Journal of Machine Learning Research 9 (2008) 1049-1082 Submitted 5/07; Revised 12/07; Published 6/08

Learning Similarity with Operator-valued Large-margin Classifiers

Andreas Maurer ANDREASMAURER@COMPUSERVE.COM
Adalbertstr. 55
D-80799 München
Germany

Editor: Peter Bartlett

Abstract

A method is introduced to learn and represent similarity with linear operators in kernel induced Hilbert spaces. Transferring error bounds for vector valued large-margin classifiers to the setting of Hilbert-Schmidt operators leads to dimension free bounds on a risk functional for linear representations and motivates a regularized objective functional. Minimization of this objective is effected by a simple technique of stochastic gradient descent. The resulting representations are tested on transfer problems in image processing, involving plane and spatial geometric invariants, handwritten characters and face recognition.

Keywords: learning similarity, similarity, transfer learning

1. Introduction

Similarity seems fundamental to perception and reasoning. The precise meaning of the word ”similarity” however is elusive and a corresponding Google search reveals a plethora of definitions ranging from ”the property of being similar” to the analyses of Wittgenstein, Russel or Carnap. This is not surprising: Pairs of triangles may or may not be similar, two poems may induce similar or radically different moods in similar people, pairs of wines, bird-calls, weather patterns, approaches to cognitive science, movies and definitions of similarity themselves may each be similar or not similar or similar to a varying degree. It is the very universality of the concept which explains the lack of a universal definition, beyond the structural feature that similarity is a property possessed by pairs of objects and the vague feeling that it is somehow related to geometric proximity.

What cannot be defined may still be learned. Even if we cannot explain the meaning of a concept, as long as it has observable manifestations we can hope to infer models to predict future observations. These models will generally be domain-dependent and not in the form of definitions, but rather vectors of synaptic efficacies, transformation coefficients or other constructions which often defy verbalization, they are more akin to feeling than to rationality and they will be judged by their predictive power rather than by logical clarity and comprehensibility.

This paper introduces a technique to learn and to represent similarity, an analysis of generalization performance, an algorithmic realization and some experimental results.

1.1 A General Framework

Assumption 1: There is a measurable space $\mathcal{X}$ and a probability measure ρ on $\mathcal{X}^2 \times \{-1,1\}$, the *pair oracle*.

The interpretation is as follows: $\mathcal{X}$ is the input space containing the objects in question. Whenever we call the oracle it will return the triplet $(x,x',r) \in \mathcal{X}^2 \times \{-1,1\}$ with probability $\rho(x,x',r)$. If the oracle returns $(x,x',1)$ it asserts that x and x' are *similar*, if it returns $(x,x',-1)$ it asserts that x and x' are *dissimilar*. The assumption that similarity is a binary property, which can only have the values of true or false, is a simplification which can in principle be removed (see Section 4.1 below). Beyond this restriction arbitrary definitions of similarity can be substituted, we do not require "obvious properties" such as $\rho(x,x',r) = \rho(x',x,r)$ or $\rho(x,x,-1) = 0$. We will however use an intuitive property of similarity, a kinship to closeness or geometric proximity, in the choice of our hypothesis spaces below.

Assumption 2: $\mathcal{X} \subset H$, where H is a real separable finite- or infinite-dimensional Hilbert space, and $diam(\mathcal{X}) \leq 1$.

Either the inputs are already members of some Euclidean space (vectors of neural activations, pixel vectors or vectors of feature-values), or we identify them with their images under some feature map, which may be realized by some fixed pre-processor such as a fixed weight neural network or by a positive definite kernel on $\mathcal{X}$. In the more detailed description of the proposed algorithm and experimental results we will be more explicit about these feature maps. The bound on the diameter of the input space is a convenience for the statement of our theoretical results.

Assumption 3: There is a training sample

$$S = \left(\left(x_1,x_1',r_1\right),...,\left(x_m,x_m',r_m\right)\right) \in \left(\mathcal{X}^2 \times \{-1,1\}\right)^m,$$

generated in m independent, identical trials of ρ, that is, $S \sim \rho^m$.

The assumptions of independence and stationarity are crucial and very strong: The oracle is without memory of our previous calls and not affected by the passage of time. It will not deliberately help nor mislead the learner. The training sample contains all the information available to the learner who wants to find a rule to predict the oracles behavior.

Definition 1: A pair hypothesis is a function $f : \mathcal{X}^2 \to \{-1,0,1\}$. Its risk is

$$R(f) = \Pr_{(x,x',r)\sim\rho}\left\{f\left(x,x'\right) \neq r\right\}.$$

The pair hypothesis attempts to predict the similarity value of a pair (x,x') and its risk is its error probability as measured by the pair oracle. We allowed the value 0 to account for the possibility that f may refuse to make a decision. Any such refusal is counted as an error by the risk functional.

1.2 Risk Bounds and Regularized Objectives

So far we have only used the structural condition that similarity is a property possessed by pairs of objects. In the choice of the hypothesis space we will follow the intuition that similarity is related to closeness or geometrical proximity. We will consider hypotheses from the set

$$\mathcal{H} = \left\{f_T : \left(x,x'\right) \mapsto sgn\left(1 - \left\|Tx - Tx'\right\|\right) : T \in \mathcal{L}_0(H)\right\},$$

where $\mathcal{L}_0(H)$ is the set of linear operators of finite rank on H. A transformation $T \in \mathcal{L}_0(H)$ thus defines a hypothesis f_T which regards a pair of inputs as similar if the distance between their respective images under T is smaller than one, and as dissimilar if this distance is larger than one. Fixing the threshold to one causes no loss of generality, because any other positive threshold could be absorbed as a factor of the transformation.

This choice of the hypothesis space combines the geometric connotation of similarity with the simplicity of linear representations. The transformations which parametrize our hypothesis space are in some ways more interesting and useful than the hypotheses themselves. A choice of $T \in \mathcal{L}_0(H)$ also implies a choice of the Mahalanobis distance $d^2(x,x') = \langle T^*T(x-x'), x-x'\rangle$ and the positive semidefinite kernel $\kappa(x,x') = \langle T^*Tx, x'\rangle$.

The risk of the hypothesis f_T induces a risk functional on $\mathcal{L}_0(H)$

$$R(T) = R(f_T) = \Pr_{(x,x',r)\sim\rho}\left\{r\left(1 - \left\|Tx - Tx'\right\|^2\right) \leq 0\right\}.$$

Since we can write $\|Tx - Tx'\|^2 = \langle T^*T(x-x'), x-x'\rangle = \langle T^*T, Q_{x-x'}\rangle_2$, where $Q_{x-x'}$ and $\langle .,.\rangle_2$ are respectively the outer product operator and the Hilbert-Schmidt inner product (see Section 2.1), the last expression is reminiscent of the risk of a classifier defined by a linear function thresholded at 1. This provides the intuition underlying the proposed technique: We will look for a linear large-margin classifier whose weight vector is the positive operator T^*T.

Let $\psi : \mathbb{R} \to \mathbb{R}$, $\psi \geq 1_{(-\infty,0]}$ with Lipschitz constant L. Given our training sample $S = ((x_1, x_1', r_1), \ldots, (x_m, x_m', r_m))$ we define the empirical risk estimate

$$\hat{R}_\psi(T,S) = \frac{1}{m}\sum_{i=1}^{m}\psi\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right). \tag{1}$$

We then have the following theorem, a proof of which will be given in Section 3.

Theorem 1 $\forall\delta > 0$, *with probability greater* $1-\delta$ *in a sample* $S \sim \rho^m$
$\forall T \in \mathcal{L}_0(H)$ *with* $\|T^*T\|_2 \geq 1$

$$R(T) \leq \hat{R}_\psi(T,S) + \frac{4L\ \|T^*T\|_2 + \sqrt{(1/2)\ln\left(2\|T^*T\|_2/\delta\right)}}{\sqrt{m}}.$$

where $\|A\|_2 = Tr(A^*A)^{1/2}$ *is the Hilbert-Schmidt- or Frobenius- norm of* A.

The theorem gives a high-probability-bound on the true risk valid for all transformations in terms of the empirical risk estimate and a complexity term. Because of the uniform nature of the bound, a principled approach could search for $T \in \mathcal{L}_0(H)$ to minimize the right side of the bound. This is just what we propose to do, with two practical modifications:

- The term $\sqrt{(1/2)\ln\left(2\|T^*T\|_2/\delta\right)}$ is neglected, the one linear in $\|T^*T\|_2$ being regarded as the dominant contribution.
- The factor $4L$ is replaced by an adjustable regularization parameter $\lambda > 0$. This allows to compensate the fact that the difficult estimates in such generalization bounds overestimate the estimation error.

As a Lipschitz function ψ we use h_γ, the hinge-loss with margin γ, which has the value $1 - t/\gamma$ for $t < \gamma$, and the value 0 otherwise. This leads to the regularized objective function

$$\Lambda_{h_\gamma,\lambda}(T) := \frac{1}{m}\sum_{i=1}^{m}h_\gamma\left(r_i\left(1 - \left\|T\left(x_i - x_i'\right)\right\|^2\right)\right) + \frac{\lambda\|T^*T\|_2}{\sqrt{m}},$$

which is convex in T^*T. Details related to the minimization of $\Lambda_{\gamma,\lambda}$ are given in Section 3.3. It follows from the nature of the objective function that the minimizing operator will have rank $\leq m$.

1.3 The Multi-category Problem and Learning-to-learn

In many classification problems occurring in a complex environment a learner cannot hope to obtain examples for all potential categories. An example is furnished by the recognition of human faces, since nobody can be expected to ever see all the faces which possibly need to be distinguished in the future and the total number of categories is itself uncertain.

Human learning appears to cope with these empirically deficient problems: In addition to the recognition of the already known categories human learners develop meta-techniques to improve the learning and recognition of future categories. As a child learns to recognize parents, family, friends, neighbors, classmates and teachers it learns to memorize, recognize and distinguish faces in general, an ability which leads to a reliable recall often already on the basis of a single training image. The earlier learning process leading to an improvement of future learning performance is often referred to as *learning-to-learn* or *meta-learning*.

The practical utility of such mechanisms for machine-learning is obvious, and theoretically well founded models of learning-to-learn may also be interesting from the point of view of cognitive science or psychology (see Robins, 1998; Thrun, 1998; Baxter, 1998 for surveys, theoretical and experimental contributions).

Here we exploit the fact that partial empirical knowledge of some of the categories of a domain implies partial empirical knowledge of an underlying principle of similarity. By a very crude operational definition similarity is a property of two phenomena which makes them belong to the same category of some domain, and dissimilarity is the negation of similarity. Following this idea we can define a pair oracle which regards pairs of inputs as similar if and only if they come with the same label, use this oracle to generate a large number of examples and train our algorithm to obtain a similarity rule, together with a representing operator T.

Subsequently, once a novel category is represented by a first example, any new phenomenon can be classified as belonging to the given category if and only if it is similar to the representing example. Let us call this decision rule the *elementary verifier* generated from the example. It follows from our analysis, that we can be confident that a single randomly chosen example for a future (possibly previously unseen) category will give an elementary verifier with expected error bounded by the bound in Theorem 1. These constructions and some related questions will be discussed in Section 5.

A related concept is *transfer*, where a representation trained from the data-set of a training-task is applied to facilitate the learning of another, presumably related, target-task. Corresponding experiments were carried out on a number of problems in image recognition, such as the recognition of handwritten characters, rotation and scale invariant character recognition and the recognition of human faces. In all these cases the representations generated from the training-tasks yielded a considerable performance improvement for single-sample nearest neighbor classifiers on the target-tasks.

This paper is organized as follows: Section 2 gives notation and theoretical background, Section 3 introduces operator valued large-margin classifiers and a corresponding algorithm to train representations, Section 4 derives an alternative algorithm related to PCA, Section 5 discusses the application of our method to meta-learning and transfer, Section 6 describes experimental results and Section 7 gives a brief review of some related approaches in the literature.

A precursor of this paper appeared in the NIPS’06 workshop on ”Learning to Compare Examples” (Maurer, 2006c).

2. Notation, Definitions and Preliminary Results

In this section we introduce some notation and necessary theoretical background. For the readers convenience there is also an appendix with a tabular summary of most of the notation used in this paper.

2.1 Hilbert Space and Hilbert Schmidt Operators

The letter H denotes a finite- or infinite-dimensional real, separable Hilbert space, with inner product $\langle .,. \rangle$ and norm $\|.\|$. Inner products and norms on other spaces will be identified with subscripts. If $\left(\mathcal{S}, \|.\|_{\mathcal{S}}\right)$ is a generic normed space and $E \subseteq \mathcal{S}$, we will use the notation

$$\|E\|_{\mathcal{S}} = \sup_{x \in E} \|x\|_{\mathcal{S}} .$$

If H' is another Hilbert space, with inner product $\langle .,. \rangle'$ and norm $\|.\|'$, then $\mathcal{L}_\infty(H,H')$ denotes the Banach space of linear transformations $T : H \to H'$ such that

$$\|T\|_\infty = \sup_{x \in H, \|x\| \leq 1} \|Tx\|' < \infty.$$

For $T \in \mathcal{L}_\infty(H,H')$, $Ker(T)$ is the subspace $\{x : Tx = 0\}$ and T^* is the unique member of $\mathcal{L}_\infty(H',H)$ satisfying $\langle Tx, y \rangle' = \langle x, T^* y \rangle$, $\forall x \in H, y \in H'$. If $S \subset H$ then $S^\perp$ is the subspace $\{x : \langle x, y \rangle = 0, \forall y \in S\}$. A transformation $U \in \mathcal{L}_\infty(H,H')$ is called a partial isometry if $\|Ux\| = \|x\|$, $\forall x \in Ker(U)^\perp$.

We write $\mathcal{L}_\infty(H) = \mathcal{L}_\infty(H,H)$. An operator $T \in \mathcal{L}_\infty(H)$ is called symmetric if $T = T^*$ and positive if it is symmetric and $\langle Tx, x \rangle \geq 0$, $\forall x \in H$. We use $\mathcal{L}_\infty^*(H)$ and $\mathcal{L}_\infty^+(H)$ to denote the sets of symmetric and positive members of $\mathcal{L}_\infty(H)$ respectively. For every $T \in \mathcal{L}_\infty^+(H)$ there is some unique $T^{1/2} \in \mathcal{L}_\infty^+(H)$ with $T = T^{1/2} T^{1/2}$. Evidently for every $T \in \mathcal{L}_\infty(H,H')$ we have $T^* T \in \mathcal{L}_\infty^+(H)$ and we denote with $|T|$ the positive operator $(T^*T)^{1/2}$. For every $T \in \mathcal{L}_\infty(H,H')$ there is a polar decomposition $T = U|T|$ for a unique partial isometry $U \in \mathcal{L}_\infty(H,H')$.

For $\mathcal{V} \subseteq \mathcal{L}_\infty(H)$ we use the notation $\mathcal{V}^* \mathcal{V} = \left\{T^* T : T \in \mathcal{V}\right\}$. The set of finite rank operators on H is denoted by $\mathcal{L}_0(H)$. Any orthonormal basis establishes a one-to-one correspondence between $\mathcal{L}_0(H)$ and $\bigcup_{n=1}^\infty \{T : H \to \mathbb{R}^n : T \text{ linear}\}$.

With $\mathcal{L}_2(H)$ we denote the real vector space of operators $T \in \mathcal{L}_\infty(H)$ satisfying $\sum_{i=1}^\infty \|Te_i\|^2 \leq \infty$ for every orthonormal basis $(e_i)_{i=1}^\infty$ of H. The members of $\mathcal{L}_2(H)$ are compact and called *Hilbert Schmidt operators*. For $S, T \in \mathcal{L}_2(H)$ and an orthonormal basis (e_i) the series $\sum_i \langle Se_i, Te_i \rangle$ is absolutely summable and independent of the chosen basis. The number

$$\langle S, T \rangle_2 = \sum_i \langle Se_i, Te_i \rangle$$

defines an inner product on $\mathcal{L}_2(H)$, making it a Hilbert space. We denote the corresponding norm with $\|.\|_2$ (see Reed and Simon, 1980 for background on functional analysis). $\mathcal{L}_2^*(H)$ and $\mathcal{L}_2^+(H)$ denote the sets of symmetric and positive members of $\mathcal{L}_2(H)$ respectively. For every member of $\mathcal{L}_2^*(H)$ there is a complete orthonormal basis of eigenvectors, and for $T \in \mathcal{L}_2^*(H)$ the norm $\|T\|_2$ is just the ℓ_2-norm of its sequence of eigenvalues. In the finite dimensional case the norm $\|T\|_2$ is the Frobenius norm of the matrix of T in an orthonormal representation.

The set of d-dimensional, orthogonal projections in H is denoted with $\mathcal{P}_d$. It is easy to verify that $\mathcal{P}_d \subset \mathcal{L}_2^+(H)$ and if $P \in \mathcal{P}_d$ then $\|P\|_2 = \sqrt{d}$ and $P^2 = P$.

Definition 2 *For $x \in H$ define an operator Q_x on H by $Q_x z = \langle z, x \rangle x$, $\forall z \in H$.*

The map $x \to Q_x$ is an embedding of H in $\mathcal{L}_2^+(H)$ which is homogeneous of degree two (i.e., $Q_{\lambda x} = \lambda^2 Q_x$, $\lambda \in \mathbb{R}$, $x \in H$). In matrix terminology Q_x is the 'outer product' of x with itself. The following simple lemma is crucial for our proofs (see also Maurer, 2006b).

Lemma 3 *Let $x, y \in H$ and $T \in \mathcal{L}_2(H)$. Then*
(i) $Q_x \in \mathcal{L}_2^+(H)$ and $\|Q_x\|_2 = \|x\|^2$.
(ii) $\langle Q_x, Q_y \rangle_2 = \langle x, y \rangle^2$.
(iii) $\langle T, Q_x \rangle_2 = \langle Tx, x \rangle$.
(iv) $\langle T^*T, Q_x \rangle_2 = \|Tx\|^2$.
(v) For $\alpha \in \mathbb{R}$, $Q_{\alpha x} = \alpha^2 Q_x$.
(vi) For $P \in \mathcal{P}_d$ we have $\|PTP\|_2 \leq \|T\|_2$.

Proof For $x = 0$ all assertions are trivial. Otherwise extend $x/\|x\|$ to an orthonormal basis of H and use this basis in the definition of $\langle T, Q_x \rangle_2$ to obtain $\langle T, Q_x \rangle_2 = \langle Tx/\|x\|, Q_x x/\|x\| \rangle = \|x\|^{-2} \langle Tx, \langle x, x \rangle x \rangle = \langle Tx, x \rangle$, which is (iii). (ii),(iv) and the second half of (i) follow immediately, the first part of (i) then follows from (iii) with $T = Q_x$. (v) is trivial. To prove (vi) complete a basis $\{e_i\}_{i=1,..,d}$ for the range of P to a basis e_i for H to get $\|PTP\|_2^2 = \sum_{ij} \langle PTPe_i, e_j \rangle^2 = \sum_{i,j \leq d} \langle Te_i, e_j \rangle^2 \leq \sum_{ij} \langle Te_i, e_j \rangle^2 = \|T\|_2^2$. ■

2.2 Rademacher Complexities

To derive the uniform laws of large numbers we need for Theorem 1 we will use Rademacher averages as complexity measures for function classes:

Definition 4 *Let $\mathcal{F}$ be a real-valued function class on a space $\mathcal{X}$. Let $\{\sigma_i : i \in \{1, \ldots, m\}\}$ be a collection of independent random variables, distributed uniformly in $\{-1, 1\}$. The empirical Rademacher complexity of $\mathcal{F}$ is the function $\hat{\mathcal{R}}_m(\mathcal{F})$ defined on $\mathcal{X}^m$ by*

$$\hat{\mathcal{R}}_m(\mathcal{F})(\mathbf{x}) = \mathbb{E}_\sigma \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^m \sigma_i f(x_i) \right].$$

If $\mathbf{X} = (X_i)_{i=1}^m$ is a vector of $\mathcal{X}$-valued independent random variables then the expected Rademacher complexity of $\mathcal{F}$ is

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{\mathbf{X}} \left[\hat{\mathcal{R}}_m(\mathcal{F})(\mathbf{X}) \right].$$

Theorem 5 *Let $\mathcal{F}$ be a $[0,1]$-valued function class on a space $\mathcal{X}$, and $\mathbf{X} = (X_i)_{i=1}^m$ a vector of $\mathcal{X}$-valued independent, identically distributed random variables. Fix $\delta > 0$.*

With probability greater than $1 - \delta$ we have for all $f \in \mathcal{F}$

$$E[f(X_1)] \leq \frac{1}{m} \sum_{i=1}^m f(X_i) + \mathcal{R}_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

We also have with probability greater than $1-\delta$ *for all* $f \in \mathcal{F}$*, that*

$$E\left[f\left(X_1\right)\right] \leq \frac{1}{m}\sum_{i=1}^{m} f\left(X_i\right) + \hat{\mathcal{R}}_m\left(\mathcal{F}\right)\left(\mathbf{X}\right) + \sqrt{\frac{9\ln\left(2/\delta\right)}{2m}}.$$

For a proof see Bartlett and Mendelson (2002) or Maurer (2006b). We will also use the following result from Bartlett et al. (2005):

Theorem 6 *Let* $\mathcal{F}$ *be a class of real-valued functions on a space* $\mathcal{X}$ *and suppose that* $\psi : \mathbb{R} \to \mathbb{R}$ *has Lipschitz constant L. Let* $\psi \circ \mathcal{F} = \{\psi \circ f : f \in \mathcal{F}\}$*. Then* $\hat{\mathcal{R}}_m\left(\psi \circ \mathcal{F}\right) \leq L\, \hat{\mathcal{R}}_m\left(\mathcal{F}\right)$*.*

2.3 Bounds for Vector-valued Processes

In this section we review some bounds for vector valued processes and linear classifiers. The bounds are not at all original (they are taken from Koltchinskii and Panchenko, 2002; Bartlett and Mendelson, 2002; Shawe-Taylor and Christianini, 2003), nor are they necessarily the tightest possible, because they are uniformly valid on the chosen function classes (compare with Bartlett et al., 2005). Because the proofs are easy we provide them for the readers convenience.

Using Lemma 3 all these results can be easily transferred from the Hilbert space H to the Hilbert space $\mathcal{L}_2\left(H\right)$. This simple step, together the geometrical interpretation implied by Lemma 3 (iv), is the principal theoretical contribution of this paper.

Lemma 7 *Let* $V \subset H$ *and* $\mathcal{F} = \{x \in H \mapsto \langle x, v\rangle : v \in V\}$*. Then for any* $\mathbf{x} = \left(x_1, \dots, x_m\right) \in H^m$

$$\hat{\mathcal{R}}_m\left(\mathcal{F}\right)\left(\mathbf{x}\right) \leq \frac{2\left\|V\right\|}{m}\left(\sum_{i=1}^{m}\left\|x_i\right\|^2\right)^{1/2}.$$

Proof From Schwartz' and Jensen's inequality and linearity we obtain

$$\begin{aligned}\hat{\mathcal{R}}_m\left(\mathcal{F}\right)\left(\mathbf{x}\right) &= \mathbb{E}_\sigma\left[\sup_{v\in V}\frac{2}{m}\sum_{i=1}^{m}\sigma_i\left\langle x_i, v\right\rangle\right] = \mathbb{E}_\sigma\left[\sup_{v\in V}\frac{2}{m}\left\langle\sum_{i=1}^{m}\sigma_i x_i, v\right\rangle\right] \\ &\leq \frac{2\left\|V\right\|}{m}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|\right] \leq \frac{2\left\|V\right\|}{m}\left(\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right]\right)^{1/2},\end{aligned}$$

but by the properties of the σ_i we have $\mathbb{E}_\sigma\left[\sigma_i\sigma_j\right] = \delta_{ij}$, so we get

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^{m}\sigma_i x_i\right\|^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}_\sigma\left[\sigma_i\sigma_j\right]\left\langle x_i, x_j\right\rangle = \sum_{i=1}^{m}\left\|x_i\right\|^2.$$

■

Theorem 8 *Let (X,r) be a random variable with values in $H \times \{-1,1\}$ and let $(\mathbf{X},\mathbf{r}) = ((X_1,r_1), \ldots, (X_m,r_m))$ be a vector of m iid copies of (X,r). Let $\psi : \mathbb{R} \to \mathbb{R}$, $\psi \geq 1_{(-\infty,0]}$ with Lipschitz constant L and $V \subset H$. For $v \in V$ denote $err(v) = \Pr\{sign(1 - \langle X,v\rangle) \neq r\}$ and*

$$err\hat{r}_{\psi}(v,(\mathbf{X},\mathbf{r})) = \frac{1}{m}\sum_{i=1}^{m}\psi(r_i(1-\langle X_i,v\rangle)).$$

Let $\delta > 0$. Then with probability greater than $1-\delta$ we have for all $v \in V$

$$err(v) \leq err\hat{r}_{\psi}(v,(\mathbf{X},\mathbf{r})) + \frac{2L\ \|V\|}{m}\left(\sum_{i=1}^{m}\|X_i\|^2\right)^{1/2} + \sqrt{\frac{9\ln(2/\delta)}{2m}}.$$

If $\|X\| \leq 1$ a.s. then with probability greater than $1-\delta$ we have for all $v \in V$ that

$$err(v) \leq err\hat{r}_{\psi}(v,(\mathbf{X},\mathbf{r})) + \frac{2L\ \|V\|}{\sqrt{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

Proof If we can prove the Theorem for $\psi : \mathbb{R} \to [0,1]$, then it will follow for general ψ, because it will also be true for $\min\{\psi,1\}$ which has Lipschitz constant bounded by L and $\mathrm{er}\hat{\mathrm{r}}_{\min\{\psi,1\}}(v,(\mathbf{X},\mathbf{r})) \leq \mathrm{er}\hat{\mathrm{r}}_{\psi}(v,(\mathbf{X},\mathbf{r}))$. We can thus assume $\psi : \mathbb{R} \to [0,1]$ and since $\psi \geq 1_{(-\infty,0]}$ we have $R(v) = \mathbb{E}\left[1_{(-\infty,0]}(r(1-\langle X,v\rangle))\right] \leq \mathbb{E}[\psi(r(1-\langle X,v\rangle))]$. In view of Theorem 5 it therefore suffices to prove that

$$\hat{\mathcal{R}}_m(\psi \circ \mathcal{F})(\mathbf{X}) \leq \frac{2L\ \|V\|}{m}\left(\sum_{i=1}^{m}\|X_i\|^2\right)^{1/2}, \tag{2}$$

where $\mathcal{F}$ is the function class

$$\mathcal{F} = \{(x,r) \in H \times \{-1,1\} \mapsto r(1-\langle x,v\rangle) : v \in V\}.$$

By Theorem 6 we have $\hat{\mathcal{R}}_m(\psi \circ \mathcal{F}) \leq L\ \hat{\mathcal{R}}_m(\mathcal{F})$ and one verifies easily that $\hat{\mathcal{R}}_m(\mathcal{F}) = \hat{\mathcal{R}}_m(x \mapsto \langle x,v\rangle : v \in V)$, so (2) follows from Lemma 7. ■

To derive our bounds for hyperbolic PCA in Section 4 we need the following lemma. A similar statement can be found in Shawe-Taylor and Christianini (2003).

Lemma 9 *Let $V,W \subset H$ be and suppose that $X_1,\ldots,X_m$ are independent, identically distributed, zero-mean random variables with values in W. Then for ε and m such that $\|W\|\,\|V\| < \sqrt{m}\varepsilon$ we have*

$$\Pr\left\{\sup_{v\in V}\left|\frac{1}{m}\sum_{i=1}^{m}\langle v,X_i\rangle\right| > \varepsilon\right\} \leq \exp\left(\frac{-\left(\sqrt{m}\varepsilon - \|V\|\,\|W\|\right)^2}{2\left|\langle V,W\rangle\right|^2}\right).$$

Proof Consider the average $\bar{\mathbf{X}} = (1/m)\sum_1^m X_i$. With Jensen's inequality and using independence we obtain

$$\left(\mathbb{E}\left[\|\bar{\mathbf{X}}\|\right]\right)^2 \leq \mathbb{E}\left[\|\bar{\mathbf{X}}\|^2\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\|X_i\|^2\right] \leq \|W\|^2/m.$$

Now let $f : W^m \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \sup_{v \in V} |(1/m) \sum_1^m \langle v, x_i \rangle|$. We have to bound the probability that $f > \varepsilon$. By Schwartz' inequality and the above bound we have

$$\mathbb{E}[f(\mathbf{X})] = \mathbb{E}\left[\sup_{v \in V} \left|\left\langle v, \bar{\mathbf{X}} \right\rangle\right|\right] \leq \|V\| \, \mathbb{E}\left[\left\|\bar{\mathbf{X}}\right\|\right] \leq \left(1/\sqrt{m}\right) \|V\| \, \|W\| \,. \tag{3}$$

Let $\mathbf{x} \in W^m$ be arbitrary and $\mathbf{x}' \in W^m$ be obtained by modifying a coordinate x_k of $\mathbf{x}$ to be an arbitrary $x_k' \in W$. Then

$$\left| f(\mathbf{x}) - f\left(\mathbf{x}'\right) \right| \leq \frac{1}{m} \sup_{v \in V} \left| \langle v, x_k \rangle - \left\langle v, x_k' \right\rangle \right| \leq \frac{2}{m} \left| \langle V, W \rangle \right| .$$

By (3) and the bounded-difference inequality (see McDiarmid, 1998) we obtain for $t > 0$

$$\Pr\left\{ f(\mathbf{X}) > \frac{\|V\| \, \|W\|}{\sqrt{m}} + t \right\} \leq \Pr\left\{ f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] > t \right\} \leq \exp\left(\frac{-mt^2}{2 \left| \langle V, W \rangle \right|^2} \right) .$$

The conclusion follows from setting $t = \varepsilon - (1/\sqrt{m}) \|V\| \, \|W\|$ ■

We will also use the following model-selection lemma taken from (Anthony and Bartlett, 1999, Lemma 15.5):

Lemma 10 *Suppose*

$$\{F(\alpha_1, \alpha_2, \delta) : 0 < \alpha_1, \alpha_2, \delta \leq 1\}$$

is a set of events such that:

(i) For all $0 < \alpha \leq 1$ *and* $0 < \delta \leq 1$,

$$\Pr\{F(\alpha, \alpha, \delta)\} \leq \delta.$$

(ii) For all $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$*and* $0 < \delta_1 \leq \delta \leq 1$,

$$F(\alpha_1, \alpha_2, \delta_1) \subseteq F(\alpha, \alpha, \delta) .$$

Then for $0 < a, \delta < 1$,

$$\Pr\left(\bigcup_{\alpha \in (0,1]} F(\alpha a, \alpha, \delta\alpha(1-a)) \right) \leq \delta.$$

3. Operator Valued Large Margin Classifiers

In this section we derive our algorithm. We give a proof of Theorem 1, then discuss the derivation of an objective functional and finally some issues related to its minimization.

3.1 Generalization Bounds

We first give a version of Theorem 1 for a fixed hypothesis space given by Hilbert-Schmidt operators of uniformly bounded norm and then derive from it a regularized version applying to all Hilbert-Schmidt operators.

Recall that the pair oracle is a probability measure ρ on $\mathcal{X}^2 \times \{-1,1\}$, the set of labeled pairs (sometimes also called equivalence constraints; Bar-Hillel et al., 2005), and that we assumed the input space $\mathcal{X}$ to be embedded in the Hilbert space H such that $diam(\mathcal{X}) \leq 1$. We will apply Theorem 8 to the Hilbert space $\mathcal{L}_2(H)$ instead of H and replace the random variable (X,r) of Theorem 8 by the random variable $(Q_{x-x'},r)$ with values in $\mathcal{L}_2(H) \times \{-1,1\}$, where (x,x',r) are distributed according to the pair oracle ρ. The training sample $S = ((x_1,x_1',r_1),\dots,(x_m,x_m',r_m))$ corresponds to m independent realizations $\left(\left(Q_{x_1-x_1'},r_1\right),\dots,\left(Q_{x_m-x_m'},r_m\right)\right)$ of this random variable. Since $diam(\mathcal{X}) \leq 1$ we have $\|Q_{x-x'}\|_2 = \|x-x'\|^2 \leq 1$ a.s. by virtue of Lemma 3 (i).

Recall the definition of the risk associated with a transformation T. Using Lemma 3 (iv) we have

$$\begin{aligned} R(T) &= \Pr\left\{r\left(1-\left\|Tx-Tx'\right\|^2\right) \leq 0\right\} = \Pr\{r(1-\langle Q_{x-x'},T^*T\rangle_2) \leq 0\} \\ &= \operatorname{err}(T^*T). \end{aligned}$$

Here err(T^*T) (see Theorem 8) is the expected error of the linear classifier defined by the vector T^*T in $\mathcal{L}_2(H)$ thresholded at the value 1 and applied to a random labeled data point $(Q_{x-x'},r) \in \mathcal{L}_2(H) \times \{-1,1\}$. Also note that the empirical estimator (1) can be rewritten

$$\hat{R}_\psi(T,S) = \frac{1}{m}\sum_{i=1}^m \psi\left(r_i\left(1-\left\langle Q_{x_i-x_i'},T^*T\right\rangle_2\right)\right).$$

With corresponding substitutions Theorem 8 becomes

Theorem 11 *Let $\psi : \mathbb{R} \to \mathbb{R}$, $\psi \geq 1_{(-\infty,0]}$ with Lipschitz constant L and $\mathcal{V} \subset \mathcal{L}_2(H)$. Let $\delta > 0$.*

(i) With probability greater than $1-\delta$ we have for all $T \in \mathcal{L}_\infty(H)$ such that $T^*T \in \mathcal{V}$

$$R(T) \leq \hat{R}_\psi(T,S) + \frac{2L\ \left\|\mathcal{V}\right\|_2}{m}\left(\sum_{i=1}^m \left\|X_i - X_i'\right\|^4\right)^{1/2} + \sqrt{\frac{9\ln(2/\delta)}{2m}}.$$

(ii) With probability greater than $1-\delta$ we have for all $T \in \mathcal{L}_\infty(H)$ such that $T^*T \in \mathcal{V}$

$$R(T) \leq \hat{R}_\psi(T,S) + \frac{2L\ \left\|\mathcal{V}\right\|_2}{\sqrt{m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

Typically $\mathcal{V}$ would be a ball of some fixed radius, say c, about the origin in $\mathcal{L}_2(H)$, so that $\left\|\mathcal{V}\right\|_2 = c$. Our application of the vector valued generalization Theorem 8 introduced a certain looseness: Theorem 8 gives bounds on err(A) for all $A \in \mathcal{V}$, while we only require the bounds for those $A \in \mathcal{V}$ which are of the form $A = T^*T$, that is we are only using the linear functionals in $\mathcal{V} \cap \mathcal{L}_2^+(H)$. This raises the question if we might not get much better bounds for $\mathcal{V} \cap \mathcal{L}_2^+(H)$ than for $\mathcal{V}$. The following proposition shows that this will not work using Rademacher averages.

Proposition 12 *Let $\mathcal{F}$ and $\mathcal{F}^+$ be the function classes on $\mathcal{L}_2(H)$ given by*

$$\begin{aligned}\mathcal{F} &= \left\{B \in \mathcal{L}_2(H) \mapsto \langle B, A\rangle_2 : \|A\|_2 \leq 1\right\},\\ \mathcal{F}^+ &= \left\{B \in \mathcal{L}_2(H) \mapsto \langle B, A\rangle_2 : \|A\|_2 \leq 1,\ A \in \mathcal{L}_2^+(H)\right\}.\end{aligned}$$

Then $\hat{\mathcal{R}}_m(\mathcal{F}^+) \leq \hat{\mathcal{R}}_m(\mathcal{F}) \leq 4\hat{\mathcal{R}}_m(\mathcal{F}^+)$.

Proof The first inequality is obvious since $\mathcal{F}^+ \subset \mathcal{F}$ (see Theorem 12 in Bartlett and Mendelson, 2002). Suppose $\|A\|_2 \leq 1$. With $A_1 = (A + A^*)/2$ and $A_2 = (A + A^*)/2$ we can write $A = A_1 + A_2$ with A_i symmetric and $\|A_i\|_2 \leq 1$. By symmetry of the A_i and the spectral theorem we can write $A_i = A_{i1} - A_{i2}$ with $A_{ij} \in \mathcal{L}_2^+(H)$ and $\left\|A_{ij}\right\|_2 \leq 1$. So any $f \in \mathcal{F}$ can be written in the form $f = f_{11} - f_{12} + f_{21} - f_{22}$ with $f_{ij} \in \mathcal{F}^+$, or $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^+ + \mathcal{F}^+ - \mathcal{F}^+$, whence the second inequality also follows from Theorem 12 in Bartlett and Mendelson (2002), or from an application of the triangle inequality. ■

By stratification over balls in $\mathcal{L}_2(H)$ we arrive at a generalization bound valid for *all* bounded operators. If specialized to $\mathcal{L}_0(H)$, the second conclusion below becomes Theorem 1 in the introduction.

Theorem 13 *Let $\psi : \mathbb{R} \to \mathbb{R}$, $\psi \geq 1_{(-\infty,0]}$ with Lipschitz constant L and $\delta > 0$.*

1. *With probability greater than $1 - \delta$ we have for all $T \in \mathcal{L}_\infty(H)$*

$$R(T) \leq \hat{R}_\psi(T, S) + \frac{4L\ \|T^*T\|_2}{m}\left(\sum_{i=1}^m \left\|X_i - X_i'\right\|^4\right)^{1/2} + \sqrt{\frac{9\ln(4\|T^*T\|_2/\delta)}{2m}}.$$

2. *With probability greater than $1 - \delta$ we have for all $T \in \mathcal{L}_\infty(H)$*

$$R(T) \leq \hat{R}_\psi(T, S) + \frac{4L\ \|T^*T\|_2}{\sqrt{m}} + \sqrt{\frac{\ln(2\|T^*T\|_2/\delta)}{2m}}.$$

Proof We will use Lemma 10. For $\alpha \in (0, 1]$ let $\mathcal{V}(\alpha) = \{T \in \mathcal{L}_\infty(H) : \|T^*T\|_2 \leq 1/\alpha\}$ and consider the events

$$\begin{aligned}F(\alpha_1, \alpha_2, \delta) = \Big\{\exists T \in \mathcal{V}(\alpha_2)\ &\text{such that}\\ &R(T) > \hat{R}(T, S) + \frac{2L}{\alpha_1 m}\left(\sum_{i=1}^m \left\|Q_{X_i - X_i'}\right\|^2\right)^{1/2} + \sqrt{\frac{9\ln(2/\delta)}{2m}}\Big\}.\end{aligned}$$

By the first conclusion of Theorem 11 the events $F(\alpha_1, \alpha_2, \delta)$ satisfy hypothesis (i) of Lemma 10, and it is easy to see that (ii) also holds. If we set $a = 1/2$ and replace α by $1/\|T^*T\|_2$, then the conclusion of Lemma 10 becomes the first conclusion above. The second conclusion is proved similarly. ■

3.2 The Objective Functional

Because of the complicated estimates involved, risk bounds such as Theorem 13 have a tendency to be somewhat loose, so that a learning algorithm relying on naive minimization of the bounds may end up with suboptimal hypotheses. On the other hand in the absence of other helpful information such a bound provides a valuable guiding principle, as long as it is transformed to an objective functional which can be minimized in practice and allows for a flexible parametrization of the slack suspected in the bound.

Departing from the simpler of the two conclusions of Theorem 13, a naive approach would look for some $T \in \mathcal{L}_0(H)$ to minimize the objective functional

$$\hat{R}_\psi(T,S) + \frac{4L \ \|T^*T\|_2}{\sqrt{m}} + \sqrt{\frac{\ln(2\|T^*T\|_2/\delta)}{2m}}.$$

Our first modification is to discard the last term on the grounds that it will be dominated by second one. This is only justified if we exclude extremely small values of the confidence parameter δ and work with operators of reasonably large norm, so that $\|T^*T\|_2$ is substantially greater than $\sqrt{\ln\|T^*T\|_2/\delta}$. The new objective functional reads

$$\hat{R}_\psi(T,S) + \frac{4L \ \|T^*T\|_2}{\sqrt{m}}.$$

Our second modification is to replace the factor $4L$ in the second term by an adjustable regularization parameter $\lambda > 0$. On the one hand this just absorbs the Lipschitz constant L of the function ψ (which is yet to be fixed), on the other hand it expresses the belief that $\|T^*T\|_2/\sqrt{m}$ gives the right order of the true estimation error. We will stick to this belief, even though it can be successfully argued that it is naive, because the estimation error may decay much more rapidly than $m^{-1/2}$ as shown by Bartlett et al. (2005) and several other works. There were in fact experimental indications, that in our case the decay is not much better than $m^{-1/2}$, because the same value of λ appeared to work very well for different applications with rather different values of m (see Section 6.1).

The objective functional now depends on ψ and λ and has the form

$$\Lambda_{\psi,\lambda}(T,S) = \hat{R}_\psi(T,S) + \frac{\lambda\|T^*T\|_2}{\sqrt{m}}. \tag{4}$$

We still have to fix the Lipschitz function ψ, satisfying $\psi \geq 1_{(-\infty,0]}$. We want it to be as small as possible to reduce slack, but it should be convex for practical reasons. It is easy to show that any convex function ψ with $\psi \geq 1_{(-\infty,0]}$ and Lipschitz constant L satisfies $\psi \geq h_{L^{-1}}$, where h_γ is the hinge loss with margin $\gamma > 0$, defined by

$$h_\gamma(t) = \begin{cases} 1 - t/\gamma & \text{if} \quad t \leq \gamma \\ 0 & \text{if} \quad t > \gamma \end{cases}.$$

We settle for the hinge loss, not only because it is optimal with respect to convexity, the Lipschitz condition and the lower bound constraint, but because of its simplicity. Our final objective function thus depends on the two parameters λ and γ and reads

$$\Lambda_{h_\gamma,\lambda}(T,S) = \frac{1}{m}\sum_{i=1}^m h_\gamma\left(r_i\left(1 - \left\|Tx_i - Tx_i'\right\|^2\right)\right) + \frac{\lambda\|T^*T\|_2}{\sqrt{m}},$$

for a sample $S = ((x_1, x_1', r_1), \ldots, (x_m, x_m', r_m))$ generated in m independent, identical trials of the pair oracle ρ. The proposed algorithm searches for $T \in \mathcal{L}_\infty(H)$ to minimize $\Lambda_{h_\gamma,\lambda}(T,S)$.

While the hinge loss is used in most of our experiments, there are other choices. It is inherent to the problem of similarity learning that one is led to consider an asymmetric response to similar and dissimilar examples (see, for example, the approaches of Bar-Hillel et al., 2005 and Xing et al., 2002). This is implemented by making the Lipschitz function ψ dependent on the parameter r which indicates similarity or dissimilarity. We thus consider two functions $\psi_1, \psi_{-1} : \mathbb{R} \to \mathbb{R}$, $\psi_i \geq 1_{(-\infty,0]}$ and the objective functional

$$\Lambda_{\psi_1,\psi_{-1},\lambda}(T,S) = \frac{1}{m} \sum_{i=1}^{m} \psi_{r_i}\left(r_i \left(1 - \left\| T\left(x_i - x_i'\right) \right\|^2 \right) \right) + \frac{\lambda \|T^*T\|_2}{\sqrt{m}}.$$

Note that our generalization bounds remain valid for the empirical risk estimate using ψ_1 and ψ_{-1} as long as we use for L the Lipschitz constant of $\min\{\psi_1, \psi_{-1}\}$. Using $\psi_i = h_{\gamma_i}$ we are effectively considering an *inside margin* γ_1, which applies to similar pairs, and an *outside margin* γ_{-1}, which applies to dissimilar pairs.

The use of different margins, or more generally different functions ψ_i in response to similar and dissimilar examples is nonsensical from the point of view of our bounds, which would always assume a smaller value if we used $\min\{\psi_1, \psi_{-1}\}$ to begin with. These bounds however were imported from the inherently symmetric vector valued case to a situation which is inherently asymmetric, because $\langle Q_{X-X'}, T^*T \rangle_2 = \|TX - TX'\|^2 \geq 0$ for all possible solutions T. The asymmetric margins may therefore have their merit and an instance of asymmetric margins is described in Section 4.

Even with symmetric margins the response will be asymmetrical: If we use $\gamma = 1$ (as in fact we did in our experiments) every similar pair (x_i, x_i') will make a contribution to the objective function unless $Tx_i = Tx_i'$, whereas dissimilar pairs will only contribute if $\|Tx_i - Tx_i'\|^2 < 2$, which can exclude many examples of dissimilarity, in particular if the regularization parameter λ is small.

Why not use the trace-norm $\|T^*T\|_1 = \|T\|_2^2$ as a regularizer? Since $\|T^*T\|_2 \leq \|T^*T\|_1$ the trace-norm could be substituted in our bounds and the regularization part of the algorithm in Table 1 below would simplify considerably if we used $\|T\|_2^2$ instead of $\|T^*T\|_2$. Regularization with the trace-norm is also believed to enforce sparsity in the sense of a low rank of T.

The trace-norm was not used for three reasons:

1. The bounds become looser upon substitution of $\|T^*T\|_1$. While this is obvious, one could argue, that there may be other bounds which work well with $\|T^*T\|_1$. Besides, the idea of minimizing bounds has to be approached with caution, so this argument is not decisive.

2. The trace-norm does not work as well in practice. In all experiments the trace norm was tried and performance found to be slightly inferior (while still comparable) to the use of the Hilbert-Schmidt norm $\|T^*T\|_2$ (see Section 6).

3. Regularization with the trace norm can cause too much sparsity and instability of the learning algorithm. This can be seen by simplifying the empirical part of the objective to be linear in $V = T^*T$ (that this is a valid approximation is shown in Proposition 15 below). Then there is an empirical operator A (made explicit in Section 4.2) such that an objective functional can be written as
$$-\langle V, A \rangle_2 + \lambda \|T^*T\|_p^p,$$

to be minimized with positive operators V. In this case the minimizers can be given explicitly in terms of A. For $p=2$ (the Hilbert-Schmidt case) one finds that the minimizer is a multiple of the positive part A_+ of the empirical operator A (the source of the sparsity observed in our experiments), and stable under small perturbations of the eigenvalues of A. For $p=1$ however the minimizer V will be a multiple of the projection to the subspace spanned by the eigenvectors corresponding to the largest positive eigenvalue of A. This space will be generically one-dimensional, making it useless for many data-representations. If it is more than one-dimensional then it will be unstable under perturbations of the eigenvalues of A.

3.3 Minimization of the Objective Functional

Throughout this section fix a sample $S=((x_1,x_1',r_1),\ldots,(x_m,x_m',r_m))$ and assume that $\psi:\mathbb{R}\to\mathbb{R}$ is convex, satisfying $\psi\geq 1_{(-\infty,0]}$. For $\lambda>0$ consider the functional $\Omega_{\psi,\lambda}:\mathcal{L}_2^+(H)\to\mathbb{R}$

$$\Omega_{\psi,\lambda}(A)=\frac{1}{m}\sum_{i=1}^{m}\psi\left(r_i\left(1-\left\langle Q_{x_i-x_i'},A\right\rangle_2\right)\right)+\frac{\lambda\|A\|_2}{\sqrt{m}}. \tag{5}$$

Comparison with (4) reveals that $\Lambda_{\psi,\lambda}(T)=\Omega_{\psi,\lambda}(T^*T)$, so that any operator T is a minimizer of $\Lambda_{\psi,\lambda}$ if and only if T^*T is a minimizer of $\Omega_{\psi,\lambda}$ in $\mathcal{L}_2^+(H)$, a simple fact which has several important consequences. Note that $\Omega_{\psi,\lambda}$ is convex if ψ is convex, and since $\mathcal{L}_2^+(H)$ is a convex set we obtain a convex optimization problem. The situation somewhat resembles that of an SVM (in particular if ψ is the hinge loss), but solutions cannot be sought in all of $\mathcal{L}_2(H)$ but must lie in the cone $\mathcal{L}_2^+(H)$, a positivity constraint which makes the optimization problem quite different.

Denote with M the linear span of $\{x_i-x_i':i=1,\ldots,m\}$ in H and define a map from M^m to $\mathcal{L}_2^+(H)$ by

$$A_{\mathbf{v}}=\sum_{i=1}^{m}Q_{v_i}\text{ where }\mathbf{v}=(v_1,\ldots,v_m)\in M^m.$$

That $A_{\mathbf{v}}\in\mathcal{L}_2^+(H)$ follows from Lemma 3. We also define a linear transformation $T_{\mathbf{v}}:H\to\mathbb{R}^m$ by setting

$$(T_{\mathbf{v}}z)_k=\langle z,v_k\rangle\text{ for }k=1,\ldots,m\text{ and }z\in H. \tag{6}$$

Then we have $T_{\mathbf{v}}^*T_{\mathbf{v}}=A_{\mathbf{v}}$. Also note that $\mathbf{v}\leftrightarrow T_{\mathbf{v}}$ establishes a continuous bijection between M^m and the set of all linear transformations $T:H\to\mathbb{R}^m$ with $M^\perp\subseteq Ker(T)$.

Suppose we can find some $\mathbf{v}\in M^m$ such that $\Omega_{\psi,\lambda}(A_{\mathbf{v}})\leq\Omega_{\psi,\lambda}(A)$ for all $A\in\mathcal{L}_2^+(H)$. Then $\Lambda_{\psi,\lambda}(T_{\mathbf{v}})=\Omega_{\psi,\lambda}(T_{\mathbf{v}}^*T_{\mathbf{v}})=\Omega_{\psi,\lambda}(A_{\mathbf{v}})$ is also optimal, so $T_{\mathbf{v}}$ will be an optimal pre-processor and we are done. To find such an optimal $\mathbf{v}\in M^m$ we plan to do gradient descent of $\Omega_{\psi,\lambda}(A_{\mathbf{v}})$ in the parameter $\mathbf{v}$ which automatically ensures the positivity constraint and keeps us comfortable in an at most m-dimensional environment. Since $\Omega_{\psi,\lambda}(A_{\mathbf{v}})$ is not generally convex in $\mathbf{v}$, even if $\Omega_{\psi,\lambda}$ is convex, one might worry about the existence of local minima.[1] The following theorem excludes this possibility:

Theorem 14 *Assume that ψ is convex. For $\mathbf{v}\in M^m$ define $\Phi(\mathbf{v})=\Omega_{\psi,\lambda}(A_{\mathbf{v}})$. If Φ has a stable local minimum at $\mathbf{v}\in M^m$ then $A_{\mathbf{v}}$ is a global minimizer of $\Omega_{\psi,\lambda}$ in $\mathcal{L}_2^+(H)$.*

[1] Recall that a real function f on a topological space $\mathcal{X}$ has a local minimum at $x\in\mathcal{X}$ if there is an open set $O\ni x$ such that $f(x)\leq f(y)\ \forall y\in O$

Proof Abbreviate $\Omega_{\psi,\lambda}$ to Ω and use P to denote the orthogonal projection onto M. We first claim that

$$\forall A \in \mathcal{L}_2^+(H) \text{ we have } \Omega(PAP) \leq \Omega(A). \tag{7}$$

This follows from $\left\langle Q_{x_i - x_i'}, PAP \right\rangle_2 = \left\langle Q_{x_i - x_i'}, A \right\rangle_2$ and Lemma 3 (vi) by inspection of (5). Next consider the identity

$$\{A_{\mathbf{v}} : \mathbf{v} \in M^m\} = \left\{PAP : A \in \mathcal{L}_2^+(H)\right\}. \tag{8}$$

First note that the inclusion from left to right follows from $A_{\mathbf{v}} = PA_{\mathbf{v}}P$ for $\mathbf{v} \in M^m$. On the other hand for any $A \in \mathcal{L}_2^+(H)$ all the eigenvectors of PAP with nonzero eigenvalues have to lie in M, and enumerating the eigenvectors e_i of PAP beginning with those in M (of which there can be at most m) we have

$$PAPz = \sum_{i=1}^m \lambda_i \langle z, e_i \rangle e_i = \sum_{i=1}^m \left\langle z, \lambda_i^{1/2} e_i \right\rangle \lambda_i^{1/2} e_i = \sum_{i=1}^m Q_{\left(\lambda_i^{1/2} e_i\right)} z = A_{\mathbf{v}} z$$

for all $z \in H$, so that $PAP \in \{A_{\mathbf{v}} : \mathbf{v} \in M^m\}$ which proves (8).

To prove the theorem let Φ attain a local minimum at $\mathbf{v} \in M^m$. We will assume that Ω does not attain a global minimum at $A_{\mathbf{v}}$ and derive a contradiction. We can write $A_{\mathbf{v}} = T_{\mathbf{v}}^* T_{\mathbf{v}}$, using (6). Since Ω is convex it cannot even attain a local minimum at $A_{\mathbf{v}}$, so there is a sequence $A_n \in \mathcal{L}_2^+(H)$ such that $A_n \to A_{\mathbf{v}}$ and $\Omega(A_n) < \Omega(A_{\mathbf{v}})$. By continuity of multiplication also $PA_nP \to PA_{\mathbf{v}}P = A_{\mathbf{v}}$, by (7) $\Omega(PA_nP) \leq \Omega(A_n) < \Omega(A_{\mathbf{v}})$ and by (8) there exists $\mathbf{v}_n \in M^m$ such that $A_{\mathbf{v}_n} = PA_nP$. We thus have $A_{\mathbf{v}_n} \to A_{\mathbf{v}}$ (this does not imply that $\mathbf{v}_n \to \mathbf{v}$!) and $\Omega(A_{\mathbf{v}_n}) < \Omega(A_{\mathbf{v}})$. By continuity of the square-root $A_{\mathbf{v}_n}^{1/2} \to A_{\mathbf{v}}^{1/2} = |T_{\mathbf{v}}|$. By polar decomposition we can write $T_{\mathbf{v}} = U|T_{\mathbf{v}}|$, where U is a partial isometry, and define $T_n : M \to \mathbb{R}^m$ by $T_n = UA_n^{1/2}$. Then $T_n \to U_{\mathbf{v}}|T_{\mathbf{v}}| = T_{\mathbf{v}}$, so if $\mathbf{w}_n$ is chosen such that $T_n = T_{\mathbf{w}_n}$ we have $\mathbf{w}_n \to \mathbf{v}$, but also $\Phi(\mathbf{w}_n) = \Omega(T_n^* T_n) = \Omega\left(A_n^{1/2} U^* U A_n^{1/2}\right) = \Omega(A_n) < \Omega(A_{\mathbf{v}}) = \Phi(\mathbf{v})$, so Φ cannot attain a local minimum at $\mathbf{v}$. ■

Observe that this result justifies the gradient descent method in the presence of positivity constraints also for other convex loss functions besides the hinge loss. Nevertheless, it does not exclude the existence of points with vanishing gradients away from the global minimum. While the probability of arriving at these points during gradient descent is vanishing, the algorithm can still be slowed down considerably in their neighborhood, a possibility which we have to be prepared for (although it doesn't seem to happen in practice). The theorem therefore only proves that the gradient descent works, but not that it is efficient.

There is an alternative technique (see Xing et al., 2002) of iterative projections which avoids the problem of vanishing gradients and stays more closely to the original convex optimization problem. Briefly, one extends the functional Ω to all of $\mathcal{L}_2(H)$, so that the problem becomes equivalent to an SVM, and then alternates between gradient descent in Ω, which is convex, and projections onto $\mathcal{L}_2^+(H)$. The projection of an operator $A \in \mathcal{L}_2(H)$ to $\mathcal{L}_2^+(H)$ is effected by an eigen-decomposition and reconstruction with all the negative eigenvalues set to zero, so that only the positive part of A is retained. This method would also work for our objective functional, in fact for any convex objective constrained to positive operators. Here this technique was not chosen, because it appeared that the effort of the repeated eigen-decomposition might cancel the advantages of the method. Also the proposed gradient descent, which is easily converted to an online algorithm, appeared more elegant

```
Given sample $S$, regularization parameter $\lambda$, margin $\gamma$, learning rate $\theta$
set $\lambda' = \lambda/\sqrt{|S|}$ and $d = m$
randomly initialize $v = (v_1, ..., v_d)$
repeat
    Compute $\|A_v\|_2 = \left(\sum_{ij} \left\langle v_i, v_j \right\rangle^2\right)^{1/2}$
    For $i = 1, ..., d$ compute $w_i = 2\,\|A_v\|_2^{-1} \sum_j \left\langle v_i, v_j \right\rangle v_i$
    Fetch $(x, x', r)$ randomly from sample $S$
    For $i = 1, ..., d$ compute $a_i \leftarrow \langle v_i, x - x' \rangle$
    Compute $b \leftarrow \sum_{i=1}^{d} a_i^2$
    If $r\,(1-b) < \gamma$
        then for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta\left(\frac{r}{\gamma} a_i\,(x - x') + \lambda' w_i\right)$
        else for $i := 1, ..., d$ do $v_i \leftarrow v_i - \theta\lambda' w_i$
until convergence
```

Table 1: Learning algorithm

in its implicit adherence to the positivity constraint. The ultimate reason to stay with the proposed technique was of course its practical success.

So our algorithm will randomly initialize the vector $\mathbf{v} \in M^m$ and then follow the negative gradient of Φ, either for a specified number of steps or until some heuristic convergence criterion on the value of Φ is met. Straightforward differentiation gives for the k-th component of the gradient of Φ at $\mathbf{v} \in M^m$ the expression

$$\begin{aligned}(\nabla\Phi)_k(\mathbf{v}) &= \frac{-2}{m}\sum_{i=1}^{m} \psi'\left(r_i\left(1-\sum_{j=1}^{m} a_{ij}^2\right)\right) r_i a_{ik}\left(x_i - x_i'\right) \\ &\quad + \frac{2\lambda}{\|A_{\mathbf{v}}\|_2\sqrt{m}}\sum_{j=1}^{m} \left\langle v_k, v_j \right\rangle v_j,\end{aligned}$$

where $a_{ik} = \langle x_i - x_i', v_k \rangle$ and $\|A_{\mathbf{v}}\|_2 = \left(\sum_{i,j} \left\langle v_i, v_j \right\rangle^2\right)^{1/2}$. In Table 1 we give a corresponding algorithm of stochastic gradient descent for the case that ψ is the hinge-loss with margin γ.

In a simplified view, which disregards the regularization term (or if $\lambda = 0$), the algorithm will modify T in an attempt to bring Tx_i and Tx_i' closer if x_i and x_i' are similar (i.e., $r_i = 1$) and their distance exceeds $1-\gamma$, it will attempt to move Tx_i and Tx_i' further apart if x_i and x_i' are dissimilar (i.e., $r_i = -1$) and their distance is less than $1+\gamma$, and it will be indifferent to all other cases. This procedure can also be interpreted in terms of the effect which the individual gradient steps have on the level ellipsoid of the quadratic form induced by T^*T. Figure 1 tries to shows the simple geometrical intuition behind this construction.

There are two heuristic approximations to accelerate this algorithm. A simple time-saver is the observation that the contribution of the regularization term to the gradient changes only very little with small updates $\mathbf{v}$. It therefore doesn't need to be recomputed on every iteration, but it suffices to compute it intermittently.

Also in the experiments reported below a singular value decomposition of the optimal operator revealed that it was dimensionally sparse, in the sense that very few (generally less than 50) singular

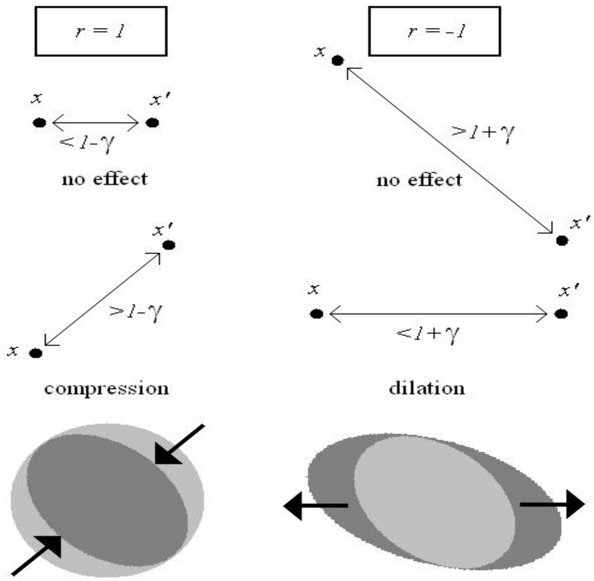

Figure 1: The effect of similar and dissimilar pairs on the level ellipsoid of the quadratic form induced by T^*T in the case of hinge loss with margin γ. If $\|Tx-Tx'\| \leq 1-\gamma$ for similar or $\|Tx-Tx'\| \geq 1+\gamma$ for dissimilar pairs there is no effect. If $\|Tx-Tx'\| < 1+\gamma$ for similar pairs, the ellipsoid is compressed in a direction parallel to the line between x and x'. If $\|Tx-Tx'\| > 1-\gamma$ for dissimilar pairs, the ellipsoid is dilated. Regularization corresponds to a shrinking of the ellipsoid.

values were significantly different from zero. This implies that $A_{\mathbf{v}}$ can be well approximated by some $A_{\mathbf{w}}$ where $\mathbf{w} \in M^d$ with $d \ll m$, and shows that the proposed algorithm effects a dimensional reduction. It also suggests that one might try gradient descent in M^d instead of M^m for $d < m$, which causes a considerable acceleration if $d \approx 10^2$ and $m \geq 10^3$. Of course the argument in Theorem 14 is then no longer valid, because finite dimensional constraints are not convex, so that local minima might become a problem. In practice this never happened for $d \approx 10^2$ and was observed only for $d \leq 5$. The heuristic is implemented by accordingly modifying the initialization $d = m$ in Table 1. In our experiments we used $d = 100$.

There is a simple practical use to an eigen-decomposition of the optimal $A_{\mathbf{v}}$ returned by the algorithm. The v_i will in general not be orthogonal, the algorithm does not enforce this in any way). If the decomposition reveals that only few eigenvalues of $A_{\mathbf{v}}$ are significantly different from zero, we can restrict ourselves to the span of the corresponding eigenvalues. This yields a representation operator T with very low dimensional range, which is easier to compute and facilitates subsequent processing.

By virtue of our dimension free bounds the algorithm is well suited for kernel implementations. Note that the search space consists of vectors $\mathbf{v} = (v_1, ..., v_d) \in M^d$ which admit a linear representation

$$v_i = \sum_{j=1}^{m} \alpha_j^i \left(x_j - x_j'\right)$$

in terms of the training sample. To add two such $\mathbf{v}$ we just add the corresponding matrices α^i_j, to compute inner products we just use the kernel function on the input space. Substituting these rules one finds that there is no problem in kernelization of the algorithm.

4. Similarity Regression and Hyperbolic PCA

In this section we consider some alternatives. The first is rather obvious and extends the method described above to continuous similarity values. The other method is derived from the risk functional R and has already been described in Maurer (2006a).

4.1 Similarity Regression

The bounds in Section 3.1 have straightforward extensions to the case, when the oracle measure is not supported on $\mathcal{X}^2 \times \{-1,1\}$ but on $\mathcal{X}^2 \times [0,1]$, corresponding to a continuum of similarity values, and $\ell : [0,1] \times \mathbb{R} \to \mathbb{R}$ is a loss function such that $\ell(y,.)$ has Lipschitz constant at most L for all $y \in [0,1]$. The corresponding risk functional to be minimized would be

$$R'(T) = \mathbb{E}_{(x,x',r)\sim\rho}\left[\ell\left(r, \left\|Tx - Tx'\right\|^2\right)\right].$$

If ℓ has the appropriate convexity properties, then an obvious modification of the proposed algorithm can be used. With a least-squares loss function the square of the norm (an unavoidable feature of our method) will lead to an overemphasis of large distances, probably an undesirable feature which can be partially compensated by a redefinition of the loss function at the expense of a large Lipschitz constant (e.g., with $\ell(y,t) = (y-t)^2 / (y-y_0)^2$ for some $y_0 > 0$).

The possibilities of this type of similarity regression (or learning of metrics) remain to be explored.

4.2 Risk Bounds with Affine Loss Functions and Hyperbolic PCA

Consider again the case of a binary oracle (similar/dissimilar) and the task of selecting an operator from some set $\mathcal{V} \subset \mathcal{L}_2(H)$.

Proposition 15 *Suppose* $1 < \left\|\mathcal{V}\right\|_\infty = c < \infty$ *and set* $\eta_1 = -1$ *and* $\eta_{-1} = 1/\left(c^2-1\right)$. *Then for all* $T \in \mathcal{V}$

$$R(T) \leq 1 + \mathbb{E}_{(x,x',r)\sim\rho}[\eta_r] - \left\langle T^*T, \mathbb{E}_{(x,x',r)\sim\rho}[\eta_r Q_{x-x'}]\right\rangle_2.$$

For a balanced oracle this becomes

$$R(T) \leq \frac{c^2}{2(c^2-1)} - \left\langle T^*T, \mathbb{E}_{(x,x',r)\sim\rho}[\eta_r Q_{x-x'}]\right\rangle_2.$$

Proof Define real functions ψ_1, ψ_{-1} by $\psi_1(t) = 1-t$ and $\psi_{-1}(t) = 1 - t/\left(c^2-1\right)$. Since $\|x-x'\| \leq 1$ a.s. and by the definition of c we have for $T \in \mathcal{V}$

$$1_{(-\infty,0]}\left(r\left(1 - \left\|TX - TX'\right\|^2\right)\right) \leq \psi_r\left(r\left(1 - \left\|TX - TX'\right\|^2\right)\right) \text{ a.s.}$$

We also have for $r \in \{-1,1\}$ that $\psi_r(t) = 1 + r\eta_r t$, whence

$$\begin{aligned} R(T) &= \mathbb{E}_{(x,x',r)\sim\rho}\left[1_{(-\infty,0]}\left(r\left(1-\left\|TX-TX'\right\|^2\right)\right)\right] \\ &\leq \mathbb{E}_{(x,x',r)\sim\rho}\left[\psi_r\left(r\left(1-\left\|TX-TX'\right\|^2\right)\right)\right] \\ &= \mathbb{E}_{(x,x',r)\sim\rho}\left[1+\eta_r\left(1-\left\|TX-TX'\right\|^2\right)\right] \\ &= 1+\mathbb{E}_{(x,x',r)\sim\rho}[\eta_r] - \left\langle T^*T, \mathbb{E}_{(x,x',r)\sim\rho}[\eta_r Q_{x-x'}]\right\rangle_2 . \end{aligned}$$

Since for a balanced oracle $1+\mathbb{E}_{(x,x',r)\sim\rho}[\eta_r] = c^2/\left(2\left(c^2-1\right)\right)$, the second conclusion is immediate. ■

Of course we can use the risk bounds of the previous section for an empirical estimator constructed from the Lipschitz functions ψ_1, ψ_{-1} used in the proof above, corresponding to an inner margin of 1 and an outer margin of c^2-1 (see Section 3.1 and 3.2). The affine nature of these functions however allows a different, more direct analysis and a different algorithmic implementation.

The operator

$$A := \mathbb{E}_{(x,x',r)\sim\rho}[\eta_r Q_{x-x'}]$$

is the expectation of the operator-valued random variable $(x,x',r) \mapsto \eta_r Q_{x-x'}$. Minimizing the bounds in Proposition 15 is equivalent to maximizing $\langle T^*T, A\rangle_2$, which is the only term depending on the operator T. Given the sample $S = ((x_1,x_1',r_1),\ldots,(x_m,x_m',r_m))$ the obvious way to try this is by maximizing the empirical counterpart $\langle T^*T,\hat{A}\rangle_2$ where $\hat{A}$ is the *empirical operator*

$$\hat{A}(S) = \frac{1}{m}\sum_{i=1}^{m} \eta_{r_i} Q_{x_i - x_i'} . \tag{9}$$

The next result addresses the issue of estimation. The result is similar to Theorem 13 but can be obtained without the use of Rademacher averages.

Theorem 16 *Under the above assumptions and if $c \geq 2$ we have for $\delta > 0$ with probability greater $1-\delta$ in the sample S that for all $T \in \mathcal{V}$*

$$\left|\langle T^*T,\hat{A}\rangle_2 - \langle T^*T,A\rangle_2\right| \leq \frac{2}{\sqrt{m}}\left(\sup_{T\in\mathcal{V}}\|T^*T\|_2 + c^2\sqrt{2\ln(1/\delta)}\right).$$

Proof Apply Lemma 9 to the $\mathcal{L}_2(H)$-valued random variable $\eta_r Q_{x-x'} - \mathbb{E}[\eta_r Q_{x-x'}]$. The corresponding substitutions give

$$\Pr\left\{\left|\langle T^*T,\hat{A}\rangle_2 - \langle T^*T,A\rangle_2\right| > \varepsilon\right\} \leq \exp\left(\frac{-\left(\sqrt{m}\varepsilon - 2\sup_{T\in\mathcal{V}}\|T^*T\|_2\right)^2}{8c^4}\right),$$

and the result follows from equating the right side to δ. ■

Of course an analogous result could have been obtained from Theorem 11.

We now specialize the space of candidate operators to $\mathcal{V} = c\mathcal{P}_d$ where $\mathcal{P}_d$ is the set of d-dimensional orthogonal projections. Then $\sup_{T\in\mathcal{V}} \|T^*T\|_2 = c^2\sqrt{d}$ in the bound above. The optimization problem now is to maximize $\left\langle P, \hat{A}\right\rangle_2$ for $P \in \mathcal{P}_d$ which is done by projecting onto a maximal eigenspace of $\hat{A}$, as shown by the following proposition.

Proposition 17 *Suppose $A \in \mathcal{L}_2(H)$ is symmetric with eigenvectors e_i and corresponding eigenvalues λ_i. Suppose that $d \in \mathbb{N}$ and that the sum in the eigen-expansion of A can be ordered in such a manner that $\lambda_i \geq \lambda_j$ for all $i \leq d < j$. Then*

$$\max_{P\in\mathcal{P}_d} \langle A, P\rangle_2 = \sum_{i=1}^{d} \lambda_i,$$

the maximum being attained by the projection onto the span of $(e_i)_{i=1}^d$.

Proof let $P \in \mathcal{P}_d$ with $v_1, \ldots, v_d$ being an orthonormal basis for the range of P. Then

$$\begin{aligned}
\langle A, P\rangle_2 &= \sum_{j=1}^{d}\sum_{i=1}^{d} \lambda_i \left\langle v_j, e_i\right\rangle^2 + \sum_{j=1}^{d}\sum_{i=d+1}^{\infty} \lambda_i \left\langle v_j, e_i\right\rangle^2 \\
&\leq \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{d} \left\langle v_j, e_i\right\rangle^2 + \lambda_d \sum_{j=1}^{d} \left(\sum_{i=d+1}^{\infty} \left\langle v_j, e_i\right\rangle^2 \right) \\
&= \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{d} \left\langle v_j, e_i\right\rangle^2 + \lambda_d \sum_{i=1}^{d} \left(1 - \sum_{j=1}^{d} \left\langle v_j, e_i\right\rangle^2 \right) \\
&\leq \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{d} \left\langle v_j, e_i\right\rangle^2 + \sum_{i=1}^{d} \lambda_i \left(1 - \sum_{j=1}^{d} \left\langle v_j, e_i\right\rangle^2 \right) = \sum_{i=1}^{d} \lambda_i,
\end{aligned}$$

which proves $\sup_{P\in\mathcal{P}_d} \langle A, P\rangle_2 \leq \sum_{i=1}^{d} \lambda_i$ (this also follows directly from Horn's theorem; Simon, 1979, Theorem 1.15). If P is the projection onto the span of $(e_i)_{i=1}^d$ we can set $v_j = e_j$ above and obtain an equality. ■

This gives an alternative algorithm to the one described in Section 3: Fix a quantity $c > 2$ and construct a matrix representation of the empirical operator $\hat{A}$ given in (9). For some fixed target-dimension d find the projection onto a d-dimensional dominant eigenspace of $\hat{A}$. We omit the rather straightforward technical details concerning to the representation of $\hat{A}$ and the implementation of a kernel. Some of these issues are discussed in Maurer (2006a) where corresponding experiments are reported.

There is an intuitive interpretation to this algorithm, which could be called hyperbolic PCA: The empirical objective functional is proportional to

$$\begin{aligned}
m\left\langle P, \hat{A}\right\rangle_2 &= \sum_{i=1}^{m} \eta_{r_i} \left\| Px_i - Px_i' \right\|^2 \\
&= \frac{1}{c^2 - 1} \sum_{i: x_i, x_i' \text{ dissimilar}} \left\| Px_i - Px_i' \right\|^2 - \sum_{i: x_i, x_i' \text{ similar}} \left\| Px_i - Px_i' \right\|^2 .
\end{aligned}$$

When similarity and dissimilarity are defined by class memberships, then maximizing this expression corresponds to maximizing inter-class- and minimizing intra-class variance, where the parameters $1/\left(c^2-1\right)$ and the proportions of similar and dissimilar pairs control the trade-off between these potentially conflicting goals. A similar proposal can be found in Thrun (1998). Typically we have $c \gg 1$ so that $1/\left(c^2-1\right) \ll 1$, so that dissimilar pairs ('negative equivalence constraints') receive a much smaller weight than similar pairs, corresponding to the intuitive counting argument given ba Bar-Hillel et al. (2005).

The method is similar to principal component analysis insofar as it projects to a principal eigenspace of a symmetric operator. In contrast to PCA, where the operator in question is the empirical covariance operator, which is always nonnegative, we will project to an eigenspace of an empirical operator which is a linear combination of empirical covariances and generally not positive. While the quadratic form associated with the covariance has elliptic level sets, the empirical operator induces hyperbolic level sets.

5. Applications to Multi-category Problems and Learning to Learn

In this section we apply similarity learning to problems where the nature of the application task is partially or completely unknown, so that the available data is used for a preparatory learning process, to facilitate future learning.

5.1 Classification Problems Involving a Large Number of Categories

A multi-category task τ with input space $\mathcal{X}$ is a pair $\tau = (\mathcal{Y}, \mu)$ where $\mathcal{Y}$ is a finite or countable alphabet of labels and μ a probability measure on $\mathcal{X} \times \mathcal{Y}$. We interpret $\mu(x,y)$ as the probability to encounter the pattern x with label y. As usual we assume $\mathcal{X}$ to be embedded in the Hilbert space H.

Let $\tau = (\mathcal{Y}, \mu)$ be such a task, $T \in \mathcal{L}_0(H)$ and suppose that we are given a *single* labeled example $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Any classifier trained on this example alone and applied to another pattern $x' \in \mathcal{X}$ can sensibly only make the decisions "*x' is of type y*" or "*x' is not of type y*" or no decision at all. Face verification is a case where such classifiers can be quite important in practice: Anyone having to verify the identity of a person on the basis of a single photograph has to learn and generalize on the basis of a single example image. A simple classifier using only the pseudo-metric induced by T is the *elementary verifier* $\varepsilon_T(x,y)$ which decides

$$\begin{array}{ccc} x' \textit{ is of type } y & \text{if} & \|T(x-x')\| < 1 \\ \textit{undecided} & \text{if} & \|T(x-x')\| = 1 \\ x' \textit{ is not of type } y & \text{if} & \|T(x-x')\| > 1 \end{array} .$$

Relative to the task $\tau = (\mathcal{Y}, \mu)$ it has the error probability (counting 'undecided' as an error)

$$\operatorname{err}_\tau(\varepsilon_T(x,y)) = \Pr_{(x',y') \sim \mu} \left\{ r\left(y,y'\right)\left(1 - \left\|T\left(x-x'\right)\right\|\right) \leq 0 \right\},$$

where the function $r : \mathcal{Y} \times \mathcal{Y} \to \{-1,1\}$ quantifies equality and inequality: $r(y,y') = 1$ if $y = y'$ and $r(y,y') = -1$ if $y \neq y'$.

There is a canonical pair oracle derived from the task τ. It is the probability measure ρ_τ on $\mathcal{X}^2 \times \{-1,1\}$ given by

$$\rho_\tau(A) = \Pr_{((x,y),(x',y')) \sim \mu^2} \left\{ \left(x, x', r\left(y,y'\right)\right) \in A \right\} \text{ for } A \subseteq \mathcal{X}^2 \times \{-1,1\}. \tag{10}$$

To generate a draw (x,x',r) from ρ_τ make two independent draws of (x,y) and (x',y') from μ and then return $(x,x',1)$ if $y=y'$ and $(x,x',-1)$ if $y\neq y'$. Then

$$\mathbb{E}_{(x,y)\sim\mu}\left[\mathrm{err}_\tau\left(\varepsilon_T\left(x,y\right)\right)\right]=R_{\rho_\tau}\left(T\right), \tag{11}$$

so we can use the risk bounds in Sections 3.1 or 4.2 to bound the expected error of the elementary verifier $\varepsilon_T\left(x,y\right)$ under a random draw of the training example (x,y).

If there are many labels appearing approximately equally likely, then dissimilar pairs will be sampled much more frequently than similar ones, resulting in a negative bias of elementary classifiers. Similar unwanted effects have been noted by Xing et al. (2002) and Bar-Hillel et al. (2005). This does not mean that our bounds are paradoxical, because the biased sampling corresponds to an equally biased error measure: Under these circumstances the non-verifier which always asserts ”*x' is not of type y*” would already have a small error.

This problem can be avoided by a simple balancing technique: In the computation of the error of the elementary classifier $\varepsilon_d\left(x,y\right)$ we assign different weights to the cases of false rejection and false acceptance. Define a balanced error

$$\begin{aligned}\bar{\mathrm{err}}_\tau\left(\varepsilon_d\left(x,y\right)\right) &= \frac{1}{2C_1}\Pr_{(x',y')\sim\mu}\left\{\left\|T\left(x-x'\right)\right\|\geq 1 \text{ and } y'=y\right\}\\ &\quad+\frac{1}{2C_{-1}}\Pr_{(x',y')\sim\mu}\left\{\left\|T\left(x-x'\right)\right\|\leq 1 \text{ and } y'\neq y\right\},\end{aligned}$$

where $C_1=\mu^2\left\{\left(\left(x,y\right),\left(x',y'\right)\right):y'=y\right\}$ and $C_{-1}=\mu^2\left\{\left(\left(x,y\right),\left(x',y'\right)\right):y'=y\right\}$ are the probabilities to obtain examples with equal and unequal labels respectively in two independent draws of μ. If we define a balanced pair oracle $\bar{\rho}$ by

$$\bar{\rho}_\tau\left(A\right)=\frac{\rho_\tau\left(A\cap\mathcal{X}^2\times\{1\}\right)}{2\rho_\tau\left(\mathcal{X}^2\times\{1\}\right)}+\frac{\rho_\tau\left(A\cap\mathcal{X}^2\times\{-1\}\right)}{2\rho_\tau\left(\mathcal{X}^2\times\{-1\}\right)},$$

where ρ_τ is defined in (10) then one again verifies that

$$\mathbb{E}_{(x,y)\sim\mu}\left[\bar{\mathrm{err}}_\tau\left(\varepsilon_d\left(x,y\right)\right)\right]=R_{\bar{\rho}_\tau}\left(T\right) \tag{12}$$

and that $\bar{\rho}_\tau$ returns similar and dissimilar pairs with equal probability. To generate a draw of (x,x',r) from $\bar{\rho}_\tau$ first draw (x,y) from μ and then flip a fair coin. On heads draw x' from the class conditional distribution of y and return $(x,x',1)$, on tails draw x' from the conditional distribution for $\mathcal{Y}\setminus\{y\}$ (or continue to draw $(x',y')\sim\mu$ until $y'\neq y$) and return $(x,x',-1)$.

Whichever of the two oracles we use: If we make m independent draws from it to obtain a sample $S=\left(\left(x_1,x_1',r_1\right),...,\left(x_m,x_m',r_m\right)\right)$ from which we generate the operator T according to the algorithm derived in Section 3, then we are essentially minimizing a bound on the expected error of elementary verifiers trained on future examples. In this way the proposed algorithm can be considered an algorithm of learning to learn.

This is particularly interesting if $m<|\mathcal{Y}|$, because we obtain performance guarantees for categories which we haven't seen before. Consider for example the case where $\mathcal{Y}$ stands for a large population of human individuals, and the inputs correspond to facial images. From a sample $S=\left(\left(x_1,x_1',r_1\right),...,\left(x_m,x_m',r_m\right)\right)$ we train an operator T and then use the elementary verifiers ε_T for face-verification on the entire population. With methods similar to the balancing technique described above the oracle can be also modified to achieve different desired penalties for false rejection and false acceptance.

5.2 Similarity as a Vehicle for Transfer

The face-verification system proposed above is somewhat naive from an economical point of view, because the example-pairs have to be obtained independently, which means that we will need images of approximately $3m/2$ individuals (with balancing), all of which will probably have to be paid to cooperate. It would be preferable to collect the data independently but conditional to a smaller subpopulation $\mathcal{Y}' \subset \mathcal{Y}$ with $|\mathcal{Y}'| < m$, so that multiple images can be gathered from each individual. This effectively replaces the original task $\tau = (\mathcal{Y}, \mu)$ with a subtask $\tau' = (\mathcal{Y}', \mu')$, where

$$\mu'(A) = \frac{\mu(A)}{\mu(\mathcal{X} \times \mathcal{Y}')} \text{ for } A \subseteq \mathcal{X} \times \mathcal{Y}'.$$

Of course events which are independent w.r.t. μ are not independent w.r.t. μ' and vice versa, so if we sample independently from μ' our generalization guarantees will only work for the task τ'. The operator T generated from the sample drawn from μ' may nonetheless work for the original task τ corresponding to the entire population.

This points to a different method to apply the proposed algorithm: Use one task $\tau = (\mathcal{Y}, \mu)$, the training task, to draw the training sample and train the representation T, and apply T to the data of the application task $\tau' = (\mathcal{Y}', \mu')$.

As the application task is unknown at the time of training T, this transfer mechanism may of course fail. Deciding between success or failure however does not have the sample complexity of learning, which is affected by a complexity penalty of the function class, but only the sample complexity of validation, which can be determined from Hoeffdings inequality. If T has been trained from τ and is subsequently tested on $\tau' = (\mathcal{Y}', \mu')$ then with probability greater $1 - \delta$ in an i.i.d. sample S' of size m drawn from μ' we have

$$R_{\rho_{\tau'}}(T) \leq \hat{R}_{1_{(-\infty,0]}}(T, S') + \sqrt{\frac{\ln(1/\delta)}{2m}},$$

which is of course much better than the bounds in Theorem 11 (here $\hat{R}_{1_{(-\infty,0]}}(T, S')$ is the empirical risk of T on S').

Of course this type of transfer can be attempted between any two binary classification tasks, but its success normally requires a similarity of the pattern classes themselves. A classifier trained to distinguish images of the characters ”a” and ”b” will be successfully applied to two classes of images if these pattern classes have some resemblance of ”a” and ”b”. A representation of similarity however can be successfully transferred if there is a similar notion of similarity applicable to both problems. It is this higher conceptual level of similarity concepts that allows them to transcend task-boundaries. This kind of ’meta-similarity” is for example present, when there is a common process of data generation, as is the case when the tasks share an invariance property.

Suppose now that the learner has trained operators $T_1, ..., T_K$ from the samples $S_1, ..., S_K$ gathered in past experience, with each S_k drawn iid from a corresponding multi-category task τ_k, and that the learner is currently confronted with a new task τ', for which a new sample S' is available. A simple union bound yields from Hoeffdings inequality the following standard result: $\forall \delta > 0$ with probability greater $1 - \delta$ in S' we have that for every $k \in \{1, ..., K\}$

$$R_{\rho_{\tau'}}(T_k) \leq \hat{R}_{1_{(-\infty,0]}}(T_k, S') + \sqrt{\frac{\ln K + \ln(1/\delta)}{2m}}.$$

Let us assume that K is not too large. Minimizing $\hat{R}_{1_{(-\infty,0]}}(T_k, S')$ over k is a simple algorithm which does two things:

1. It finds the operator T_{k^*} which optimally represents the data of τ' for the purpose of recognition on the basis of single training examples.

2. It classifies the new task τ into one of the K meta-categories defined by the old tasks $\tau_1, ..., \tau_K$. The classification is essentially carried out on the basis of compatibility of underlying notions of similarity.

The more obvious practical aspect is the first, because T_{k^*} can be put to use right away. There are however interesting strategies for ”life long learning” (Thrun, 1998) where the mechanism of task classification is also useful. If the optimal empirical risk $\hat{R}_{1_{(-\infty,0]}}(T_k, S')$ is too large one could use S' to train T' which is appended to $(T_1, ..., T_K)$ to reflect the fact that a new type of task has been discovered. If $\hat{R}_{1_{(-\infty,0]}}(T_k, S')$ is small on the other hand, one could merge the data S' of the new task with the data S_{k^*} of the most closely matching task and retrain on $S' \cup S_{k^*}$ to obtain an operator T'' to replace T_{k^*} for better generalization due to the larger sample size, and keep the number K constant.

In this context it should be noted, that ”multi-task learning”, the idea of pooling the data from various tasks to achieve a smaller estimation error (see Caruana 1998; Baxter, 2000; Evgeniou et al., 2004; Maurer 2006b), is easily implemented in the case of similarity learning by just concatenating the samples $S_1 \cup ... \cup S_K$ and training. The concatenated sample corresponds to a draw from the mixed pair oracle $\rho = \sum c_k \rho_k$ with $c_k = |S_k| / \sum_i |S_i|$.

The proposed method has been derived from the objective of minimizing the risk functional R which is connected to classification through the identities (11) and (12). It is therefore a principled technique to train representations for future learning on the basis of a *single example*. While the representation T can be used to preprocess data for other algorithms operating on larger future training examples, there is reason to believe that it will be no longer optimal if there are more examples. It is a challenging problem to define risk functionals giving optimality for algorithms operating on other future sample sizes. While the algorithm proposed in Argyriou et al. (2006) appears to have such properties, corresponding risk bounds are lacking and the relationship to the current work remains to be explored.

6. Experiments

All the experiments reported concern transfer in machine vision where a representation trained on one task is applied to another one. The experiments involved the recognition of randomly rotated-, randomly scaled-, randomly rotated and scaled-, and handwritten characters, spatially rotated objects and face recognition. Below we briefly describe the parametrization of the training algorithm, the various tasks tried and the parameters recorded in testing. An executable to reproduce most of the experiments will be made available at the web-site www.andreas-maurer.eu.

6.1 Parametrization and Experimental Setup

All the experiments with character recognition used gray-scale images of 28×28 pixels, corresponding to 784-dimensional vectors. The images of the COIL100 database had 64×64, the images of the ATT face database had 92×112 pixels. Pixel vectors were normalized to unity, otherwise there was no preprocessing. The embedding in the Hilbert space H was effected by the Gaussian RBF-kernel

$$\langle x,y\rangle = \kappa(x,y) = (1/2)\exp\left(-4\left|\frac{x}{|x|}-\frac{y}{|y|}\right|^2\right),$$

where x and y are two images, $|.|$ is the euclidean norm on pixel vectors and $\langle .,.\rangle$ is the inner product of the embedded vectors in the RKHS H (see, for example, Christianini and Shawe-Taylor, 2000, for kernel techniques).

For all tasks involved in the experiments corresponding data-sets were generated and processed in this kernel-representation.

In every transfer experiment there was a training task and an application- or test task, represented by corresponding labeled data-sets. On the data-set of the training task the algorithm given in Table 1 was used, together with the accelerating heuristics in Section 3.3, to generate a representation T. All the experiments reported below were carried out with margin $\gamma = 1$ and the regularization parameter either $\lambda = 0.005$ (for the Hilbert-Schmidt norm $\|T^*T\|_2$) or $\lambda = 0.001$ (for $\|T^*T\|_1 = \|T\|_2^2$). These values were determined by cross validation for problem of handwritten character recognition below and reused in all the other experiments. Separately optimizing parameters for each experiment using cross validation would only lead to an improvement of the results.

The gradient descent was carried out for 10^6 steps with a constant learning rate $\theta = 0.01$. For the training task we report the final values of the objective function Λ and the empirical risk $\hat{R}(T) = \hat{R}_{1_{(-\infty,0]}}(T)$. Another interesting property of T is its 'essential rank' as the number of singular values appreciably larger than 0. In the results below this is referred to as 'sparsity' and given as the number of singular values larger than 2% of the spectral norm.

The representation T is applied to the data-set of the application task, with pixel vectors equi-dimensional to those in the training data-set. Application and training data-sets had no overlapping categories.

On the application task we measured three properties related to the quality of the representation:

1. The empirical risk $\hat{R}(T) = \hat{R}_{1_{(-\infty,0]}}(T)$ as an estimator for the true risk $R(T)$. This relates to the theory above and to the performance of elementary verifiers.

2. The area under the ROC-curve (ROC area T) for the distance as a detector of class-equality. This can be regarded as an estimator for the probability that a pattern pair with equal labels is represented at a closer distance than an independently chosen pair with different labels.

3. The error (error T) of nearest neighbor classifiers when each category of the application task is represented by a *single* example, averaged over 100 runs with randomly chosen examples.

Both of the latter two quantities were also measured for the unrepresented but normalized input pixel vectors (ROC area input, error input). The values obtained using the trace-norm regularization are given in parentheses.

It has been pointed out by a referee that the theory is not exactly applicable to the experiments, because training sets and the test sets are not chosen iid from the same distribution. The experiments were carried out in the present form to illustrate the utility of similarity for transfer. Separating the same data-sets into sets of training- and test pairs would certainly have produced even better results.

6.2 Training and Application Tasks

In some of the transfer experiments the training and application tasks shared a definitive class of invariants, so that similarity of two pattern corresponds the existence of a transformation in the class of invariants, which (roughly) maps one pattern to the other.

Rotation invariant character recognition. Randomly rotated images of printed alpha characters were used for the training set and randomly rotated images of printed digits were used for the test set, the digit 9 being omitted for obvious reasons.

Scale invariant character recognition. Randomly scaled images of printed alpha characters for the training, randomly scaled images of printed digits were used for the test set. Scaling ranged over a factor of 2.

Rotation and scale invariant character recognition. Randomly rotated and scaled characters for training, randomly rotated and scaled images of printed digits were used for the test set, the digit 9 being omitted for obvious reasons. Scaling ranged over a factor of 2, the digit 9 is omitted in the test set.

Spatially rotation invariant object recognition. The COIL100 database contains images of objects rotated about an axis at an angle 60° to the optical axis. Here the invariance transformations which relate similar patterns cannot be explicitly computed from the images. The first 80 objects of the database were taken for training, the remaining 20 for testing.

In the remaining experiments the underlying notion of similarity cannot be explicitly specified and corresponds to a Gestalt-property of the domain in question.

Handwritten character recognition. The images of handwritten alpha characters from the NIST database were used for training, the handwritten digits from the MNIST database for testing.

Face recognition. The first 35 images of the ATT database for training, the last 5 for testing. Unfortunately the ATT database is at the same time very small and very clean and easy, so that the results are not very conclusive. Attempts to obtain the potentially more interesting Purdue database failed.

The images for the first three experiments are available on the web-site www.andreas-maurer.eu, the others are publicly available.

6.3 Results for Transfer

The results are summarized in Tables 2 and 3. The various row headings will be explained in the sequel.

In all experiments the representation T brings an improvement in recognition rates. This improvement is moderate (54% error downto 33% in the case of handwritten characters) to spectacular (72% downto $< 1\%$ for plane rotations). The results on the COIL and ATT databases slightly improve on the corresponding results in Fleuret and Blanchard, (2005) and Chopra et al. (2005), but the margin is so small, that this may well be a statistical artifact. What is more remarkable is that our results are at all comparable, because our method makes no assumption on the specific properties of image data, such as high correlations for neighboring pixels: In contrast to the approaches described in Chopra et al. (2005) and Fleuret and Blanchard (2005), our method would yield the same results if the images were subjected to any fixed but unknown permutation of pixel indices.

type of experiment	rotation invariance	scale invariance	rot.+scale invariance
training set	alpha	alpha	alpha
nr of categories	20	52	20
nr of examples	2000	1560	4000
$\hat{R}(T)$	0.019	0.005	0.058
$\Lambda(T)$	0.074	0.033	0.185
sparsity of T	9	42	7
test set	digits $\setminus$ 9	digits	digits $\setminus$ 9
nr of categories	9	10	9
nr of examples	900	300	1800
$\hat{R}(T)$	0.55	0.061	0.128
ROC area input	0.597	0.69	0.54
ROC area T	0.999 (0.999)	0.995 (0.99)	0.982 (0.972)
error input	0.716	0.508	0.822
error T	0.009 (0.011)	0.019 (0.035)	0.097 (0.127)

Table 2:

type of experiment	spatial rot. invariance	handw. Chars	face recognition
training set	COIL $\leq$80	NIST	ATT 1-35
nr of categories	80	52	35
nr of examples	2880	4160	350
$\hat{R}(T)$	0.003	0.038	0
$\Lambda(T)$	0.024	0.314	0.022
sparsity of T	46	19	35
test set	COIL $\geq$81	MNIST	ATT 36-40
nr of categories	20	10	5
nr of examples	720	10000	50
$\hat{R}(T)$	0.379	0.183	0.045
ROC area input	0.845	0.728	0.934
ROC area T	0.989 (0.984)	0.9 (0.891)	0.997 (0.997)
error input	0.375	0.549	0.113
error T	0.093 (0.123)	0.335 (0.383)	0 (0)

Table 3:

6.4 One Experiment in Detail

To illustrate these experiments and results we will consider the example of combined rotation and scale-invariance, corresponding to the last column in Table 2. Figure 2 shows some typical training examples, of which there are 200 representing each of the 20 categories, making a total number of 4000.

The oracle presents the learner with pairs of these images together with a similarity value $r \in \{-1, 1\}$. Similar pairs are chosen from the same category (the same column in Fig. 2) dis-

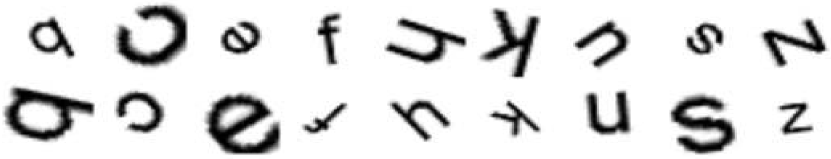

Figure 2: Randomly rotated and scaled alpha-characters

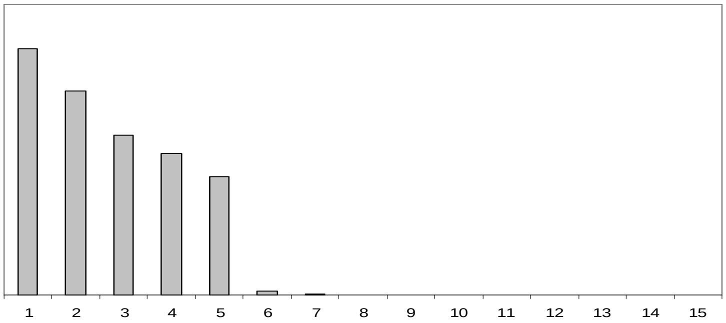

Figure 3: The 15 largest eigenvalues of T^*T in proportion

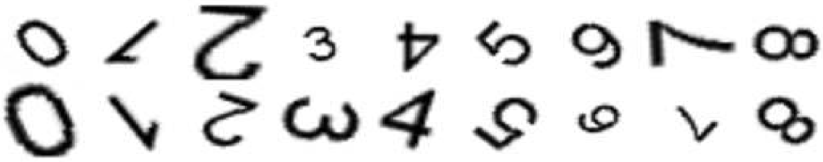

Figure 4: Randomly rotated and scaled digits

similar ones from different categories (different columns). The pairs are chosen at random under the constraint that similar pairs appear with equal frequency as dissimilar ones. These pairs are fed as input to the stochastic gradient descent algorithm in Table 1. While there are $O\left(10^6\right)$ pairs potentially generated, only about 2000 of these can be statistically independent in terms of the generation of the original sample.

The spectrum of T^*T (Fig. 3) of the resulting operator T shows a marked decrease of singular values, allowing the conclusion that the data characterizing rotation and scale invariant character categories in the chosen Gaussian kernel representation is essentially only 5-dimensional. This observed sparsity is not a consequence of the regularization with the Hilbert-Schmidt norm, because an increase in the regularization parameter λ increases the essential rank of T (a different behavior would be expected with a 1-norm regularizer), but an intrinsic property of the data which the algorithm discovers.

The representation T is then applied to the recognition of digits. Fig. 4 exemplifies the test data.

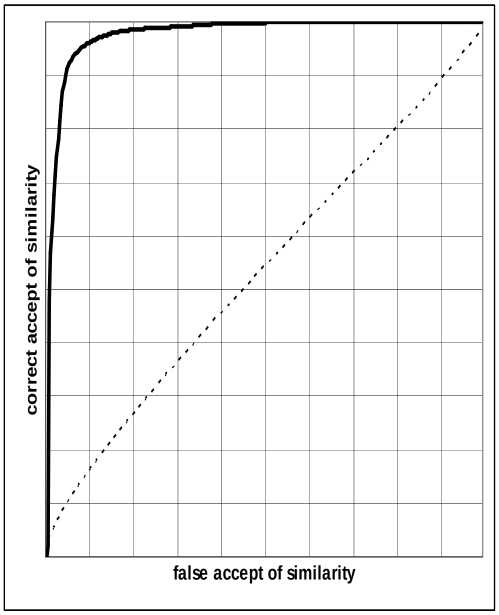

Figure 5: ROC curve for the metric as a feature for similarity. Transformed data solid, input data dotted.

To measure the empirical risk $\hat{R}(T) = \hat{R}_{1_{(-\infty,0]}}(T)$ we generate a random sequence of pairs as we did with the training task above and average the relative rates of dissimilar pairs with $\|Tx - Tx'\|_H \leq 1$ and the rate of similar pairs with $\|Tx - Tx'\|_H \geq 1$. Here the subscript H refers to the distance in the RKHS. This gives the entry 0.128 in the row labeled by 'Risk T', so the representation T organizes the data of rotation and scale invariant character categories into balls of diameter 1, up to an error of about 13%.

To parametrize a potential verification system an ROC curve for the utility of the metric as a feature for class equality is useful. Fig 5 shows these curves dotted for the metric $\|x - x'\|_{28\times 28}$ (where the distance is measured on the raw normalized pixel vectors in $\mathbb{R}^{28\times 28}$) in red and solid for $\|Tx - Tx'\|_H$. The areas under these curves correspond to the rows labeled 'ROC area input' and 'ROC area T' respectively. The values in parentheses correspond to regularization with $\|T\|_2^2$, which givel slightly inferior results.

Finally we measure the performance of a single-nearest neighbor classifier. A prototype is selected randomly from each category. The images in either one of the rows in Fig. 4 could represent such a 'training sample'. We then measure the performance of the corresponding 1-NN classifier on the test set with the training data omitted. The error rates are averaged over 100 random selections of the prototype set. These computations are carried out (with equal prototypes) for the metric $\|x - x'\|_{28\times 28}$ and the metric $\|Tx - Tx'\|_H$ and give the entries in the rows 'error input' and 'error T' respectively. Again the values in parentheses correspond to regularization with $\|T\|_2^2$.

6.5 Classification of Tasks

We report some simple results concerning the recognition of task-families on the basis of the similarity of underlying similarity concepts, as proposed in Section 5.2. In Table 4 we consider task-families with 28 × 28 pixel data. These families share the properties of rotation invariance, scale invariance, combined rotation and scale invariance and the property of being handwritten respectively. The columns correspond to alpha-characters used to train representation, the rows to digits used for testing. Each entry is the empirical risk $\hat{R}(T)$ of the operator T trained from the alpha-task heading the column, measured on the digit-task heading the row. The minimum in each row occurs on the diagonal, which shows that the underlying similarity property or invariance of a data-set is reliably recognized. The margin of this minimum is weakened only for separate rotation and scale invariances, because the representation for combined rotation and scale invariance also performs reasonably well on these data-sets, although not as well as the specialized representations. Scale invariant representations perform better on handwritten data than rotation invariant ones, probably because scale invariance is more related to the latent invariance properties of handwritten characters than full rotation invariance.

	rotation	scaling	rot.+scaling	handwritten
rotation	0.055	0.38	0.089	0.374
scaling	0.36	0.061	0.11	0.304
rot.+scaling	0.375	0.39	0.128	0.434
handwritten	0.4	0.336	0.35	0.18

Table 4:

Given the nature of the algorithm, these results are perhaps not surprising, but they seem to point in interesting new directions for the design of more autonomous systems of pattern recognition.

7. Related Work

A lot of work has been done to develop learning algorithms for data-representations with optimal metric properties. Typically a heuristically derived objective function is optimized, and often there is no discussion of generalization performance for high dimensional input data.

The classical method seems to be *Linear Discriminant Analysis* (LDA, Fukunaga, 1990) which projects onto a dominant subspace of the matrix quotient of the inter-class and intra-class empirical covariances. This can only work if the intra-class covariance operator is non-singular, and it will work poorly even if it is non-singular, but has very small eigenvalues (a generic situation in high dimensions), whence there have been several efforts to remedy the situation by optimization within the null-space of the intra-class covariance operator (NLDA) or by simultaneous diagonalisation of the intra- and inter-class covariances (OLDA). The stability problem inherent to the quotient approach is approached by a heuristic regularization prescription (ROLDA). These various extensions to LDA are presented by Ye and Xiong (2006) and have been tested in an experimental situation where the trained representation is combined with K-NN classification on the same task where the representation was trained. The performance appears to be comparable to SVM.

LDA and its extensions can be best compared to the algorithm of hyperbolic PCA presented in Section 4. In contrast to LDA hyperbolic PCA projects onto a dominant eigenspace of a weighted difference and not the quotient of the inter- and intra-class covariances. For this reason hyperbolic PCA is free of the stability problems of LDA and generalization guarantees are easily obtained.

An interesting technique has been proposed by Goldberger et al (2004). The desired representation is chosen to optimize the performance of a stochastic variant of K-NN classification on the represented data. The method, called *Neighborhood Component Analysis* (NCA) appears to admit a regularized version, with essentially the same regularizer as in this work, and it seems possible to obtain dimension-free generalization guarantees for the regularized version. Unfortunately the optimization problem underlying NCA is not convex.

There is a certain kinship of NCA to the technique presented in this work, because both approaches depart from an objective defined by performance requirements of algorithms operating on the represented data. The stochastic K-NN classifiers of NCA correspond to the elementary verifiers (see Section 5) in our approach.

The problem of similarity learning from pair oracles similar to this paper has been considered by several authors.

In the work of Bar-Hillel et al. (2005) the triplets (x,x',r) generated by the oracle are called equivalence constraints, positive if $r=1$ and negative if $r=-1$. Their algorithm, called RCA for *Relevant Component Analysis*, does not use negative equivalence constraints on the grounds that these are less informative than positive ones, a claim supported by a simple counting argument. The objective of RCA is essentially entropy maximization under the constraint that ”chunklets” of data points belonging to the same class, as inferred from the positive equivalence constraints, remain confined to balls of a fixed diameter. Under Gaussian assumptions there is a bound on the variance of the RCA-estimator, but in general it is unclear if a representation optimizing the objective for one data-set will also be nearly optimal for another one drawn from the same distribution.

Xing et al. (2002) use both positive and negative equivalence constraints and pose the following optimization problem (in our notation) for a sample $((x_1,x_1',r_1),...,(x_m,x_m',r_m))$

$$\min_T \sum_{i:r_i=1} \left\|Tx_i - Tx_i'\right\|^2 \text{ such that } \sum_{i:r_i=-1} \left\|Tx_i - Tx_i'\right\| \geq 1.$$

To solve this problem they propose an algorithm which enforces the positivity constraint for T^*T by alternating gradient descent and projection to the cone of positive operators by eigenvalue decompositions, a method which seems generally applicable when a convex objective is to be optimized under a positivity constraint.

It is surprising that both in Bar-Hillel et al. (2005) and Xing et al. (2002) the existence of a pair oracle to generate a sample of labeled pairs (or equivalence constraints) is implicitly assumed, without attempting to directly predict this oracles behavior. If there is a process which generates labeled examples (and this is what the equivalence constraints are) it seems natural to learn to predict the labels.

This idea has already been proposed by Thrun and Mitchell (1995) (see also Thrun, 1998), where also the obvious connection to transfer and meta-learning is mentioned. This work combines this with the idea of representation learning, as also proposed by Thrun (1998).

Some authors (Chopra et al., 2005; Fleuret and Blanchard 2005) have considered the utility of representation learning for the purpose of multi-category and multi-task pattern recognition on the basis of single training examples. In contrast to the more general method introduced above, these approaches are tailored to the special domain of image processing.

8. Conclusion

This work presented a technique to represent pattern similarity on the basis of data generated by a domain dependent oracle. The method works well in multi-task and multi-category environments as a preparation for future learning with minimal training sets, such as a single training example or a single training example per category.

A major theoretical problem is to explain the good performance of the method in the context of transfer, a phenomenon which doesn't seem to be completely understood.

Another important development would be a learning algorithm of optimal representations for larger future sample sizes. It is conceivable that, in the context of learning-to-learn, the learner can choose from a catalogue of previously trained representations on the basis of the size of the available training sample. The representations trained by the proposed algorithm would then constitute only one extreme entry in this catalogue, corresponding to a minimal sample size of one.

Appendix A. Notation Table

Notation	Short Description	Section
$\mathcal{X}$	input space, $\mathcal{X} \subset H, diam(\mathcal{X}) \leq 1$	1.1
ρ	pair oracle. P-measure on $\mathcal{X}^2 \times \{-1,1\}$	1.1
(x,x',r)	generic triplet $(x,x',r) \in \mathcal{X}^2 \times \{-1,1\}$	1.1
S	training sample $S \in \left(\mathcal{X}^2 \times \{-1,1\}\right)^m, S \sim \rho^m$	1.1
$R(T)$	risk of operator $= \Pr\left\{r\left(1-\|Tx-Tx'\|^2\right) \leq 0\right\}$	1.2
$\hat{R}_\psi(T,S)$	empirical risk estimate for $\psi \geq 1_{(-\infty,0]}$ $= \frac{1}{m}\sum_{i=1}^m \psi\left(r_i\left(1-\|T(x_i-x_i')\|^2\right)\right)$	1.2
$\hat{R}(T,S)$	empirical risk, $\hat{R}(T,S) = \hat{R}_{1_{(-\infty,0]}}(T,S)$	6.1
h_γ	hinge-loss with margin γ	1.2
λ	regularization parameter	3.2
$\Lambda_{\psi,\lambda}(T,S)$	objective functional $\Lambda_{\psi,\lambda}(T,S) = \hat{R}_\psi(T,S) + m^{-1/2}\lambda\|T^*T\|_2$	3.2
Ω	convex objective, $\Omega(T^*T) = \Lambda(T)$	3.3
H	real, separable Hilbert space	2.1
$\langle .,. \rangle$ and $\|.\|$	inner product and norm on H	2.1
$\|T\|_\infty$	operator norm $\|T\|_\infty = \sup_{x \in H, \|x\| \leq 1}\|Tx\|$	2.1
T^*	adjoint of the operator T	2.1
$\mathcal{L}_\infty(H)$	set of operators on H with $\|T\|_\infty < \infty$	2.1
$\mathcal{L}_\infty^*(H)$	$\{T \in \mathcal{L}_\infty(H) : T^* = T\}$	2.1
$\mathcal{L}_\infty^+(H)$	$\{T \in \mathcal{L}_\infty^*(H) : \langle Tx,x \rangle \geq 0, \forall x \in H\}$	2.1
$\mathcal{L}_0(H)$	set of finite-rank operators on H	2.1
$\|T\|_2$	Hilbert-Schmidt norm	2.1
$\mathcal{L}_2(H)$	set of operators on H with $\|T\|_2 < \infty$	2.1
$\langle T,S \rangle_2$	inner product in $\mathcal{L}_2(H)$	2.1
$\mathcal{L}_2^*(H)$, $\mathcal{L}_2^+(H)$	$\mathcal{L}_\infty^*(H) \cap \mathcal{L}_2(H)$ and $\mathcal{L}_\infty^+(H) \cap \mathcal{L}_2(H)$ resp.	2.1
$\mathcal{P}_d$	d-dimensional orthogonal projections in H	2.1
Q_x, for $x \in H$	the operator $Q_x(z) = \langle z,x \rangle x$	2.1
$\|E\|_\mathcal{S}$, $E \subseteq \mathcal{S}$	maximal norm in E, that is, $\sup_{x \in E}\|x\|_\mathcal{S}$	2.1
$\hat{\mathcal{R}}_m(\mathcal{F})$	empirical Rademacher complexity of $\mathcal{F}$	2.2
σ_i	Rademacher variables, uniform on $\{-1,1\}$	2.2

References

M. Anthony and P. Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, 1999.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2006.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

P. Bartlett, O. Bousquet and S. Mendelson. Local Rademacher complexities. Available online: http://www.stat.berkeley.edu/˜bartlett/papers/bbm-lrc-02b.pdf.

A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6: 937–965, 2005.

J. Baxter. Theoretical models of learning to learn. In *Learning to Learn*, S. Thrun and L. Pratt, Eds., Springer, 1998.

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.

R. Caruana. Multitask learning. In *Learning to Learn*, S. Thrun and L. Pratt, Eds., Springer, 1998.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proc. Conference on Knowledge Discovery and Data Mining*, 2004.

F. Fleuret and G. Blanchard. Pattern recognition from one example by chopping. In *Advances in Neural Information Processing Systems*, 2005.

K. Fukunaga. *Introduction to Statistical Pattern Classification.* Academic Press, 1990.

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. In *Advances in Neural Information Processing Systems*, 2004.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1): 1–50, 2002.

A. Maurer. Generalization bounds for subspace selection and hyperbolic PCA. In *Subspace, Latent Structure and Feature Selection. LNCS* 3940: 185–197, Springer, 2006a.

A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006b.

A. Maurer. Learning to compare using operator-valued large-margin classifiers. In *Advances in Neural Information Processing Systems*, 2006c.

C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248. Springer, Berlin, 1998.

M. Reed and B. Simon. Functional analysis. In *Methods of Mathematical Physics,* Academic Press, 1980.

A. Robins. Transfer in cognition. In *Learning to Learn*, S. Thrun and L. Pratt, Eds., Springer, 1998.

J. Shawe-Taylor and N. Christianini. Estimating the moments of a random vector. In *Proceedings of GRETSI 2003 Conference I,* pages 47–52, 2003.

B. Simon. *Trace Ideals and Their Applications*. Cambridge University Press, London, 1979.

S. Thrun and T. M. Mitchell. Learning one more thing. In *Proceedings of IJCAI*, 1995.

S. Thrun. Lifelong learning algorithms. In *Learning to Learn*, S. Thrun and L. Pratt, Eds., Springer, 1998.

E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell. Distance metric learning, with application to clustering with side information. In S. Becker, S. Thrun, and K. Obermayer, Eds., *Advances in Neural Information Processing Systems* 14, MIT Press, Cambridge, MA, 2002.

J. Ye and T. Xiong. Computational and theoretical analysis of null-space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research*, 7:1183–1204, 2006.