Generalization Bounds for Subspace Selection and Hyperbolic PCA

Andreas Maurer

Adalbertstr. 55
D-80799 München
andreasmaurer@compuserve.com

Abstract. We present a method which uses example pairs of equal or unequal class labels to select a subspace with near optimal metric properties in a kernel-induced Hilbert space. A representation of finite dimensional projections as bounded linear functionals on a space of Hilbert-Schmidt operators leads to PAC-type performance guarantees for the resulting feature maps. The proposed algorithm returns the projection onto the span of the principal eigenvectors of an empirical operator constructed in terms of the example pairs. It can be applied to meta-learning environments and experiments demonstrate an effective transfer of knowledge between different but related learning tasks.

1 Introduction

Humans can use the experience accumulated during previous learning efforts to learn novel but related tasks more efficiently, often generalizing well on the basis of a single training example (see e.g. [13]).

Here we present a machine learning algorithm designed to imitate aspects of this behaviour. It attempts to represent input data in a Euclidean space, such that the metric relations between the represented data points match semantic relations of their class labels. The most elementary semantic relations are equality and inequality, often called *equivalence constraints*, and matching these means that pairs of equally labelled input points should be mapped close to each other, while pairs of differently labelled points should be separated. To train the representing feature map such equivalence constraints can be sampled from environments encompassing many individual learning tasks. If the semantic match is good on the training data and the feature map generalizes well, then we can expect that for any - possibly novel - learning task in the environment, a classifier thresholding the distance to a single training example will have good performance.

Similar methods have received some attention recently, both from the perspective of machine learning ([16],[3]) and cognitive science ([6]). Our approach is motivated by a distribution-independent analysis of the generalization performance of elementary classifiers in a meta-learning environment. The proposed algorithm is a subspace selection technique which can be regarded as a hyperbolic extension of PCA. It utilizes both positive (equal labels) and negative (different labels) equivalence constraints.

The method has been tested in various domains of image recognition: Handwritten characters, rotation and scale invariant character recognition and the recognition of human faces. In all these cases the representations trained on one learning task resulted in a considerable performance improvement for small-sample nearest neighbour classifiers on related tasks.

The next section introduces metric threshold classifiers and corresponding risk functionals for general metric representations. Section 3 presents a probabilistic model for the generation of equivalence constraints. Section 4 specializes to the representations considered by our algorithm and gives a high probability generalization guarantee in terms of the empirical properties of a representation. Section 5 is devoted to the proof of this theorem and section 6 discusses some details of our algorithm. Some experimental results are presented in section 7.

2 Risk functionals for metric representations

Suppose that $\mathcal{E}$ is an environment of learning tasks with common input space $\mathcal{X}$ (see Baxter [4]). This means that $\mathcal{E}$ is a probability distribution on a space of learning tasks $\{(\mathcal{Y},\mu)\}$, where each $\mathcal{Y}$ is an alphabet of labels, and each μ is a probability distribution on $\mathcal{X}\times\mathcal{Y}$, $\mu(x,y)$ being the probability to encounter the pattern x carrying the label y in the context of the task $(\mathcal{Y},\mu)$.

We now define a performance measure for metric representations of $\mathcal{X}$ in terms of the expected performance of elementary threshold classifiers. Suppose $\Phi:\mathcal{X}\rightarrow\Phi(\mathcal{X})$ is such a representation in a metric space $(\Phi(\mathcal{X}),d)$, where we assume the diameter of $\Phi(\mathcal{X})$ to be bounded by 1. Consider a learning task $(\mathcal{Y},\mu)$ and a single training example $(x,y)\in\mathcal{X}\times\mathcal{Y}$. A 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 recognition is an environment where such classifiers can be quite important in practice: A police officer having to verify the identity of a person on the basis of a single passport photograph has to learn and generalize on the basis of a single example image. A simple classifier using only the metric representation is the threshold classifier $\epsilon_c(x,y)$ which decides

$$\begin{array}{cl} x' \textit{ is of type } y & \text{if } d(\Phi(x),\Phi(x'))<c \\ \textit{undecided} & \text{if } d(\Phi(x),\Phi(x'))=c\,, \\ x' \textit{ is not of type } y & \text{if } d(\Phi(x),\Phi(x'))>c \end{array}$$

where c is some distance threshold $c\in(0,1)$. Relative to the task $(\mathcal{Y},\mu)$ this classifier has the error probability (counting 'undecided' as an error)

$$\operatorname{err}(\epsilon_c(x,y))=\Pr_{(x',y')\sim\mu}\{r_{\mathcal{Y}}(y,y')(c-d(\Phi(x),\Phi(x')))\leq 0\}\,,$$

where the function $r_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \{0,1\}$ quantifies equality and inequality in $\mathcal{Y}$:

$$r_{\mathcal{Y}}(y,y') = \begin{cases} 1 & \text{if } y = y' \\ -1 & \text{if } y \neq y' \end{cases} .$$

The expected value of $\mathrm{err}(\epsilon_c(x,y))$, as a task $(\mathcal{Y},\mu)$ is selected randomly from the environment $\mathcal{E}$ and a training example (x,y) is chosen from μ, is

$$R(\varPhi,c,\mathcal{E}) = \mathbb{E}_{(\mathcal{Y},\mu)\sim\mathcal{E}}\left[\mathbb{E}_{(x,y)\sim\mu}\left[\Pr_{(x',y')\sim\mu}\{r_{\mathcal{Y}}(y,y')(c - d(\varPhi(x),\varPhi(x'))) \leq 0\}\right]\right]$$

The quantity $R(\varPhi,c,\mathcal{E})$ is a measure of the risk associated with the metric representation $\varPhi$, the assumed threshold c and the environment $\mathcal{E}$. Optimization with respect to c gives the threshold independent risk functional[1]

$$R(\varPhi,\mathcal{E}) = \inf_{c\in(0,1)} R(\varPhi,c,\mathcal{E}) . \tag{1}$$

Our algorithm will seek a metric representation $\varPhi$ with a small value of $R(\varPhi,\mathcal{E})$ where $\varPhi(\mathcal{X})$ is isometrically embedded in $\mathbb{R}^d$. Any bound on $R(\varPhi,\mathcal{E})$ is then also a bound on the expected error of threshold classifiers. This is the theoretical justification of the risk functional R, but it does not imply that we are constrained to use the simple and functionally limited threshold classifiers: Any machine learning algorithm applicable to labelled vectors in $\mathbb{R}^d$ (e.g. NN or SVM) can be used on the data which has been preprocessed by $\varPhi$.

3 Equivalence constraints

A triplet $(x,x',r) \in \mathcal{X}^2 \times \{-1,1\}$ is called an *equivalence constraint* ([3],[6]). Given an environment $\mathcal{E}$ we define a probability measure $\rho_{\mathcal{E}}$ on $\mathcal{X}^2 \times \{-1,1\}$ by the formula

$$\rho_{\mathcal{E}}(A) = \mathbb{E}_{(\mathcal{Y},\mu)\sim\mathcal{E}}\left[\Pr_{((x,y),(x',y'))\sim\mu^2}\{(x,x',r_{\mathcal{Y}}(y,y')) \in A\}\right] \text{ for } A \subseteq \mathcal{X}^2\times\{-1,1\} .$$

To draw an equivalence constraint (x,x',r) from $\rho_{\mathcal{E}}$ we first draw a task $(\mathcal{Y},\mu)$ from $\mathcal{E}$, and then make two independent draws from μ to generate the pair $((x,y),(x',y')) \in (\mathcal{X}\times\mathcal{Y})^2$. If $y = y'$ we set $r = 1$ else we set $r = -1$. We then have

$$R(\varPhi,c,\mathcal{E}) = \Pr_{(x,x',r)\sim\rho_{\mathcal{E}}}\{r(c - d(\varPhi(x),\varPhi(x'))) \leq 0\} . \tag{2}$$

The measure $\rho_{\mathcal{E}}$ is itself unknown to our algorithm, which instead has to rely on a training sample $S = ((x_1,x_1',r_1),...,(x_m,x_m',r_m)) \in \left(\mathcal{X}^2 \times \{-1,1\}\right)^m$ of m equivalence constraints generated in m independent, identical trials of $\rho_{\mathcal{E}}$ according to the above procedure, i.e. $S \sim (\rho_{\mathcal{E}})^m$.

[1] If 'undecided' was not counted as an error, this infimum would always be attained for some distance threshold $c^* \in [0,1]$, which can be regarded as a granularity of the metric representation.

The specific way in which the measure $\rho_{\mathcal{E}}$ was generated served to derive and motivate the risk functional R and is otherwise irrelevant to most of our analysis. We only require a probability measure ρ on $\mathcal{X}^2 \times \{-1,1\}$ and risk functionals R as defined by (2) and (1). There are other interesting ways to generate such measures: As pointed out by Bar-Hillel et al ([3]), equivalence constraints can be generated in an unsupervised way by observing a video sequence, regarding image pairs taken at similar times as positive and pairs at very different times as negative constraints. We will therefore axiomatically postulate the existence of the measure ρ, dropping the subscript which indicated the dependence on the environment $\mathcal{E}$. We also write $R(\varPhi, c, \mathcal{E}) = R(\varPhi, c, \rho_{\mathcal{E}})$ and $R(\varPhi, \mathcal{E}) = R(\varPhi, \rho_{\mathcal{E}})$.

Another important issue here is balancing. If the alphabets in $\mathcal{E}$ are large, with their symbols appearing approximately equally likely, then negative equivalence constraints will be sampled much more frequently than positive ones, resulting in a negative bias of elementary classifiers. This unwanted effect has been noted in [16] and [3]. A simple remedy is to define a new measure $\bar{\rho}_{\mathcal{E}}$ by

$$\bar{\rho}_{\mathcal{E}}(A) = \frac{\rho_{\mathcal{E}}(A \cap \{1\})}{2\rho_{\mathcal{E}}(X \cap \{1\})} + \frac{\rho_{\mathcal{E}}(A \cap \{-1\})}{2\rho_{\mathcal{E}}(X \cap \{-1\})} \text{ for } A \subseteq \mathcal{X}^2.$$

Then positive and negative equivalence constraints occur equally likely as measured by $\bar{\rho}$. The risk $R(\varPhi, c, \bar{\rho}_{\mathcal{E}})$ relative to $\bar{\rho}_{\mathcal{E}}$ is often more relevant than $R(\varPhi, c, \rho_{\mathcal{E}})$. Since our bounds will be valid for any probability measure ρ on $\mathcal{X}^2 \times \{-1,1\}$ they will also work with $\bar{\rho}_{\mathcal{E}}$ as long as we remember that the training sample S is also drawn from the modified measure $S \sim (\bar{\rho}_{\mathcal{E}})^m$.

4 Generalization bounds for subspace selection

Our technique is related to kernel-PCA (see [10], [14]): It requires some fixed map $\psi : \mathcal{X} \to H$ to embed the input data in a Hilbert space H. In practice the embedding ψ is realized by a positive definite kernel κ on the input space which maps onto the inner product $\langle ., . \rangle$ in the Hilbert space H (see [5]). For our results we generally require $\|\psi(x) - \psi(x')\| \leq 1$ for all inputs x and x', and we assume H to be infinite dimensional. On the basis of the training set S a d-dimensional orthogonal projection P on H is selected. The combined map of embedding and projection $\varPhi = P \circ \psi$ is then used as a metric representation for future data. Since ψ is fixed and P is completely determined by its range, our algorithm can also be considered a *subspace selection technique.*

In the following we fix the Hilbert space H and simply write x instead of $\psi(x)$, identifying $\mathcal{X}$ with its image $\psi(\mathcal{X}) \subset H$ under the kernel-map. When we discuss details of our algorithm we bring ψ back into play. It is crucial that $\operatorname{diam}(\mathcal{X}) \leq 1$.

For subspace selection the risk functionals in (2) and (1), which now depend on the projection P, read as

$$\begin{aligned} R(P,c,\rho) &= \Pr_{(x,x',r)\sim\rho}\{r(c-\|P(x-x')\|)\leq 0\} \\ R(P,\rho) &= \inf_{c\in(0,1)} R(P,c,\rho). \end{aligned}$$

To write down a sample dependent bound on R we introduce for $\gamma > 0$ the margin functions

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

and the empirical margin error $\hat{R}_\gamma$ for a sample $S = ((x_1,x_1',r_1),...,(x_m,x_m',r_m)) \in (\mathcal{X}^2\times\{-1,1\})^m$, a threshold $c>0$ and a d-dimensional projection P

$$\hat{R}_\gamma(P,c,S) = \frac{1}{m}\sum_{i=1}^m f_\gamma\left(r_i\left(c^2-\|P(x_i-x_i')\|^2\right)\right).$$

Recent results on large margin classifiers (Kolchinskii and Panchenko [7], Bartlett and Mendelson [1]), combined with a reformulation in terms of Hilbert-Schmidt operators give the following:

Theorem 1. *Fix $\gamma > 0$. For every $\delta > 0$ we have with probability greater than $1-\delta$ in a sample S drawn from ρ^m, that for every d-dimensional projection P*

$$R(P,\rho) \leq \inf_{c\in(0,1)} \hat{R}_\gamma(P,c,S) + \frac{1}{\sqrt{m}}\left(\frac{2\left(\sqrt{d}+1\right)}{\gamma} + \sqrt{\frac{\ln(1/\delta)}{2}}\right).$$

In addition to the empirical error the bound shows an estimation error, decreasing as $1/\sqrt{m}$ which is usual for this type of bound. The estimation error contains two terms: The customary dependence on the confidence parameter δ and a complexity penalty consisting of $1/\gamma$ (really the Lipschitz constant of the margin function f_γ), and the penalty $\sqrt{d}$ on the dimension of the representing projection.

We will outline a proof of a more general version of this theorem in the next section.

5 Operator-valued linear large-margin classifiers

In this section we rewrite finite dimensional projections and more general feature maps as operator valued large-margin classifiers, and use this formulation to prove a more general version of Theorem 1. We will use the following general result on linear large margin classifiers (Kolchinskii and Panchenko [7], Bartlett and Mendelson [1]):

Theorem 2. *Let (Ω, μ) be a probability space, H a Hilbert space with unit ball $B_1(H)$ and $(w, y) : \Omega \to B_1(H) \times \{-1, 1\}$ a random variable.*

Let $\Lambda \subset H$ be a set of vectors and write

$$B_\Lambda = \sup_{v \in \Lambda} \|v\| \text{ and } C_\Lambda = \sup_{v \in \Lambda, \omega \in \Omega} |\langle w(\omega), v \rangle| .$$

Fix $\gamma, \delta \in (0, 1)$. Then with probability greater than $1 - \delta$ in $S = (\omega_1, ..., \omega_m)$ drawn from μ^m we have for every $v \in \Lambda$ and every t with $|t| \leq C_\Lambda$

$$\begin{aligned} &\Pr_{\omega \sim \mu} \{y(\omega)(\langle w(\omega), v\rangle - t) \leq 0\} \\ &\leq \frac{1}{m} \sum_{i=1}^m f_\gamma(y(\omega_i)(\langle w(\omega_i), v\rangle - t)) + \frac{1}{\sqrt{m}} \left(\frac{2(B_\Lambda + C_\Lambda)}{\gamma} + \sqrt{\frac{\ln(1/\delta)}{2}} \right). \end{aligned}$$

The theorem as stated has an improved margin dependent term by a factor of 2 over the results in [1]. This results from using a slightly different definition of Rademacher complexity with a correspondingly improved bound on the complexity of function classes obtained from compositions with Lipschitz functions (Theorem A6 in [2]).

For a fixed Hilbert space H we now define a second Hilbert space consisting of *Hilbert-Schmidt operators*. With HS we denote the real vector space of symmetric operators on H satisfying $\sum_{i=1}^\infty \|Te_i\|^2 \leq \infty$ for every orthonormal basis $(e_i)_{i=1}^\infty$ of H. For $S, T \in HS$ 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 $\langle S, T \rangle_{HS} = \sum \langle Se_i, Te_i \rangle$ defines an inner product on HS, making it into a Hilbert space. We denote the corresponding norm with $\|.\|_{HS}$ (see Reed and Simon [12] for background on functional analysis).

We use HS_+ to denote the set of *positive* Hilbert-Schmidt operators,

$$HS_+ = \{T \in HS : \langle Tv, v \rangle \geq 0 \text{ for all } v \in H\} .$$

Then HS_+ is a closed convex cone in HS. Every $T \in HS_+$ has a unique positive squareroot, which is a bounded operator $T^{1/2}$ (in fact $T^{1/2} \in HS_+$) such that $T = T^{1/2} T^{1/2}$.

For every $v \in H$ we define an operator Q_v by $Q_v w = \langle w, v \rangle v$. For $v \neq 0$ chose an orthonormal basis $(e_i)_1^\infty$, so that $e_1 = v / \|v\|$. Then

$$\|Q_v\|_{HS}^2 = \sum_i \|Q_v e_i\|^2 = \|Q_v v\|^2 / \|v\|^2 = \|v\|^4 ,$$

so $Q_v \in HS_+$ and $\|Q_v\|_{HS} = \|v\|^2$. With the same basis we have for any $T \in HS$

$$\langle T, Q_v \rangle_{HS} = \sum_i \langle Te_i, Q_v e_i \rangle = \langle Tv, Q_v v \rangle / \|v\|^2 = \langle Tv, v \rangle .$$

For $T \in HS_+$ we then have

$$\langle T, Q_v \rangle_{HS} = \left\| T^{1/2} v \right\|^2 . \tag{3}$$

The set of d-dimensional, orthogonal projections in H is denoted with $\mathcal{P}_d$. We have $\mathcal{P}_d \subset HS_+$ and if $P \in \mathcal{P}_d$ then $\|P\|_{HS} = \sqrt{d}$ and $P^{1/2} = P$.

Consider the feature map given by the operator $T^{1/2}$, where T is any operator in HS_+ (this corresponds to the metric $d\left(.,.\right)_T$ considered in [16]). Its threshold dependent risk is

$$\begin{aligned} R\left(T^{1/2}, c, \rho\right) &= \Pr_{(x,x',r)\sim\rho}\left\{r\left(c^2 - \left\|T^{1/2}\left(x - x'\right)\right\|^2\right) \leq 0\right\} \\ &= \Pr_{(x,x',r)\sim\rho}\left\{r\left(c^2 - \langle T, Q_{x-x'}\rangle_{HS}\right) \leq 0\right\}, \end{aligned}$$

where we used the key formula (3). For a margin $\gamma > 0$ and a sample $S = \left(\left(x_1, x_1', r_1\right), ..., \left(x_m, x_m', r_m\right)\right)$ we define the empirical margin-error

$$\begin{aligned} \hat{R}_\gamma\left(T^{1/2}, c, S\right) &= \frac{1}{m}\sum_{i=1}^m f_\gamma\left(r_i\left(c^2 - \left\|T^{1/2}\left(x_i - x_i'\right)\right\|^2\right)\right) \\ &= \frac{1}{m}\sum_{i=1}^m f_\gamma\left(r_i\left(c^2 - \left\langle T, Q_{x_i - x_i'}\right\rangle_{HS}\right)\right). \end{aligned}$$

It is clear that the definitions of R and $\hat{R}$ coincide with those used in Theorem 1 when P is a finite dimensional orthogonal projection. These definitions are also analogous to the risk and empirical margin errors for classifiers obtained by thresholding bounded linear functionals as in Theorem 2. This leads to

Theorem 3. *Let $\mathcal{T}$ be some class of positive symmetric linear operators on H and denote*[2]

$$\|\mathcal{T}\|_{HS} = \sup_{T\in\mathcal{T}}\|T\|_{HS} \text{ and } \|\mathcal{T}\|_\infty = \sup_{T\in\mathcal{T}}\|T\|_\infty.$$

Fix $\gamma > 0$. Then for every $\delta > 0$ we have with probability greater than $1 - \delta$ in a sample $S \sim \rho^m$, that for every $T \in \mathcal{T}$ and every $c \in \left(0, \|\mathcal{T}\|_\infty^{1/2}\right)$

$$R\left(T^{1/2}, c, \rho\right) \leq \hat{R}_\gamma\left(T^{1/2}, c, S\right) + \frac{1}{\sqrt{m}}\left(\frac{2\left(\|\mathcal{T}\|_{HS} + \|\mathcal{T}\|_\infty\right)}{\gamma} + \sqrt{\frac{\ln\left(1/\delta\right)}{2}}\right).$$

Theorem 1 follows immediately from setting $\mathcal{T} = \mathcal{P}_d$, since $\|\mathcal{P}_d\|_{HS} = \sqrt{d}$ and $\|\mathcal{P}_d\|_\infty = 1$.

Proof. Note that for (x, x', r) in the support of ρ we have $\|Q_{x-x'}\|_{HS} = \|x - x'\|^2 \leq 1$, so we can apply Theorem 2 with $\Omega = \mathcal{X}^2 \times \{-1, 1\}$, $\mu = \rho$, $H = HS$, $w\left(x, x', r\right) = -Q_{x-x'}$, and $y\left(x, x', r\right) = r$ and $\Lambda = \mathcal{T}$. Then $B_\Lambda = \|\mathcal{T}\|_{HS}$ and $C_\Lambda \leq \|\mathcal{T}\|_\infty$. Substitution of the expressions for R and $\hat{R}_\gamma$ in the bound of Theorem 2 gives Theorem 3.□

[2] Here $\|T\|_\infty = \sup_{\|v\|=1}\|Tv\|$ is the usual operator norm (see [12]).

6 Hyperbolic PCA

Fix a margin γ and a training sample $S = ((x_1, x_1', r_1), ..., (x_m, x_m', r_m))$ of equivalence constraints. Since there are no other sample dependent terms in the bound of Theorem 1, we should in principle minimize the empirical margin-error

$$\hat{R}_\gamma (P, c, S) = \frac{1}{m} \sum_{i=1}^{m} f_\gamma \left(r_i \left(c^2 - \langle P, Q_{x_i - x_i'} \rangle_{HS} \right) \right).$$

over all choices of $c \in (0,1)$ and $P \in \mathcal{P}_d$, to obtain a (nearly) optimal projection P^* together with some clustering granularity c^*.

This algorithm is difficult to implement in practice. One obstacle is the nonlinearity of the margin functions f_γ. Replacing the f_γ by the convex hinge-loss does not help, because the set of d-dimensional projections itself fails to be convex. Replacing the set $\mathcal{P}_d$ of candidate maps by the set of positive operators with a uniform bound B on their Hilbert-Schmidt norms and replacing the f_γ by any convex function such as the hinge-loss results in a convex optimization problem. Its solution would be the most direct way to exploit Theorem 3 (taking us outside the domain of subspace selection). A major difficulty here is the positivity constraint on the operators chosen. It can be handled by a gradient-descent/projection technique as in [16], but this is computationally expensive, necessitating an eigen-decomposition at every projection step.

Here we take a different path, remaining in the domain of subspace selection. Fix $c \in (0,1)$ and $\gamma > 0$ and for $i \in \{-1, 1\}$ define numbers η_i by $\eta_{-1} = \min \left\{ \frac{1}{1-c^2}, \frac{1}{\gamma} \right\}$ and $\eta_1 = -\min \left\{ \frac{1}{c^2}, \frac{1}{\gamma} \right\}$. Define the empirical operator $\hat{T}(\eta, S)$ by

$$\hat{T}(\eta, S) = \frac{1}{m} \sum_{i=1}^{m} \eta_{r_i} Q_{x_i - x_i'}.$$

Then

$$\begin{aligned} \hat{R}_\gamma (P, c, S) &\leq \frac{1}{m} \sum_{i=1}^{m} \left(1 + \eta_{r_i} \left(c^2 - \langle Q_{x_i - x_i'}, P \rangle_{HS} \right) \right) \\ &= 1 + \frac{c^2}{m} \sum_{i=1}^{m} \eta_{r_i} - \left\langle \hat{T}(\eta, S), P \right\rangle_{HS}. \end{aligned}$$

The right hand side above is the smallest functional dominating $\hat{R}_\gamma$ and affine in the $Q_{x_i - x_i'}$. Minimizing it over $P \in \mathcal{P}_d$ is equivalent to maximizing $\left\langle \hat{T}(\eta, S), P \right\rangle_{HS}$ and constitutes the core step of our algorithm where it is used to generate candidate pairs (P, c) to be tried in the bound of Theorem 1, leading to a heuristic minimization of $\hat{R}_\gamma (P, c, S)$ for different values of $c \in (0,1)$. Current work seeks to replace this heuristic by a more systematic boosting scheme.

Maximization of $\left\langle \hat{T}\left(\eta, S\right), P\right\rangle_{HS}$ is carried out by solving the eigenvalue problem for $\hat{T}$ and taking for P the projection onto the span of the d eigenvectors corresponding to the largest eigenvalues of $\hat{T}$. This is similar to the situation for PCA, where the empirical operator approximating the covariance operator is

$$\hat{C}\left(S\right) = \frac{1}{m}\sum_{i=1}^{m} Q_{x_i}$$

and the x_i are the points of an unlabeled sample. The essential difference to PCA is that while $\hat{C}$ is a positive operator, the operator $\hat{T}$ is not, and it can have negative eigenvalues. The infinite dimensionality of H and the finite rank of $\hat{T}$ ensure that there is a sufficient supply of eigenvectors with nonnegative eigenvalues. Nevertheless, while the level sets of the quadratic form defined by $\hat{C}$ are always ellipsoids, those of $\hat{T}$ are hyperboloids in general, due to the contributions of positive equivalence constraints. In a world of acronyms our algorithm should therefore be called HPCA, for *hyperbolic principal component analysis.* If there are only negative equivalence constraints our algorithm is essentially equivalent to PCA.

To describe in more detail how the method works, we put the kernel map ψ back into the formulation. The empirical operator then reads

$$\hat{T}\left(\eta, S\right) = \frac{1}{m}\sum_{i=1}^{m} \eta_{r_i} Q_{\psi(x_i)-\psi\left(x_i'\right)}.$$

Clearly an eigenvector w of $\hat{T}$ must be in the span of the $\left\{\psi\left(x_i\right) - \psi\left(x_i'\right)\right\}_{i=1}^{m}$, so we can write

$$w = \sum_{i=1}^{m} \alpha_i \left(\psi\left(x_i\right) - \psi\left(x_i'\right)\right). \tag{4}$$

Substitution in the equation $\hat{T}\left(\eta, S\right) w = \lambda w$ and taking inner product with $\left(\psi\left(x_j\right) - \psi\left(x_j'\right)\right)$ gives the generalized matrix-eigenvalue problem

$$\begin{aligned} \mathit{\Gamma} D \mathit{\Gamma} \boldsymbol{\alpha} &= \lambda \mathit{\Gamma} \boldsymbol{\alpha} \\ \mathit{\Gamma}_{ij} &= \left\langle \psi\left(x_i\right) - \psi\left(x_i'\right), \psi\left(x_j\right) - \psi\left(x_j'\right)\right\rangle \\ D_{ij} &= \eta_{r_i}\delta_{ij}. \end{aligned}$$

Evidently all these quantities can be computed from the kernel matrix $\left\langle \psi\left(x_i\right), \psi\left(x_j\right)\right\rangle$. The d solutions $\boldsymbol{\alpha}_k = \left(\alpha_i\right)_k$ corresponding to the largest eigenvalues λ are substituted in (4), the resulting vectors w_k are normalized and the projection corresponding to largest eigenvalues is computed. Notice how this algorithm resembles PCA if there are only negative equivalence constraints, because then D becomes the identity matrix.

There is an interesting variant of this method, which is useful in practice even though it does not completely fit the probabilistic framework described above. Suppose we are given an ordinary sample of labelled data $S =$

$((x_1, y_1), ..., (x_m, y_m))$ and we want to exploit *all* the equivalence constraints implied by S, that is to maximize $\left\langle \hat{T}\left(\eta, S^{(2)}\right), P\right\rangle_{HS}$ with $S^{(2)} = ((x_i, x_j, r(y_i, y_j)))_{i \neq j}$. One might be led to think that this would require solving the eigenvalue problem of an $m^2 \times m^2$-matrix, which would be of order m^6, making it computationally impractical even for moderate sample sizes. The problem may however be reduced to the eigenvalue problem of an $m \times m$-matrix, thus of order m^3:

The empirical operator now reads (with $r_{ij} = r(y_i, y_j)$)

$$\hat{T}\left(\eta, S^{(2)}\right) = \frac{1}{m^2} \sum_{i,j=1}^{m} \eta_{r_{ij}} Q_{\psi(x_i) - \psi(x_j)}.$$

Substituting an eigenvector $w = \sum_1^m \gamma_k \psi(x_k)$ in the eigenvalue-equation and taking the inner product with some $\psi(x_l)$, and using the fact that the matrix $a_{ij} = \eta_{r_{ij}}/m^2$ is symmetric we get

$$\begin{aligned}
&\lambda \sum_{k=1}^{m} \gamma_k \left\langle \psi(x_k), \psi(x_l)\right\rangle \\
&= \frac{1}{m^2} \sum_{i,j=1}^{m} \eta_{r_{ij}} \left\langle Q_{\psi(x_i) - \psi(x_j)} w, \psi(x_l)\right\rangle \\
&= \sum_{k=1}^{m} \gamma_k \sum_{i,j=1}^{m} a_{ij} \left\langle \psi(x_k), \psi(x_i) - \psi(x_j)\right\rangle \left\langle \psi(x_i) - \psi(x_j), \psi(x_l)\right\rangle \\
&= 2 \sum_{k=1}^{m} \gamma_k \sum_{ij=1}^{m} \left(\delta_{ij} \sum_{n=1}^{m} a_{in} - a_{ij}\right) \left\langle \psi(x_k), \psi(x_i)\right\rangle \left\langle \psi(x_j), \psi(x_l)\right\rangle.
\end{aligned}$$

Using G to denote the ordinary Gramian or kernel-matrix $G_{ij} = \langle \psi(x_i), \psi(x_j)\rangle$ we again obtain a generalized $m \times m$ eigenvalue problem

$$GAG\boldsymbol{\gamma} = \lambda G\boldsymbol{\gamma}.$$

A is not diagonal in this case, but given by the symmetric matrix

$$A_{ij} = 2\left(\delta_{ij} \sum_{n=1}^{m} a_{in} - a_{ij}\right) = \frac{2}{m^2}\left(\delta_{ij} \sum_{n=1}^{m} \eta_{r_{in}} - \eta_{r_{ij}}\right).$$

The sample $S^{(2)}$ does not fit into our probabilistic framework, because it has not been generated by m^2 *independent* draws of equivalence constraints, in fact only $O(m)$ of the pairs in S' can be independent. We nevertheless used this variant of the algorithm to exploit all the information in the training samples for the experiments reported below. The worst possible effect of the use of $S^{(2)}$ is that the number m^2 of equivalence constraints must be replaced by m in our bounds.

7 Experiments

The experiments are designed to test the transfer capabilities of our subspace selection algorithm: We use the data of one set of learning tasks to train a projection, and then check how it facilitates the learning of a new and unknown task.

In practice we take a sample S from a single multiclass learning task with alphabet $\mathcal{Y}$ (this could easily be extended to a collection of tasks) and employ the algorithm described at the end of the previous section to generate projections from all the equivalence constraints implied by S for different values $\eta_1$and η_{-1}, selecting the projection P^* giving the smallest empirical risk $\hat{R}_{0.01}\left(P^*, S^{(2)}\right)$. The optimal values are reported below for each experiment[3]. Here the balanced version of the risk is used to eliminate the effects of alphabet-sizes.

The projection P^* is applied to a *target task* (with alphabet $\mathcal{Y}'$) for which a test-sample S' is available. The empirical distribution of S' is used to estimate the balanced risk R^* of P^* in the new task (reported below for each case, together with the optimal distance threshold c^*). In addition the feature map is tested with nearest neighbour classification: From S' a *single* example per class is chosen as training data for a nearest neighbour classifier and the error rate of this classifier is recorded for both the metric induced by the feature map (projected data) and the original Euclidean metric on normalized pixel vectors (raw data). This experiment is repeated over all possible choices of training data (in the manner of a leave-$(n-1)$-out test) and the resulting error rates are reported.

The pixel vectors were normalized to unit length. The raw data below already refers to these unit vectors. The embedding ψ was realized by the RBF-kernel $\kappa(x,y) = 2^{-1}\exp\left(-C\,|x-y|^2\right)$, with $C = 16$ for the handwritten digits and $C = 8$ in all other cases [4]. Note that the normalization of the kernel is chosen to bound the diameter of the embedded input vectors by 1, as required by our bounds.

We tried five learning environments, two realistic ones involving handwritten characters and face recognition, and three slightly artificial ones defined by the respective invariances of rotation, scaling and combined rotation and scaling.

For handwritten characters we used images of upper and lower case *letters* in the NIST database to train P^*, and a subset of the MNIST database of *digits* for testing. For face recognition we used the images of 31 subjects in the AT&T Face-Database for training and the remaining 9 subjects for testing.

For rotation invariant character recognition randomly rotated images of printed lower case letters were used for training, randomly rotated images of printed digits (with '9' omitted) for testing. For scale invariant character recognition randomly scaled (from 50% to 150%) images of printed capitals and lowercase

[3] Theorem 1 overestimates the estimation error. This is why a small value for γ is chosen, even though this may make the bound of Theorem 1 trivial.

[4] Here and in the definition of the kernel $|.|$ refers to the euclidean norm of the pixel vectors.

letters were used for training, randomly scaled images of printed digits for testing. For combined rotation and scale invariant character recognition the images in the rotation invariant dataset were also randomly scaled (from 50% to 150%). Again the projection was trained from the letters and tested with digits. The following table summarizes the results of these experiments:

	handw. chars	faces	rotated chars	scaled chars	rotated +scaled chars
$\lvert\mathcal{Y}\rvert$ (Training task)	52	31	20	44	20
$\lvert S\rvert$	4160	310	2000	1320	4000
$\lvert\mathcal{Y}'\rvert$ (Testing task)	10	9	9	10	9
$\lvert S'\rvert$	500	90	900	300	1800
$d = \dim(P^*)$	24	20	18	24	18
η_{-1} ($\eta_1 = 1$, balanced)	0.052	0.016	0.019	0.22	0.19
R^* balanced	0.188	0.05	0.022	0.02	0.068
c^* (balanced)	0.26	0.45	0.3	0.36	0.25
1-NNError on raw data	0.549	0.116	0.716	0.472	0.803
1-NNError on projected data	0.318	0.043	0.014	0.008	0.072

Table 1. Summary of experimental results.

The classification error on the projected data correlates well with the risk R^* and the projection leads to a significant improvement in all cases, handwritten character recognition being the most difficult environment. In the case of face recognition the data set used to train the projection is rather small and further improvements are to be expected for larger, perhaps more difficult data sets than AT&T. In the cases, where the environment corresponds to a class of specific geometric invariances, the projection spectacularly reduces the classification error by orders of magnitude.

It seems promising to extend these experiments to the recognition of spatially rotated objects. A very interesting possible line of possible experiments involves unsupervised learning through the observation of a continuous process. A pair consisting of the present observable vector and a recent memory would be treated as a positive equivalence constraint, a pair of the current vector and a distant memory a negative one. A correspondingly trained projection should map temporal proximity to spatial proximity in its feature space. The observation of continuously and quickly rotating objects which are occasionally being replaced could then lead to a nearly rotation invariant preprocessor. Some experiments pointing in a similar direction have been made by Bar-Hillel et al [3].

References

1. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 2002.

2. P.Bartlett, O.Bousquet and S.Mendelson. Local Rademacher complexities. Available online: http://www.stat.berkeley.edu/~bartlett/papers/bbm-lrc-02b.pdf.
3. A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall. Learning a Mahalanobis Metric from Equivalence Constraints. *Journal of Machine Learning Research* 6: 937-965, 2005.
4. J.Baxter, A Model of Inductive Bias Learning, *Journal of Artificial Intelligence Research* 12: 149-198, 2000
5. Nello Cristianini and John Shawe-Taylor, Support Vector Machines, *Cambridge University Press*, 2000.
6. R. Hammer, T. Hertz, S. Hochstein, D. Weinshall. Category learning from equivalence constraints. XXVII Conference of Cognitive Science Society (CogSci2005), available online.
7. V. Koltchinskii and D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *The Annals of Statistics*, Vol. 30, No 1, 1-50.
8. M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.
9. Colin McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.
10. S.Mika, B.Schölkopf, A.Smola, K.-R.Müller, M.Scholz and G.Rätsch. Kernel PCA and De-noising in Feature Spaces, in *Advances in Neural Information Processing Systems* 11, 1998.
11. J. Shawe-Taylor, N. Christianini, Estimating the moments of a random vector, *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.
12. Michael Reed and Barry Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
13. A. Robins, Transfer in Cognition, in *Learning to Learn*, S. Thrun, L. Pratt Eds. Springer 1998.
14. John Shawe-Taylor, Christopher K. I. Williams, Nello Cristianini, Jaz S. Kandola: On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory* 51(7): 2510-2522, 2005
15. S.Thrun, Lifelong Learning Algorithms, in *Learning to Learn*, S.Thrun, L.Pratt Eds. Springer 1998
16. E. P. Xing, A. Y. Ng, M. I. Jordan, S. Russel. Distance metric learning, with application to clustering with side information. In S. Becker, S. Thrun, K. Obermayer, eds, *Advances in Neural Information Processing Systems* 14, Cambridge, MA, 2002. MIT Press.