. E. (1988). Decison rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*(1), 33-53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 50-71.

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

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83-9.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154-179.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363-78.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 516-556.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211-227.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*, 127-142.

Beeck, H. Op de, Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, *4*(12), 1244-52.

Beeck, H. Op de, Wagemans, J., & Vogels, R. (2004). A diverse stimulus representation underlies shape categorization by primates (abstract). *Journal of Vision*, *4*(8), 518a.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford: Oxford University Press.

Blumenthal, L. M. (1953). *Theory and applications of distance geometry.* Oxford: Clarendon Press.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications.* New York: Springer.

Bourne, L. E., & Restle, F. (1959). Mathematical theory of concept identification. *Psychological Review*, *66*, 278-296.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499-526.

Bradley, R. A. (1976). Science, statistics and paired comparisons. *Biometrics*, *32*, 213-32.

Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias-variance tradeoff. In R. Sun, N. Miyake, & C. Schunn (Eds.), *Proceedings of the 28th annual conference of the cognitive science society.*

Brown, J. S. (1965). Generalization and discrimination. In D. I. Mostofsky (Ed.), *Stimulus generalization* (p. 7-23). Stanford, CA: Stanford University Press.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences USA*, *89*(1), 60-4.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*, 413-423.

Carroll, J. D., & Wish, M. (1974). Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press.

Chater, N., & Vitanyi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*, 346-369.

Christianini, N., & Schölkopf, B. (2002). Support vector machines and kernel methods, the new generation of learning machines. *AI Magazine*, *23*(3), 31-41.

Christianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge: Cambridge University Press.

Cooke, T., Jäkel, F., Wallraven, C., & Bülthoff, H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, *45*, 484-495.

David, H. A. (1988). *The method of paired comparisons.* London: Charles Griffin and Company Ltd.

Debreu, G. (1960). Review of Luce's Individual Choice Behavior. *The American Economic Review*, *50*(1), 186-188.

Drösler, J. (1994). Color similarity represented as a metric of color space. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (p. 19-37). Berlin: Springer.

Dzhafarov, E. N., & Colonius, H. (2006). Reconstructing distances among objects from their discriminability. In H. Colonius & E. N. Dzhafarov (Eds.), *Measurement and representation of sensations.* Mahwah, NJ: Erlbaum Publ. Co.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*, 449-498.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, *38*, 467-474.

Ennis, D. M., Palen, J. J., & Mullen, K. (1988). A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology*, *32*, 449-465.

Fass, D., & Feldman, J. (2003). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15.* Cambridge, MA: MIT Press.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Franks, J. J., & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology*, *90*, 65-74.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorial representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312-316.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience*, *23*(12), 5235-46.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 234-257.

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

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *65*.

Görür, D., Jäkel, F., & Rasmussen, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the 23rd international conference on machine learning* (p. 8-15). Pittsburgh, PA.

Graf, A. B. A., & Wichmann, F. A. (2004). Insights from machine learning applied to human visual classification. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16, p. 905-912). Cambridge, MA: MIT Press.

Graf, A. B. A., Wichmann, F. A., Bülthoff, H. H., & Schölkopf, B. (2006). Classification of faces in man and machine. *Neural Computation*, *18*, 143-165.

Gregson, R. A. M. (1975). *Psychometrics of similarity.* New York: Academic Press.

Hahn, U., & Ramscar, M. (2001). *Similarity and categorization.* Oxford University Press.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (chap. 1). New York: Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York: Springer.

Hinton, G., & Roweis, S. (2003). Stochastic neighbor embedding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (p. 833-840). Cambridge, MA: MIT Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Indow, T. (1994). Metrics in color spaces: Im Kleinen und im Großen. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* New York: Springer.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 1307-1322.

Koldobsky, A., & Koenig, H. (2001). Aspects of isometric theory of Banach spaces. In W. B. Johnson & J. Lindenstrauss (Eds.), *Handbook of the geometry of Banach spaces* (p. 899-939). Amsterdam: Elsevier.

Krantz, D. H. (1967). Rational distance functions for multidimensional scaling. *Journal of Mathematical Psychology*, *4*, 226-245.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*, 4-34.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478-492.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In *Concepts core readings* (chap. 1). Boston, MA: MIT Press.

Liter, J. C., & Bülthoff, H. H. (1998). An introduction to object recognition. *Zeitschrift für Naturforschung*, *53c*, 610-621.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*(5), 401-14.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552-63.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309-32.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1961). A choice theory analysis of similarity judgements. *Psychometrika*, *26*, 151-163.

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

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215-233.

McFadden, D. L. (2003). Economic choice. In T. Persson (Ed.), *Nobel lectures, economics 1996-2000* (p. 330-364). Singapore: World Scientific.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128-48.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 294-317.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254-278.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *1*, 355-368.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775-99.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275-92.

Mostofsky, D. I. (Ed.). (1965). *Stimulus generalization.* Stanford, CA: Stanford University Press.

Murphy, G. L., & Medin, D. L. (1999). The role of theories in conceptual coherence. In E. Margolis & S. Laurence (Eds.), *Concepts core readings* (chap. 19). Boston, MA: MIT Press. (Reprinted from *Psychological Review*, 1985, *92*)

Navarro, D. J. (2002). *Representing stimulus similarity.* Unpublished doctoral dissertation, University of Adelaide.

Nilsson, N. J. (1965). *Learning machines.* New York: McGraw-Hill.

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

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*(1), 87-108.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393-418.

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition.* Hilldale, NJ: Erlbaum.

Ohl, F. W., Scheich, H., & Freeman, W. J. (2001). Change in pattern of ongoing cortical activity with auditory category learning. *Nature*, *412*, 733-736.

Orr, J., & Müller, K.-R. (Eds.). (1998). *Neural networks: Tricks of the trade* (Vol. 1524). Heidelberg: Springer.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291-303.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2005). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57-83.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472-491.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, *LV*, 899-910.

Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, *431*(7010), 768-774.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263-6.

Poggio, T., & Girosi, F. (1989). *A theory of networks for approximation and learning* (Tech. Rep. No. A. I. Memo No. 1140). Cambridge, MA: MIT AI LAB and Center for Biological Information Processing Whitaker College.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature*, *428*, 419-422.

Poggio, T., & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, *50*(5), 537-544.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis.* New York: Springer.

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

Rey, G. (1999). Concepts and stereotypes. In E. Margolis & S. Laurence (Eds.), *Concepts core readings* (chap. 12). Boston, MA: MIT Press. (Reprinted from *Cognition*, 1983, *15*)

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition. *Nature Neuroscience*, *2*(11), 1019-1025.

Riesenhuber, M., & Poggio, T. (2000, november). Models of object recognition. *Nature Neuroscience Supplement*, *3*, 1199-1204.

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

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

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

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178-210.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & R. J. McClelland (Eds.), (p. 318-362). Cambridge, MA: MIT Press.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, *44*(3), 522-536.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. In D. Helmbold & R. Williamson (Eds.), *Computational learning theory* (Vol. 2111, p. 416-426). Berlin: Springer.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*(5), 1299-1319.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1-17.

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

Shepard, R. N. (1958). Stimulus response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*(4), 242-256.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, *27*(2), 125-140.

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

Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D. I. Mostofsky (Ed.), *Stimulus generalization* (p. 94-110). Stanford, CA: Stanford University Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General*, 58-61.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classification. *Journal of Experimental Psychology*, *65*(1), 94-102.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13), 1-42.

Sigala, N., Gabbiani, F., & Logothetis, N. K. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, *14*(2), 187-98.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318-320.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411-1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3-27.

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 659-680.

Spence, K. W. (1937). The differential response in animals to stimuli varying in a single dimension. *Psychological Review*, *44*, 430-444.

Stevens, S. S. (1965). On the uses of poikilitic functions. In D. I. Mostofsky (Ed.), *Stimulus generalization* (p. 24-29). Stanford, CA: Stanford University Press.

Strang, G. (1988). *Linear algebra and its applications* (3 ed.). Orlando, FL: Harcourt Brace & Company.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Reviews Neuroscience*, *19*, 109-139.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based objecti recognition in man, monkey and machine. *Cognition*, *67*, 1-20.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-640.

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

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, *17*, 401-419.

Townsend, J. T., Solomon, B., & Smith, J. S. (2001). The perfect Gestalt: Infinite dimensional Riemannian face space and other aspects of face perception. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives of facial cognition: Contexts and challenges* (p. 39-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Townsend, J. T., & Thomas, R. D. (1993). On the need for a general quantitative theory of pattern similarity. In S. C. Masin (Ed.), *Foundations of perceptual theory.* Amsterdam: Elsevier.

Train, K. E. (2003). *Discrete choice methods with simulation.* Cambridge: Cambridge University Press.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*(4), 281-299.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123-154.

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*, 572-596.

Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.

Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition*, *32*(3), 379-89.

Wichmann, F. A., Graf, A. B. A., Simoncelli, E. P., Bülthoff, H. H., & Schölkopf, B. (2005). Machine learning applied to perception: decision-images for classification. In Saul L. K., Weiss Y., & L. Bottou (Eds.), *Advances in neural* *information processing systems 17* (p. 1489-1496). Cambridge, MA: MIT Press.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293-1313.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19-21.

Zhang, J. (2006). Referential duality and representational duality in the scaling of multidimensional and infinite-dimensional stimulus space. In H. Colonius & E. N. Dzhafarov (Eds.), *Measurement and representation of sensations* (p. 131-156). Mahwah, NJ: Lawrence Erlbaum Associates.