Unsupervised Slow Subspace-Learning from Stationary Processes

Andreas Maurer

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

Abstract. We propose a method of unsupervised learning from stationary, vector-valued processes. A projection to a low-dimensional subspace is selected on the basis of an objective function which rewards data-variance and penalizes the variance of the velocity vector, thus exploiting the short-time dependencies of the process. We prove bounds on the estimation error of the objective in terms of the β-mixing coefficients of the process. It is also shown that maximizing the objective minimizes an error bound for simple classification algorithms on a generic class of learning tasks. Experiments with image recognition demonstrate the algorithms ability to learn geometrically invariant feature maps.

1 Introduction

Some work has been done to extend the results of learning theory from independent, identically distributed input variables to more general stationary processes ([11], [23], [10], [20], [5]). For suitably mixing processes this extension is possible, with an increase in sample complexity caused by dependencies which slow down the estimation process. But some of these dependencies also provide important information on the environment generating the process and can be turned from a curse to a blessing, in particular in the case of unsupervised learning, when side information is scarce and the sample complexity is not as painfully felt.

Consider a stationary stochastic process modeling the evolution of a complex environment by a sequence S_t of random variables, taking values in some set $\mathbf{\Omega}$ of states. A realization $S_t = s \in \mathbf{\Omega}$ entails complete knowledge of the environmental state at time t.

In practice the information defining a state is not available to the learner. Instead there is a sensory system ϕ which maps any state s to a stimulus $\phi(s)$ in some linear space H of stimuli, which we assume to be a real separable Hilbert space. The stationary stochastic process

$$X = \{X_t\}_{t \in \mathbb{Z}} \text{ with } X_t = \phi(S_t)$$

models the evolution of stimuli and is accessible to the learner. We will assume the sensory system to be bounded, in the sense that $\|\phi(s)\| \leq 1/2$ and centered relative to S_t in the sense that $\mathbb{E}[X_t] = \mathbb{E}[\phi(S_t)] = 0$.

The representation of a state s by the stimulus $\phi(s)$ is burdened with potentially irrelevant information and one seeks to find a more concise and efficient description. Let $\mathcal{P}_d$ be the class of d-dimensional orthogonal projections in H. From observation of $S_0, ..., S_m$ the learner searches for some $P \in \mathcal{P}_d$ such that the composed map $P \circ \phi$ provides an optimal perspective on the state-space $\mathbf{\Omega}$. To guide this search we will invoke two principles of common sense.

The first principle states that *relevant signals should have a large variance.* In view of the zero-mean assumption this classical idea suggests to maximize $\mathbb{E}\left[\|P\phi(S_0)\|^2\right] = \mathbb{E}\left[\|PX_0\|^2\right]$. This coincides with the objective of PSA[1] ([12], [14], [19], [24]) to give the perspective with the broadest view of the distribution.

The second principle, the principle of *slowness* (introduced by Földiak [4], promoted and developed by Wiskott and Sejnowski [21]), states that *sensory signals vary more quickly than relevant state properties.* Consider the visual impressions caused by a familiar complex object, like a tree on the side of the road or a person acting in a movie. Any motion or deformation of the object will cause rapid changes in the states of retinal photoreceptors (or pixel-values). Yet the identities of the tree and the person in the movie remain unchanged. When a person speaks, the communicated ideas vary much more slowly than individual phonemes, let alone the air pressure amplitudes of the transmitted sound signal.

The slowness principle suggests to minimize $\mathbb{E}\left[\|P\phi(S_0) - P\phi(S_{-1})\|^2\right] = \mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right]$ (here $\dot{X}$ is the velocity process $\dot{X}_t = X_t - X_{t-1}$), and combining both principles leads to the objective function

$$L_\alpha(P) = \mathbb{E}\left[\alpha\|PX_0\|^2 - (1-\alpha)\left\|P\dot{X}_0\right\|^2\right],$$

to be maximized, where the parameter $\alpha \in [0,1]$ controls the trade-off between two potentially conflicting goals. In section 4 we will further justify the use of this objective function and show that for $\alpha \in (0,1)$ maximizing L_α minimizes an error bound for a simple classification algorithm on a generic class of classification problems, and that $\sqrt{\alpha}$ can be interpreted as a typical scale of semantic clusters. When there is no ambiguity we write $L = L_\alpha$.

As the details of the process X are generally unknown, the optimization has to rely on an empirical basis. Let $(X)_0^m = (X_0, ..., X_m)$ be $m+1$ consecutive observations of the process X and define an empirical analogue $\hat{L}(P)$ of the objective function L

$$\hat{L}(P) = \frac{1}{m}\sum_{i=1}^{m}\left(\alpha\|PX_i\|^2 - (1-\alpha)\left\|P\dot{X}_i\right\|^2\right).$$

[1] Principal Subspace Analysis, sometimes Principal Component Analysis (PCA) is used synonymously

We now propose to select $P \in \mathcal{P}_d$ to maximize $\hat{L}(.)$. This optimization problem, its analysis, algorithmic implementation and some experimental results are the contributions of this paper.

Existence of Solutions. Given the observations $(X_0, ..., X_m)$ of the process, how do we choose P to maximize the empirical objective functional $\hat{L}(.)$? Fix some trade-off parameter $\alpha \in [0,1]$ and define an operator T on H by

$$Tz = \mathbb{E}\left[\alpha \langle z, X\rangle X - (1-\alpha)\left\langle z, \dot{X}\right\rangle \dot{X}\right] \text{ for } z \in H. \tag{1}$$

Then $T = \alpha C_X - (1-\alpha) C_{\dot{X}}$, where C_X and $C_{\dot{X}}$ are the covariance operators corresponding to X and $\dot{X}$ respectively. The empirical counterpart to T is $\hat{T}$ defined by

$$\hat{T}z = \frac{1}{m}\sum_{i=1}^{m}\left(\alpha \langle z, X_i\rangle X_i - (1-\alpha)\left\langle z, \dot{X}_i\right\rangle \dot{X}_i\right). \tag{2}$$

The operators T and $\hat{T}$ are central objects of the proposed method. They are both symmetric and compact, T is trace-class and $\hat{T}$ has finite rank. If $\alpha \in (0,1)$ they will tend to have both positive and negative eigenvalues. The following theorem (see section 2.2) shows that a solution of our optimization problem can be obtained by projecting onto a dominant eigenspace of $\hat{T}$.

Theorem 1. *Suppose that $\alpha \in [0,1]$, that there are d eigenvalues $\lambda_1, ..., \lambda_d$ of $\hat{T}$ (counting multiplicities) such that $\lambda_i \geq \lambda$ for all other eigenvalues λ of $\hat{T}$, and that (e_i) is the sequence of associated eigenvectors. Then*

$$\max_{P \in \mathcal{P}_d} \hat{L}(P) = \sum_{i=1}^{d} \hat{\lambda}_i,$$

the maximum being attained when P is the orthogonal projection onto the span of $e_1, ..., e_d$.

This leads to a straightforward batch algorithm: Observe and store a realization of $(X_0, ..., X_m) = (\phi(S_0), ..., \phi(S_m))$, construct a matrix for $\hat{T}$, find eigenvectors and eigenvalues and project onto the span of d orthonormal eigenvectors corresponding to the largest eigenvalues.

Such a solution P need not be unique. In fact, if $\alpha = 0$ and $\dim(H) = \infty$, then $\hat{T}$ is a non-positive operator with infinite dimensional null-space, and there is an infinity of mutually orthogonal solutions, from which an arbitrary choice must be made. This can hardly be the way to extract meaningful signals, and the utility of the objective function with $\alpha = 0$ is questionable for high-dimensional input spaces. Except for very pathological cases, this extreme degeneracy is absent in the case $\alpha > 0$. In the generic, noisy case all nonzero eigenvalues will be distinct and if m is large then there are more than d positive eigenvalues of $\hat{T}$, so that the solution will be unique.

Estimation. Having found P to maximize the empirical objective $\hat{L}(.)$, can we be confident that the true objective $L(P)$ is also nearly maximal, and how does this confidence improve with the sample size?

These questions are complicated by the interdependence of observations, in particular by the possibility of being trapped in an unrepresentative corner of the state space for longer periods of time. Since we want to estimate an expectation on the basis of a temporal average, some sort of ergodicity property of the process S will be relevant. Our bounds are expressed in terms of the mixing coefficients $\beta(\tau)$, which roughly bound the interdependence of past and future variables separated by a time interval of duration τ (see section 2.1). Combining the techniques developed in [15] and [23] we arrive at the following result:

Theorem 2. *With the assumptions already introduced above, fix $\delta > 0$ and let $m, \tau \in \mathbb{N}$, $\tau < m/2$ and $l = \lfloor m/(2\tau) \rfloor$ and $\beta(\tau) < \delta/(2l)$. Then with probability greater $1-\delta$ in the observation of $(X_0, ..., X_m)$ we have*

$$\sup_{P \in \mathcal{P}_d} \left| \hat{L}(P) - L(P) \right| \leq \frac{4}{\sqrt{l}} \left(\sqrt{d} + \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta(\tau - 1)}} \right).$$

If the mixing coefficients β are known, then the right hand side can be minimized with an appropriate choice of τ, which in general depends on the sample size (or *total learning time*) m. For easy interpretation assume $\beta(\tau) = 0$ for $\tau \geq \tau_0$. Then we can interpret τ_0 as the mixing time beyond which all correlations vanish. If we set $\tau = \tau_0 + 1$ above, the resulting bound resembles the bound for the iid case (see Lemma 5), with an effective sample size $l = \lfloor m/(2(\tau_0 + 1)) \rfloor$. We can distinguish two time-scales:

- The smeared present in units of order 1. In this paper we use the variance of the velocity process, but any correlation on a time-scale $\ll \tau_0$ can in principle be exploited by the learner.
- The learning time in units of order τ_0. Dependencies over time scales $> \tau_0$ disappear, the process behaves like a sequence of iid variables and the learner can estimate expected properties of the smeared present.

Often the mixing coefficients are unknown, but one knows (or assumes or hopes) that S is absolutely regular, that is $\beta(\tau) \to 0$ as $\tau \to \infty$. We can then still establish convergence in probability:

Theorem 3. *If X is absolutely regular then for every $\epsilon > 0$ we have*

$$\lim_{m \to \infty} \Pr \left\{ \sup_{P \in \mathcal{P}_d} \left| \hat{L}(P) - L(P) \right| > \epsilon \right\} = 0.$$

We will prove both theorems in section 3. With $\alpha = 1$ these results specialize to generalization guarantees for PSA of weakly dependent processes.

Motivation. Now that we know how to maximize the empirical objective $\hat{L}_\alpha$, and that maximizing $\hat{L}_\alpha$ approximately maximizes the true objective L_α, can we give a more precise description of the benefit incurred by maximizing L_α?

Consider an unknown partitioning of the state space into disjoint categories and the simple rule which assigns the same category to two states s and s' if and only if $\|P\phi(s) - P\phi(s')\| < \sqrt{\alpha}$. We are interested in a bound on the error probability *Err* of this rule as s and s' are drawn independently from the stationary distribution of the process S. Clearly such a bound depends on the relationship of the process to the categories in question. In section 4 we define a corresponding property, called autoergodicity, and prove that for any partitioning into autoergodic categories

$$\mathrm{Err} \leq \frac{1}{1-\alpha}\left(1 - \frac{2}{\alpha} L_\alpha(P)\right) - R,$$

where R is the probability that two independently chosen states belong to the same category. Maximizing L_α therefore minimizes an error bound for any future partitioning as long as the future categories satisfy the autoergodicity requirement. The bound above also implies a rule for the choice of the parameter α. In section 5 we give some examples of β-mixing processes and autoergodic partitions.

Experiments. A practical problem caused by large observation times is the accumulating memory requirement to store the sample data, as long as we adhere to the batch algorithm sketched above. For this reason we use an online-algorithm for our experiments with image recognition. The algorithm, a modification of an algorithm introduced by Oja [12], is briefly introduced in section 6.1. We apply it either directly to the image data or to train the second layer of a two-layered radial-basis-function network.

Some of the experiments reported in section 6 involve processes with specific geometric invariants: Consider rapidly rotating views of a slowly changing scene. The projection returned by our algorithm then performs well as a preprocessor for rotation invariant recognition. An analogous behavior was observed for scale-invariance, and it might be conjectured that similar mechanisms could account for the ubiquity of scale invariant perception in biological vision.

Other experiments were made with face recognition. Using the ATT face dataset a process was generated which typically presents many successive images of the same person before a random change to another person is made. The corresponding projection then performs very well as a preprocessor for the images of the other subjects not involved in the process, but also in the same dataset.

A similar technique has been proposed by Wiskott and Sejnowski [21]. It is missing an analogue of a positive variance term in the objective function. The problem of potentially trivial solutions is circumnavigated by an orthonormalization prescription (whitening) of the covariance matrix prior to the subspace search, which then essentially seeks out a minimal subspace of the velocity covariance. In high (or infinite) dimensions minimal subspace analysis of (compact positive) operators should cause the above-mentioned degeneracy problem, because the eigenvalues will concentrate at zero. In [21] a corresponding problem is in fact mentioned. Also the orthonormalization increases the norms of the input vectors as the dimension grows, making it difficult to analyse the generalization behavior. In our approach all these problems are eliminated by a positive variance term, corresponding to $\alpha > 0$.

A shorter precursor of this article is [8]. This version of the paper is more self-contained and gives a broader discussion of autoergodicity (termed *continuity* in [8]) and of β-mixing processes.

2 Preliminaries

In this section we introduce some general assumptions, definitions and techniques to be used in the following. A small appendix to this paper gives a tabular summary of the more frequently used notation.

2.1 Stationary processes, mixing coefficients and inequalities

Throughout this paper $S = (S_t)_{t\in\mathbb{Z}}$ will denote a sequence of random variables with values in some measurable space of states $(\mathbf{\Omega}, \Sigma)$. For $I \subseteq \mathbb{Z}$ let $\Sigma^I = \otimes_{i\in I}\Sigma$ and use μ_I to denote the joint distribution of $(S_t)_{t\in I}$ on $\left(\mathbf{\Omega}^I, \Sigma^I\right)$.

We assume that S is *strictly stationary*, that is $\mu_I = \mu_{I+t}$, for all $I \subseteq \mathbb{Z}$ and $t \in \mathbb{Z}$. In particular all the S_t will have the same distribution $\mu_{\{t\}} = \mu_{\{0\}}$ on $(\mathbf{\Omega}, \Sigma)$. We will call $\mu_{\{0\}}$ the *stationary distribution.*

The assumption of stationarity replaces the assumption of identical distribution, one of the "i's" in the iid-assumption usually made in learning theory. It allows us to infer information on the future behavior of the process from past observations.

The assumption of independence of the S_t (the other "i") is often unrealistic, and, as we have claimed in the introduction, dependencies on a small time scale can actually be exploited to the learners benefit. On a larger time scale however we need some approximate independence to ensure that a finite number of consecutive observations covers a representative portion of the state space. The β-mixing coefficients ([17], [3], [23]) provide a way to control the error made by the assumption of independence over larger time scales.

Definition 1. *For $\tau \in \mathbb{N}$ define the β-mixing coefficient*

$$\beta_S(\tau) = \sup_{B\in\Sigma^{\{t\leq 0\}}\otimes\Sigma^{\{t\geq\tau\}}} \left|\mu_{\{t\leq 0\}\cup\{t\geq\tau\}}(B) - \mu_{\{t\leq 0\}} \times \mu_{\{t\geq\tau\}}(B)\right|$$

The process S is called absolutely regular or β-mixing if $\beta_S(\tau) \to 0$ as $\tau \to \infty$. It is called exponentially β-mixing if there are constants C and $c > 0$ such that $\beta_S(\tau) \leq Ce^{-c\tau}$, $\forall\tau \in \mathbb{N}$.

The interpretation is as follows: Let B be any statement depending on the past $\{t \leq 0\}$ and the remote future $\{t \geq \tau\}$, that is $B \in \Sigma^{\{t \leq 0\}} \otimes \Sigma^{\{t \geq \tau\}}$. So $\mu_{\{t \leq 0\} \cup \{t \geq \tau\}}(B)$ is the true probability of the event B and $\mu_{\{t \leq 0\}} \times \mu_{\{t \geq \tau\}}(B)$ would be the probability if past and future were independent. Then $\beta_S(\tau)$ is the maximal error incurred by this approximation for any such event B. The smaller $\beta_S(\tau)$, the more nearly independent are past and τ-future.

If in the above definition the supremum was constrained to events of the form $B \in \Sigma^{\{t \leq 0\}} \times \Sigma^{\{t \geq \tau\}}$, instead of $B \in \Sigma^{\{t \leq 0\}} \otimes \Sigma^{\{t \geq \tau\}}$, we would obtain the definition of α-mixing coefficients and α-mixing processes. The weaker α-mixing coefficient measures directly the maximal dependence of pairs of past and future events and is somewhat more intuitive than the β-mixing coefficient which measures the maximal error introduced by the assumption of independence for arbitrary events. We believe that an analogue of Theorem 3 is also true for α-mixing. Nevertheless it seems much easier to deal with the β-mixing coefficients, and absolute regularity can often be proven to hold for realistic α-mixing processes.

Several processes are absolutely regular: All countable-state, irreducible and aperiodic Markov chains and the standard ARMA processes are β-mixing. A strictly stationary ergodic and aperiodic (possibly continuous-state) Markov chain S_t is exponentially β-mixing if it satisfies the so called Doeblin condition (see [23], [10], [3])

$$\begin{aligned}
&\exists A \subseteq \mathbf{\Omega}, \text{with } \Pr\{S_0 \in A\} = 1,\ \exists \epsilon \in (0,1),\ \exists n \geq 1 \text{ such that} \\
&\forall x \in A,\ \forall B \subseteq \mathbf{\Omega} \text{ with } \Pr\{S_0 \in B\} \leq \epsilon, \text{ one has that} \\
&\Pr\{S_n \in B | S_0 = x\} \leq 1 - \epsilon\ . \qquad (3)
\end{aligned}$$

For any strictly stationary Markov chain it follows from the Markov property that the β-mixing coefficients are given by

$$\beta_S(\tau) = \sup_{B \in \Sigma \otimes \Sigma} \left| \mu_{\{0,\tau\}}(B) - \mu_{\{0\}} \times \mu_{\{0\}}(B) \right| .$$

We will use this simpler condition to verify β-mixing for the examples in section 5.

Let ϕ be a measurable map from $(\mathbf{\Omega}, \Sigma)$ to some other measurable space $(\mathbf{\Omega}', \Sigma')$. From Definition 1 it is easy to see that $\beta_{\phi \circ S}(\tau) \leq \beta_S(\tau)$, $\forall \tau \in \mathbb{N}$, and that $\phi \circ S$ is absolutely regular whenever S is.

The mixing coefficients can also be used to control the approximation of the law μ by a product measure involving more than two factors (see also Bin Yu [23]):

Lemma 1. *Let $B \in \Sigma_{\{1,...,m\}}$. Then*

$$\left| \mu_{\{1,...,m\}}(B) - \left(\mu_{\{0\}}\right)^m (B) \right| \leq (m-1)\, \beta_S(1) .$$

Proof. By stationarity, Fubini's Theorem and Definition 1, we have for $1 \leq k < m$, that

$$\left| \mu_{\{1\}} \times ... \times \mu_{\{k,...,m\}}(B) - \mu_{\{1\}} \times ... \times \mu_{\{k\}} \times \mu_{\{k+1,...,m\}}(B) \right| \leq \beta_S(1) .$$

Then, again with stationarity and a telescopic expansion,

$$\begin{aligned}
&\left|\mu_{\{1,...,m\}}(B) - \left(\mu_{\{0\}}\right)^m(B)\right| \\
&= \left|\mu_{\{1,...,m\}}(B) - \mu_{\{1\}} \times ... \times \mu_{\{m\}}(B)\right| \\
&\leq \sum_{k=1}^{m-1} \left|\mu_{\{1\}} \times ... \times \mu_{\{k,...,m\}}(B) - \mu_{\{1\}} \times ... \times \mu_{\{k\}} \times \mu_{\{k+1,...,m\}}(B)\right| \\
&\leq (m-1)\,\beta_S(1) \quad \square
\end{aligned}$$

We will also need the following lemma of Vidyasagar [20, Lemma 3.1]:

Lemma 2. *Suppose $\beta(\tau) \downarrow 0$ as $\tau \to \infty$. It is possible to choose a sequence $\{\tau_m\}$ such that $\tau_m \leq m$, and with $l_m = \lfloor m/\tau_m \rfloor$ we have that $l_m \to \infty$ while $l_m \beta(\tau_m) \to 0$ as $m \to \infty$.*

2.2 Hilbert Schmidt operators

For the next sections H will be a real separable Hilbert space with norm $\|.\|$ and inner product $\langle ., . \rangle$. In practice H will be very high dimensional. Our bounds are dimension free and also hold in the limiting case of infinite-dimensionality.

We assume that the learner observes the state space by means of a fixed sensory system or feature map $\phi : \mathbf{\Omega} \to H$. The function ϕ can hide important properties of the states, such as the backside view of spatial objects. In addition to the sensory measurements, ϕ can include forms or fixed neural processing of the sensory outputs or kernel-induced feature maps.

We require ϕ to be Σ-measurable and normalized and centered w.r.t. S in the sense that $\|\phi(s)\| \leq 1/2$, $\forall s \in \mathbf{\Omega}$ and $\mathbb{E}[\phi(S_t)] = 0$. With $X = (X_t)_{t \in \mathbb{Z}}$ we denote the H-valued process $X_t = \phi(S_t)$, and the velocity process $\dot{X} = \left(\dot{X}_t\right)_{t \in \mathbb{Z}}$ is given by $\dot{X}_t = X_t - X_{t-1}$.

With H_2 we denote the real vector space of symmetric operators on H satisfying $\sum_{i=1}^{\infty} \|Te_i\|^2 < \infty$ for every orthonormal basis $(e_i)_{i=1}^{\infty}$ of H. For $T_1, T_2 \in H_2$ the number $\langle T_1, T_2 \rangle_2 = \sum_{i=1}^{\infty} \langle T_1 e_i, T_2 e_i \rangle$ is independent of the chosen basis and defines an inner product on H_2, making it into a Hilbert space with norm $\|T\|_2 = \langle T, T \rangle_2^{1/2}$. The members of H_2 are compact and called *Hilbert-Schmidt operators* (see Reed and Simon [16] for background on functional analysis). For every $v \in H$ we define an operator Q_v by

$$Q_v x = \langle x, v \rangle v \text{ for all } x \in H.$$

The set of d-dimensional, orthogonal projections in H is denoted with $\mathcal{P}_d$. The following facts are easily verified (see also [7]):

Lemma 3. *Let $x, y \in H$ and $P \in \mathcal{P}_d$. Then (i) $Q_x \in H_2$ and $\|Q_x\|_2 = \|x\|^2$, (ii) $\langle Q_x, Q_y \rangle_2 = \langle x, y \rangle^2$, (iii) $\langle P, Q_x \rangle_2 = \|Px\|^2$ and (iv) $\|P\|_2 = \sqrt{d}$.*

Proof. If $x = 0$ then (i)-(iii) are trivial. If $x \neq 0$ let $(e_i)_{i=1}^{\infty}$ be an orthonormal basis with $e_1 = x / \|x\|$. Then

$$\langle Q_x, Q_y \rangle_2 = \sum_i \langle Q_x e_i, Q_y e_i \rangle = \langle Q_x e_1, Q_y e_1 \rangle = \langle x, y \rangle^2 ,$$

which proves (ii), from which (i) follows immediately. In the same basis

$$\langle P, Q_x \rangle_2 = \sum_i \langle P e_i, Q_x e_i \rangle = \langle P e_1, Q_x e_1 \rangle = \langle Px, x \rangle = \langle Px, Px \rangle = \|Px\|^2 ,$$

which is (iii). Letting $(e_i)_{i=1}^{\infty}$ be an orthonormal basis such that $(e_i)_{i=1}^{d}$ are a basis for the range of P we get

$$\|P\|_2^2 = \sum_i \langle P e_i, P e_i \rangle = \sum_{i=1}^{d} \langle e_i, e_i \rangle = d,$$

which gives (iv) □

If $T \in H_2$ is symmetric, then it follows from the spectral theorem [16] for compact operators that

$$T = \sum_{i=1}^{\infty} \lambda_i Q_{e_i}, \tag{4}$$

where (λ_i) is the sequence of real eigenvalues and (e_i) the (complete, orthonormal) sequence of eigenvectors of T. The series is convergent in the H_2-norm and $\sum \lambda_i^2 = \|T\|_2^2$.

Lemma 4. *Suppose $T \in H_2$ is symmetric, $d \in \mathbb{N}$ and the sum in (4) can be arranged that $\lambda_i \geq \lambda_j$ for all $i \leq d < j$. Then*

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

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

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

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

which proves $\sup_{P \in \mathcal{P}_d} \langle T, P\rangle_2 \leq \sum_{i=1}^{d} \lambda_i$ (this also follows directly from Horn's theorem [18, 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 □

In terms of the Q-operators we can rewrite the operators T and $\hat{T}$ in (1) and (2) as

$$T = \mathbb{E}\left[\alpha Q_X - (1-\alpha) Q_{\dot{X}}\right] \text{ and } \hat{T} = \frac{1}{m} \sum_{i=1}^{m} \left(\alpha Q_{X_i} - (1-\alpha) Q_{\dot{X}_i}\right).$$

Using Lemma 3, (iii) above, the objective functionals $L(.)$ and $\hat{L}(.)$ become

$$L(P) = \langle T, P\rangle_2 \text{ and } \hat{L}(P) = \left\langle \hat{T}, P \right\rangle_2.$$

We also have $\|T\|_2 \leq \mathbb{E}\left[\alpha \|Q_X\|_2 + (1-\alpha)\|Q_{\dot{X}}\|_2\right] \leq 1$ and similarly $\left\|\hat{T}\right\|_2 \leq 1$. The operators T and $\hat{T}$ are both in H_2, $\hat{T}$ has finite rank.

Proof (of Theorem 1). The conclusion follows from Lemma 4 and the identity $\hat{L}(P) = \left\langle \hat{T}, P \right\rangle_2$ □

These arguments are fairly standard, but there are some potential pitfalls resulting from non-positivity. For example the above is not generally true for the operator T corresponding to the true objective functional L in the infinite dimensional case, because it may happen that T has fewer than d nonnegative eigenvalues, or none at all. Since all negative eigenvalues converge to 0, the supremum might not be attained.

3 Generalization

In this section the previously introduced techniques are used to derive a uniform bound on the estimation difference between the empirical and true objective functionals $\hat{L}$ and L. We first prove a general result for vector-valued processes, which is then applied to operator-valued processes to give a relatively easy proof of Theorems 2 and 3.

For two subsets $V, W \subseteq H$ of a Hilbert space H we introduce the following notation

$$\|V\| = \sup_{v \in V} \|v\| \text{ and } |\langle V, W \rangle| = \sup_{v \in V, w \in W} |\langle v, w \rangle| .$$

Theorem 4. *Let $V, W \subset H$ and $X = \{X_t\}_{t \in \mathbb{Z}}$ a stationary, mean zero process with values in V.*

1. Fix $\delta > 0$ and let $m, \tau \in \mathbb{N}$, $\tau < m/2$ and $l = \lfloor m/(2\tau) \rfloor$ and $\beta(\tau) < \delta/(2l)$. Then with probability greater than $1 - \delta$ we have

$$\sup_{w \in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| \leq \frac{2}{\sqrt{l}} \left(\|V\| \|W\| + |\langle V, W \rangle| \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta_X(\tau)}} \right).$$

2. If X is absolutely regular then for every $\epsilon > 0$

$$\Pr \left\{ \sup_{w \in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| > \epsilon \right\} \to 0 \text{ as } m \to \infty.$$

If we let W be the unit ball in H we immediately obtain the following

Corollary 1. *Under the first assumptions of Theorem 4 we have with probability greater $1 - \delta$ that*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} X_i \right\| \leq \frac{2 \|V\|}{\sqrt{l}} \left(1 + \sqrt{\frac{1}{2} \ln \frac{1}{\delta/2 - l\beta_X(\tau)}} \right).$$

If in addition X_t is absolutely regular then $\|(1/m) \sum_{i=1}^{m} X_i\| \to 0$ in probability.

Here is a practical reformulation with trivial proof:

Corollary 2. *Theorem 4 and Corollary 1 remain valid if the mean-zero assumption is omitted, X_i is replaced by $X_i - \mathbb{E}[X_1]$ and $\|V\|$ and $|\langle V, W \rangle|$ are replaced by $2\|V\|$ and $2|\langle V, W \rangle|$ respectively.*

To prove Theorem 4 we first establish an analogous result for iid X_i (essentially following [15]) and then adapt it to dependent variables.

Lemma 5. *Let $V, W \subset H$ be and $X_1, ..., X_m$ iid zero-mean random variables with values in V. Then for ϵ and m such that $\|W\| \|V\| < \sqrt{m}\epsilon$ we have*

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

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

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

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

$$\mathbb{E}\left[f(\mathbf{X})\right] = \mathbb{E}\left[\sup_{w \in W}\left|\left\langle w, \bar{\mathbf{X}}\right\rangle\right|\right] \leq \|W\| \,\mathbb{E}\left[\left\|\bar{\mathbf{X}}\right\|\right] \leq \left(1/\sqrt{m}\right)\|W\| \|V\|. \tag{5}$$

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

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

By (5) and the bounded-difference inequality (see [9]) we obtain for $t > 0$

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

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

The proof of Theorem 4 now uses the techniques introduced by Yu [23] (see also Meir [10] and Lozano et al [5]).

Proof (of Theorem 4). Select a time-scale $\tau \in \mathbb{N}$, $2\tau < m$ and represent the discrete time axis as an alternating sequence of blocks

$$\mathbb{Z} = \left(..., H_{-1}, T_{-1}, H_0, T_0, H_1, T_1, ..., H_k, T_k, ...\right),$$

where each of the H_k and T_k has length τ,

$$H_k = \{2k\tau, ..., 2k\tau + \tau - 1\} \text{ and } T_k = \{(2k+1)\tau, ..., (2k+1)\tau + \tau - 1\}.$$

We now define the blocked processes X^H and X^T with values in the convex hull $\mathrm{co}(V)$ by $X_t^H = (1/\tau)\sum_{j \in H_t} X_j$ and $X_t^T = (1/\tau)\sum_{j \in T_t} X_j$. By stationarity the

X_i^H and X_i^T are identically distributed and themselves stationary. Because of the gaps of size τ we have $\beta_{X^H}(1) = \beta_{X^T}(1) = \beta_X(\tau)$. We can now write

$$(1, ..., m) = (H_1, T_1, H_2, T_2, ..., H_l, T_l, R),$$

where the number l of block-pairs is chosen so as to minimize the size of the remainder R, so $l = \lfloor m/(2\tau) \rfloor$ and $|R| < 2\tau$. For arbitrary $\epsilon > 0$ we obtain

$$\begin{aligned}
&\Pr\left\{ \sup_{w\in W} \left| \frac{1}{2\tau l} \sum_{i=1}^{2\tau l} \langle w, X_i \rangle \right| > \epsilon \right\} \\
&= \Pr\left\{ \sup_{w\in W} \left| \frac{1}{2l} \sum_{i=1}^{l} \left\langle w, X_i^H \right\rangle + \frac{1}{2l} \sum_{i=1}^{l} \left\langle w, X_i^T \right\rangle \right| > \epsilon \right\} \\
&\leq \Pr\left\{ \sup_{w\in W} \left| \frac{1}{2l} \sum_{i=1}^{l} \left\langle w, X_i^H \right\rangle \right| + \sup_{w\in W} \left| \frac{1}{2l} \sum_{i=1}^{l} \left\langle w, X_i^T \right\rangle \right| > \epsilon \right\} \\
&= 2\Pr\left\{ \sup_{w\in W} \left| \frac{1}{l} \sum_{i=1}^{l} \left\langle w, X_i^H \right\rangle \right| > \epsilon \right\} \\
&\leq 2\exp\left(\frac{-\left(\sqrt{l}\epsilon - \|V\|\,\|W\|\right)^2}{2\,|\langle V, W \rangle|^2} \right) + 2l\beta_X(\tau).
\end{aligned}$$

The last inequality follows from the mixing Lemma 1, $\beta_{X^H}(1) = \beta_X(\tau)$, the iid case Lemma 5 and the fact that $\|\mathrm{co}(V)\| = \|V\|$ and $|\langle \mathrm{co}(V), W \rangle| = |\langle V, W \rangle|$. To deal with the remainder R, note that

$$\Pr\left\{ \sup_{w\in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| > \epsilon \right\} \leq \Pr\left\{ \sup_{w\in W} \left| \frac{1}{2\tau l} \sum_{i=1}^{2\tau l} \langle w, X_i \rangle \right| + \frac{\|V\|\,\|W\|}{l} > \epsilon \right\}.$$

We thus obtain

$$\begin{aligned}
&\Pr\left\{ \sup_{w\in W} \left| \frac{1}{m} \sum_{i=1}^{m} \langle w, X_i \rangle \right| > \epsilon \right\} \\
&\leq 2\exp\left(\frac{-\left(\sqrt{l}\epsilon - \left(1 + \frac{1}{\sqrt{l}}\right)\|V\|\,\|W\|\right)^2}{2\,|\langle V, W \rangle|^2} \right) + 2l\beta_X(\tau). \qquad (6)
\end{aligned}$$

Solving for ϵ and using $\left(1 + 1/\sqrt{l}\right) \leq 2$ gives the first conclusion.

If X is absolutely regular then $\beta(\tau) \downarrow 0$ as $\tau \to \infty$. Choosing a subsequence τ_m as in Lemma 2 we have $l_m = \lfloor m/(2\tau_m) \rfloor \to \infty$ and $l_m \beta(\tau_m) \to 0$. Substituting l_m for l and τ_m for τ above, the bound (6) will go to zero as $m \to \infty$, which proves the second conclusion □

Now it is easy to prove the bounds in the introduction by applying Theorem 4 to the stationary operator-valued stochastic process

$$A_t = \alpha Q_{X_t} - (1-\alpha)\, Q_{\dot{X}_t}, \tag{7}$$

which we reinterpret as a vector-valued process with values in the Hilbert space H_2 of Hilbert-Schmidt operators. Note that $T = \mathbb{E}[A_1]$ and $\hat{T} = (1/m)\sum_1^m A_i$.

Proof (of Theorem 2 and Theorem 3). : First note that $\beta_A(\tau) = \beta_X(\tau - 1)$, because A_t depends also on X_{t-1} through the velocity process, and that A is absolutely regular if X is. Set $W = \mathcal{P}_d$ and define $V \subset H_2$ by

$$V = \{\alpha Q_x - (1-\alpha)\, Q_y : \|x\| \leq 1 \text{ and } \|y\| \leq 1\}.$$

Then $A_t \in V$ a.s. By Lemma 3 (i), V is contained in the unit ball in H_2 and

$$\begin{aligned} |\langle V, W\rangle_2| &= \sup_{P\in\mathcal{P}_d} \sup \left\{ \left|\langle P, \alpha Q_x - (1-\alpha)\, Q_y\rangle_2\right| : \|x\| \leq 1,\ \|y\| \leq 1\right\} \\ &\leq \sup_{P\in\mathcal{P}_d} \sup \left\{\alpha \|Px\|^2 + (1-\alpha)\|Py\|^2\right\} \leq 1. \end{aligned}$$

By Lemma 3 (iv) $\|W\|_2 = \sqrt{d}$. We also have

$$\sup_{P\in\mathcal{P}_d} \left|\hat{L}(P) - L(P)\right| = \sup_{P\in\mathcal{P}_d} \left|\frac{1}{m}\sum_{i=1}^m \langle P, A_i - \mathbb{E}[A_1]\rangle_2\right|.$$

Applying Corollary 2 to the process $A_t - \mathbb{E}[A_1]$ gives both Theorem 2 and 3 □.

4 A Generic Error Bound

The previous sections have shown how we can find a projection P to approximately maximize the true objective functional

$$L_\alpha(P) = \alpha\mathbb{E}\left[\|P\phi(S_0)\|^2\right] - (1-\alpha)\,\mathbb{E}\left[\|P\phi(S_1) - P\phi(S_0)\|^2\right],$$

with $\alpha \in [0,1]$. In this section we investigate the utility of the map $P \circ \phi$ as a preprocessor for future classification problems, where the state-space $\mathbf{\Omega}$ is partitioned into a finite or countable number of disjoint categories

$$\mathbf{\Omega} = \bigcup_k E_k.$$

Since P has been trained in an unsupervised way from the process S, this requires that the categories E_k themselves are somehow compatible to the process S. To motivate the definition and result given below we give a second heuristic derivation of the functional L_α starting from a given (but unknown) partition $\{E_k\}$, for which we require some general (and rather vague) common sense properties.

With a large number of categories it is very unlikely that two independently drawn states belong to the same category, so they should be mapped at a large distance from each other. The projection P should therefore be chosen to maximize

$$(1/2)\,\mathbb{E}_{SS'}\left[\left\|P\phi\left(S_0\right)-P\phi\left(S_0'\right)\right\|^2\right]=\mathbb{E}\left[\left\|P\phi\left(S_0\right)\right\|^2\right], \tag{8}$$

where we have used the mean-zero property of $\phi\left(S_0\right)$.

By some assumed continuity property of common-sense categories, two consecutively observed states are very likely to belong to the same category (a variant of the slowness postulate, just think of a weather forecast for the next minute), and should therefore be mapped close to each other. Thus P should minimize

$$\mathbb{E}_S\left[\left\|P\phi\left(S_1\right)-P\phi\left(S_0\right)\right\|^2\right]. \tag{9}$$

Combining the two objectives with the trade-off parameter α gives the functional L_α and our algorithm.

To convert this type of reasoning into a solid error-bound, we first have to decide on a distribution on $\mathbf{\Omega}$ which is underlying the classification problem and relative to which the expectations of errors have to be measured. The obvious and only natural candidate is the invariant distribution $\mu_{\{0\}}$, which models the frequency of states throughout the process.

Then (8) just gives the total variance, to which (in the language of linear discriminant analysis) the *inter-class variance* should make the dominant contribution. Unfortunately the other expression (9) which arises from the slowness postulate cannot be directly used as a bound on the *intra-class variance*, because the expectation in (9) is relative to the marginal measure $\mu_{\{0,1\}}$ (c.f. section 2.1 for notation), while the intra-class variance would be the same expression using the measure $\mu_{\{0\}}\times\mu_{\{0\}}$, restricted to $\cup_k E_k\times E_k$. We therefore need to bound $\mu_{\{0\}}\times\mu_{\{0\}}$ in terms of $\mu_{\{0,1\}}$ on the σ-algebra of sub-events of $\cup_k E_k\times E_k$. This motivates the following definition.

Definition 2. *A measurable set $E\subset\mathbf{\Omega}$ is called autoergodic w.r.t. the process S if for all measurable $A,B\subseteq E$ with $A\cap B=\emptyset$ we have*

$$\mu_{\{0\}}\left(A\right)\mu_{\{0\}}\left(B\right)\leq\frac{1}{2}\left(\mu_{\{0,1\}}\left(A\times B\right)+\mu_{\{0,1\}}\left(B\times A\right)\right). \tag{10}$$

A finite or countable disjoint partition of the state space into measurable sets E_k

$$\mathbf{\Omega}=\bigcup_k E_k \text{ with } E_k\cap E_l=\emptyset$$

is called auto-ergodic w.r.t. S if all the E_k are autoergodic w.r.t. S.

So if E is autoergodic w.r.t. S, then any two mutually exclusive events A and B implying E are more likely to be observed consecutively (averaged over the two possible orders of succession), than in two independent observations.

Every subset of an autoergodic set is autoergodic, and so is every singleton set $\{s\}$. In the discrete case a set E is autoergodic iff

$$\mu_{\{0\}}(s)\,\mu_{\{0\}}(s') \leq (1/2)\left(\mu_{\{0,1\}}(s,s') + \mu_{\{0,1\}}(s',s)\right)$$

for all $s, s' \in E$ with $s \neq s'$. If $\mu_{\{0,1\}}$ is absolutely continuous w.r.t. $\mu_{\{0\}}^2$, then $\mu_{\{0,1\}}$ is given by a density function ρ and the autoergodic sets are those measurable sets $E \subset \mathbf{\Omega}$ which satisfy $(1/2)(\rho(s,s') + \rho(s',s)) \geq 1$ for almost all $s, s' \in E$ with $s \neq s'$.

A sufficient condition for a set E to be autoergodic is that

$$\mu_{\{0\}}(A)\,\mu_{\{0\}}(B) \leq \mu_{\{0,1\}}(A \times B) \tag{11}$$

for all measurable $A, B \subset E$. This is simpler than (10) but a corresponding definition would exclude many of the more realistic cases.

If the process is α-mixing another sufficient condition for autoergodicity is that the sequence

$$\mu_{\{0,\tau\}}(A \times B) + \mu_{\{0,\tau\}}(B \times A)\,,\ \tau \in \mathbb{N}$$

be nonincreasing for all disjoint $A, B \subset E$, because then

$$\begin{aligned}\mu_{\{0,1\}}(A \times B) + \mu_{\{0,1\}}(B \times A) &\geq \lim_{\tau\to\infty}\left(\mu_{\{0,\tau\}}(A \times B) + \mu_{\{0,\tau\}}(B \times A)\right)\\ &= 2\mu_{\{0\}}(A)\,\mu_{\{0\}}(B)\,.\end{aligned}$$

Is autoergodicity a general requirement for common-sense categories E in realistic processes? Not quite. Consider the following statements about image frames in movies (at normal rates of frame-repetition): $E =$"a child is crossing the street" and $A, B \subset E$ given by $A =$"a little girl is crossing the street" and $B =$"a little boy is crossing the street". For E to be autoergodic A and B must be observed at least as likely one frame apart as in two independently chosen frames, which seems impossible, unless we allow a nonvanishing probability for the girl turning into a boy or vice versa in the middle of the street. For a similar reason A fails to be autoergodic, just consider A', $A'' \subset A$ where the little girl wears a red or a blue shirt respectively.

This type of problem appears when E can be split into sub-events without excess mutual transition probability. Autoergodicity requires this to be impossible (hence the name).

As a more positive example consider a sequence of facial portraits of the speakers in a talk show with many participants, sampled at a rate slow enough to allow arbitrary changes of facial expression or camera perspective, but fast enough to have every participants turn represented by a large number of consecutive portraits. Consider the event $E =$"it is Fred" and $A, B \subset E$ such as $A =$"Fred is smiling" and $B =$"Fred is frowning" and imagine the situation where we are about to unveil an image. Without any other information it will be improbable to observe B, because of the many other participants. If, on the other hand, we are shown the previous image, where we observe A, then we already know that we are in the middle of Fred's turn, making it also more likely to expect B. This means $\Pr(B) \leq \Pr(S_1 \in B | S_0 \in A)$ which is the same as (11) and shows that E would be autoergodic if all subsets of E behaved like A and B.

While autoergodicity is not a property of common-sense categories in realistic processes, these examples indicate that it approximates some aspects of such categories. In the next section we give some examples of β-mixing processes and autoergodic categories which will be relevant for our experiments. Before that we derive the most important consequences for our algorithm.

To obtain Lemma 6 below we make the following technical assumption: $\mathbf{\Omega}$ is a σ-compact Hausdorff space (a countable union of compact Hausdorff spaces, examples are $\mathbb{Z}$, or $\mathbb{R}^N$, see [2]), $\varSigma$ contains the Borel-field on $\mathbf{\Omega}$ and the feature map $\phi : \mathbf{\Omega} \to H$ is continuous. The reader who is troubled with this assumption can simply assume the state-space to be countable.

Lemma 6. *Suppose* $g : \mathbf{\Omega} \times \mathbf{\Omega} \to \mathbb{R}$ *is symmetric, nonnegative, continuous and vanishes on the diagonal. If* $E \subset \mathbf{\Omega}$ *is autoergodic then*

$$\mathbb{E}_{\mu^2_{\{0\}}}\left[g\ 1_{E\times E}\right] \leq \mathbb{E}_{\mu_{\{0,1\}}}\left[g\ 1_{E\times E}\right].$$

Proof. A technical construction relying on the σ-compactness of $\mathbf{\Omega}$ and the continuity of g produces a sequence of simple functions $\psi_n = \sum_{i,j=1}^{m_n} c^n_{ij} 1_{A^n_i \times A^n_j}$ with $A^n_i \in \varSigma_{\{0\}}$, $A^n_i \subseteq \dot{E}$, $A^n_i \cap A^n_j = \emptyset$ for $i \neq j$ and $c^n_{ij} = c^n_{ji} \geq 0$, such that $\psi_n \uparrow g\ 1_{E\times E}$. Since $\psi_n \leq g$ and g vanishes on the diagonal $c^n_{ii} = 0$. For arbitrary $\epsilon > 0$ monotone convergence and the autoergodicity of E imply that, for sufficiently large n,

$$\begin{aligned}
\mathbb{E}_{\mu^2_{\{0\}}}\left[g\ 1_{E\times E}\right] - \epsilon \leq \mathbb{E}_{\mu^2_{\{0\}}}\left[\psi_n\right] &= \sum_{i\neq j}^{m_n} c^n_{ij} \mu_{\{0\}}\left(A^n_i\right) \mu_{\{0\}}\left(A^n_j\right) \\
&\leq \sum_{i\neq j}^{m_n} c^n_{ij} \frac{1}{2} \left(\mu_{\{0,1\}}\left(A^n_i \times A^n_j\right) + \mu_{\{0,1\}}\left(A^n_j \times A^n_i\right)\right) \\
&= \frac{1}{2}\left(\sum_{i\neq j}^{m_n} c^n_{ij} \mu_{\{0,1\}}\left(A^n_i \times A^n_j\right) + \sum_{i\neq j}^{m_n} c^n_{ij} \mu_{\{0,1\}}\left(A^n_j \times A^n_i\right)\right) \\
&\leq \mathbb{E}_{\mu_{\{0,1\}}}\left[g\ 1_{E\times E}\right]. \quad \square
\end{aligned}$$

Suppose now that $\{E_k\}$ is a finite or countable partition of $\mathbf{\Omega}$ with each E_k defining some pattern class. Given a pair (s, s') drawn from $\mu^2_{\{0\}}$ we have to decide if s and s' belong to the same class, that is to decide if there is some k such that $s \in E_k$ and $s' \in E_k$. Fix $\alpha > 0$. In the absence of other known structure we use a simple metric decision rule based on the projected input and the distance threshold $\sqrt{\alpha}$.

$$s \text{ and } s' \text{ are in the same class iff } \left\|P\phi(s) - P\phi(s')\right\| < \sqrt{\alpha}.$$

The probability that this rule fails for two states s and s', independently drawn from the stationary distribution, is the quantity

$$\begin{aligned} Err = \mu_{\{0\}}^{2} \big\{ & s \text{ and } s' \text{ are in the same } E_k, \text{ but } \left\| P\phi\left(s\right) - P\phi\left(s'\right) \right\| > \sqrt{\alpha}, \text{ or} \\ & s \text{ and } s' \text{ are in different categories, but } \left\| P\phi\left(s\right) - P\phi\left(s'\right) \right\| < \sqrt{\alpha} \big\} . \end{aligned}$$

Bounds on Err can be converted into error bounds for simple metric classifiers, whenever we are provided with examples for the various E_k. It is interesting that the trade-off parameter α, which had been introduced in an ad hoc manner, now assumes a geometric role.

Theorem 5. *With $\alpha \in (0,1)$, if $\{E_k\}$ is autoergodic w.r.t. S, then for every projection P the error probability for the above rule, as s and s' are drawn independently from $\mu_{\{0\}}$, is bounded by*

$$Err \leq \frac{1}{1-\alpha}\left(1 - \frac{2}{\alpha} L_\alpha\left(P\right)\right) - R$$

where $R = \sum_k \left(\mu_{\{0\}}\left(E_k\right)\right)^2$.

Selecting P to minimize this bound is equivalent to maximizing $L_\alpha\left(P\right)$, the objective of our algorithm. The next section shows that the bound can be arbitrarily tight. The theorem also implies a rule to select the trade-off (or threshold) parameter α: It should be chosen to minimize the first term in the bound above, so α should be close to 0, but a positive value for $L_\alpha\left(P\right)$ should still be obtained, corresponding to positive eigenvalues of the operator T.

It should also be borne in mind that, apart from the autoergodicity assumption, the partition $\{E_k\}$ is largely arbitrary, so that the maximization of L_α learns a feature-map P for an entire family of classification problems, not just a single one. It will be shown in section 5.2 that such families sometimes contain all classification problems with certain invariance properties.

Proof. We introduce the distortion function $\mathit{\Delta}_P : \mathbf{\Omega} \times \mathbf{\Omega} \to \mathbb{R}$ given by

$$\mathit{\Delta}_P\left(s, s'\right) = \left\| P\phi\left(s\right) - P\phi\left(s'\right) \right\|^2 .$$

Then $\mathit{\Delta}_P$ is symmetric, nonnegative, continuous, vanishes on the diagonal and satisfies the hypotheses of Lemma 6. Since the feature map ϕ maps to a sphere of diameter ≤ 1 and projections are norm-decreasing, we also have $\mathit{\Delta}_P\left(s, s'\right) \leq 1$. Then

$$\begin{aligned} Err &= \sum_{k,l:k\neq l} \mathbb{E}_{\mu_{\{0\}}^2}\left[1_{\mathit{\Delta}_P < \alpha} 1_{E_k \times E_l}\right] + \sum_k \mathbb{E}_{\mu_{\{0\}}^2}\left[1_{\mathit{\Delta}_P \geq \alpha} 1_{E_k \times E_k}\right] \\ &= \mathbb{E}_{\mu_{\{0\}}^2}\left[1_{\mathit{\Delta}_P < \alpha}\right] + 2\sum_k \mathbb{E}_{\mu_{\{0\}}^2}\left[1_{\mathit{\Delta}_P \geq \alpha} 1_{E_k \times E_k}\right] - R \\ &\leq \mathbb{E}_{\mu_{\{0\}}^2}\left[\frac{1-\mathit{\Delta}_P}{1-\alpha}\right] + 2\sum_k \mathbb{E}_{\mu_{\{0\}}^2}\left[\frac{\mathit{\Delta}_P}{\alpha}\; 1_{E_k \times E_k}\right] - R \\ &\leq \frac{1}{1-\alpha} - \frac{1}{1-\alpha}\mathbb{E}_{\mu_{\{0\}}^2}\left[\mathit{\Delta}_P\right] + \frac{2}{\alpha}\sum_k \mathbb{E}_{\mu_{\{0,1\}}}\left[\mathit{\Delta}_P\; 1_{E_k \times E_k}\right] - R. \end{aligned}$$

The first inequality uses the bounds $1_{\Delta_P<\alpha} \leq (1-\Delta_P)/(1-\alpha)$ and $1_{\Delta_P\geq\alpha} \leq \Delta_P/\alpha$, which hold since $\Delta_P \in [0,1]$. The other inequality uses the autoergodicity of the E_k-system and Lemma 6. Now we use

$$\sum_k \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\, 1_{E_k\times E_k}\right] \leq \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\right] = \mathbb{E}\left[\left\|P\dot{X}_1\right\|^2\right] = \mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right] \tag{12}$$

and the identity $\mathbb{E}_{\mu^2_{\{0\}}}[\Delta] = 2\mathbb{E}\left[\|PX_0\|^2\right]$, which follows from the mean-zero assumption, to obtain

$$\begin{aligned} Err &\leq \frac{1}{1-\alpha} - \frac{2}{1-\alpha}\mathbb{E}\left[\|PX_0\|^2\right] + \frac{2}{\alpha}\mathbb{E}\left[\left\|P\dot{X}_0\right\|^2\right] - R \\ &= \frac{1}{1-\alpha}\left(1 - \frac{2L_\alpha(P)}{\alpha}\right) - R \quad \square \end{aligned}$$

It follows from the proof that the conclusion holds for every projection P satisfying $\Delta_P(s,s') \leq 1$ a.s., even if ϕ does not map to a sphere of diameter 1.

Also note that the slack in the inequality (12) can be bounded by

$$\sum_{k\neq l} \mathbb{E}_{\mu_{\{0,1\}}}\left[\Delta\, 1_{E_k\times E_l}\right] \leq \Pr\left\{\exists k,l,k\neq l : S_0 \in E_k, S_1 \in E_l\right\},$$

which is the probability to change classes in one time increment. The smaller this probability, the tighter is our bound. On the other hand this implies longer mixing times, so that more observations are necessary to estimate a good projection.

5 Examples of β-mixing processes and autoergodic partitions

This section gives two examples of β-mixing Markov chains, where the state-space can be naturally partitioned into autoergodic categories learnable by our algorithm. The first is modeled after the talk-show example in the previous section, the second is related to the unsupervised learning of invariant categories.

5.1 Driving the process from a multi-class learning task

Let $(\mathcal{X},\mathcal{Y},p)$ be a supervised learning task. $\mathcal{X}$ is the usual input space, assumed countable for simplicity, $\mathcal{Y}$ a finite or countable alphabet of labels, and p a distribution on $\mathcal{X}\times\mathcal{Y}$, such that all labels have a nonvanishing probability, that is $p(k) := p(\mathcal{X}\times\{k\}) > 0, \forall k \in \mathcal{Y}$.

As a state-space we take $\mathbf{\Omega} = \mathcal{X}\times\mathcal{Y}$, so that knowledge of the state implies knowledge of the label. The feature map ϕ however shall depend only on the $\mathcal{X}$-coordinate x of a state $(x,k) \in \mathbf{\Omega}$, so that knowledge of the labels is hidden to the learner.

Fix a parameter $\lambda \in (0,1)$ and define the transition probability from a state $(y,l) \in \mathbf{\Omega}$ at time 0 to a state $(x,k) \in \mathbf{\Omega}$ at time 1 by

$$\Pr\left(S_1 = (x,k)\left|S_0 = (y,l)\right.\right) = (1-\lambda)\, p\left(x|k\right) \delta_{kl} + \lambda p\left(x,k\right).$$

So with probability $1-\lambda$ the label is left unchanged and the new input is drawn from the class-conditional distribution of the old label, and with probability λ a new input-label pair is chosen from p.

It is easy to verify that the above formula defines a transition probability which extends to a Markov chain S with invariant distribution $\mu_{\{0\}} = p$. An easy induction argument leads to the formula

$$\Pr\left(S_\tau = (x,k)\left|S_0 = (y,l)\right.\right) = (1-\lambda)^\tau\, p\left(x|k\right) \delta_{kl} + \left(1-(1-\lambda)^\tau\right) p\left(x,k\right)$$

for any $\tau \in \mathbb{N}$. It follows that for any $B \in \Sigma \otimes \Sigma$ we have

$$\left|\mu_{\{0\}\cup\{\tau\}}\left(B\right) - p \times p\left(B\right)\right| \leq 2\left(1-\lambda\right)^\tau.$$

We conclude that $\beta_S\left(\tau\right) \leq 2\left(1-\lambda\right)^\tau$. The process S is therefore exponentially β-mixing with a characteristic mixing time-scale of order $-\left(\ln\left(1-\lambda\right)\right)^{-1}$ which behaves like λ^{-1} for small λ.

The state-space $\mathbf{\Omega} = \mathcal{X} \times \mathcal{Y}$ has a natural decomposition into disjoint categories

$$\mathbf{\Omega} = \bigcup_{k\in\mathcal{Y}} \mathcal{X} \times \{k\}.$$

We claim that the sets $E_k = \mathcal{X} \times \{k\}$ are autoergodic w.r.t. the process S. Indeed if (x,k) and (y,k) are in E_k, then

$$\begin{aligned}\mu_{\{0,1\}}\left(\left(y,k\right),\left(x,k\right)\right) &= p\left(y,k\right)\left(\left(1-\lambda\right)p\left(x|k\right)\delta_{kk} + \lambda p\left(x,k\right)\right) \\ &= p\left(y,k\right)\left(\frac{\left(1-\lambda\right)}{p\left(k\right)}p\left(x,k\right) + \lambda p\left(x,k\right)\right) \\ &\geq p\left(y,k\right)p\left(x,k\right).\end{aligned}$$

So S is absolutely regular and the E_k are autoergodic and all the results derived above apply.

In the experiments reported below the ATT face-dataset was used to drive such a process. It is interesting that the sequence of facial images, if presented according to the above stochastic rule, makes a remarkably smooth and natural impression, very much like the heuristic talk-show example in the previous section.

For the simplest possible realization of such a process let $\mathcal{X} = \mathcal{Y} = \{-1,1\}$ and $p = (1/2)\left(\delta_{(-1,-1)} + \delta_{(1,1)}\right)$. For the Hilbert space we take ℓ_2 and define $\phi\left(x,k\right) = \left(x/2, 0, 0, ...\right)$. If P_1 is the (evidently optimal) projection onto the first coordinate in ℓ_2 we have that $\mathbb{E}\left[\left\|P_1\phi\left(S_0\right)\right\|^2\right] = 1/4$. Also $\left\|P_1\phi\left(S_1\right) - P_1\phi\left(S_0\right)\right\|^2$

is zero with probability $1-\lambda$ and one with probability λ, so that $L_{\alpha}\left(P_1\right) = \alpha/4 - \lambda\left(1-\alpha\right)$ and the bound in Theorem 5 becomes

$$\mathrm{Err} \leq \frac{\alpha}{2\left(1-\alpha\right)} + \frac{2\lambda}{\alpha},$$

which is of order $\lambda^{1/2}$ if α is chosen of order $\lambda^{1/2}$. This shows that the bound becomes tight as $\lambda \to 0$, and since $\lambda \to 0$ implies that the mixing time $\lambda^{-1} \to \infty$, this illustrates the second of the two remarks after the proof of Theorem 5. It follows from the other remark, that the bound above remains unchanged for any amount of noise distributed orthogonal to the range of P_1 in ℓ_2. Such noise makes the estimation part nevertheless more difficult, and the number of observations necessary to find P_1 will increase.

5.2 Diffusion on a compact group

In this section we adopt the point of view that the evolution of the stimuli received by the learner does not arise from a fixed perspective on a randomly changing environment, but from a randomly changing perspective on a fixed environment. The possible perspectives are parametrized by a compact group, and the changes in perspective are modeled by a generalized diffusion process.

Let G be a compact group with invariant normalized Haar measure m. This means that G is a group and also a topological space, where every point is a closed set and the map $(x,y) \in G \times G \mapsto xy^{-1}$ is continuous. The Haar measure satisfies $m\left(G\right) = 1$ and

$$\int_G f\left(xy\right) dm\left(x\right) = \int_G f\left(yx\right) dm\left(x\right) = \int_G f\left(x^{-1}\right) dm\left(x\right)$$

for every $y \in G$ and every $f \in L_1\left(G, m\right)$. The convolution of two functions $f, g \in L_1\left(G, m\right)$ is defined by

$$f * g\left(x\right) = \int_G f\left(y\right) g\left(y^{-1}x\right) dm\left(y\right),$$

and the n-fold convolution $f^{(n)}$of f with itself is defined recursively by $f^{(1)} = f$ and $f^{(n+1)} = f * f^{(n)}$. Convolution is associative and linear, but not necessarily commutative, unless G is abelian (see [13] for more background).

We will take G as our state-space. Let $\kappa : G \to \mathbb{R}$ be continuous , $\kappa \geq 0$ and $\int_G \kappa dm = 1$. The function κ will be a kernel to generate the process S. Define

$$\Pr\left\{S_t \in A | S_{t-1} = y\right\} = \int_A \kappa\left(xy^{-1}\right) dm\left(x\right).$$

It follows from the invariance properties of m that this is indeed a transition probability, defining a Markov chain S which has the Haar measure as a stationary distribution, $\mu_{\{0\}} = m$. We also have the formula

$$\Pr\left\{S_t \in A | S_0 = y\right\} = \int_A \kappa^{(t)}\left(xy^{-1}\right) dm\left(x\right).$$

A *generating neighborhood* of G is a neighborhood U of the identity in G such that every x in G can be written as $x = u_1 u_2 ... u_m$ for some finite sequence $u_1, ..., u_m \in U$.

Theorem 6. *If $\kappa > 0$ on a generating neighborhood of G then there are constants $C \geq 0$ and $c > 0$ such that*

$$\left\| \kappa^{(n)} - 1 \right\|_\infty \leq C e^{-cn}, \forall n \in \mathbb{N}. \tag{13}$$

Proof. We only need to prove this if κ is not identically 1. First assume that that $\kappa > 0$ throughout G. By continuity and compactness, κ attains its minimum $\lambda \in (0,1)$. We claim that for all $x \in G$

$$1 - (1-\lambda)^n \leq \kappa^{(n)}(x) \leq 1 + (1-\lambda)^{n-1} \left(\|\kappa\|_\infty - 1 \right).$$

Proceeding by induction we first note that this is evident for $n = 1$ and assume it to hold for some $n \in \mathbb{N}$. We can write $\kappa = (1-\lambda)\kappa_0 + \lambda 1$ for some continuous $\kappa_0 \geq 0$ with $\int \kappa_0 dm = 1$. Then

$$\begin{aligned} \kappa_0 * \kappa^{(n)}(x) &= \int_G \kappa_0(y) \kappa^{(n)}\left(y^{-1}x\right) dm(y) \\ &\leq \max_G \kappa^{(n)} \int_G \kappa_0(y)\, dm(y) = \max_G \kappa^{(n)} \end{aligned} \tag{14}$$

and similarly $\kappa_0 * \kappa^{(n)}(x) \geq \min_G \kappa^{(n)}$. Since $1 * \kappa^{(n)} = \int \kappa^{(n)} dm = 1$ it follows that

$$\begin{aligned} \kappa^{(n+1)}(x) &= (1-\lambda)\kappa_0 * \kappa^{(n)} + \lambda 1 * \kappa^{(n)} \\ &\in \left[(1-\lambda) \min \kappa^{(n)} + \lambda, (1-\lambda) \max \kappa^{(n)} + \lambda \right] \\ &\subseteq \left[1 - (1-\lambda)^{n+1}, 1 + (1-\lambda)^n \left(\|\kappa\|_\infty - 1 \right) \right], \end{aligned}$$

where the induction hypothesis was used in the last step. This proves the claim and also the inequality (13) with $C = 1 + \left(\|\kappa\|_\infty - 1 \right) / (1-\lambda)$ and $c = -\ln(1-\lambda)$.

Now consider the general case where $\kappa \geq 0$ with $\kappa > 0$ on a generating neighborhood U of G. We claim that $\kappa^{(n)} > 0$ on $U^n := \{u_1 u_2 ... u_n : u_i \in U\}$. Proceeding by induction again, we first note that the case $n = 1$ is part of the hypotheses and assume the claim to be valid for arbitrary n. Let $x \in U^{n+1}$, $x = zu$ with $z \in U^n$ and $u \in U$. We have

$$\kappa^{(n+1)}(x) = \int_G \kappa^{(n)}\left(zuy^{-1}\right) \kappa(y)\, dm(y).$$

The integrand on the right is nonnegative and, by the induction hypothesis and continuity, strictly positive for y in some neighborhood of u. Since, by compactness, non-empty open sets have positive Haar measure we conclude that $\kappa^{(n+1)}(x) > 0$ on U^{n+1}, proving the claim by induction.

Since G is compact we have $G = \bigcup_{n=1}^{N} U^n$ for some integer N. So $\kappa^{(N)} > 0$ throughout G and by the first part

$$\left\|\kappa^{(nN)} - 1\right\|_\infty \leq C'e^{-c'n}, \forall n \in \mathbb{N},$$

for some constants C' and c'. By an argument analogous to (14) the range of $\kappa^{(n)} = \kappa^{(n-\lfloor n/N \rfloor N)} * \kappa^{(\lfloor n/N \rfloor N)}$ is contained in $\left[\min \kappa^{(\lfloor n/N \rfloor N)}, \max \kappa^{(\lfloor n/N \rfloor N)}\right]$, so that

$$\left\|\kappa^{(n)} - 1\right\|_\infty \leq \left\|\kappa^{(\lfloor n/N \rfloor N)} - 1\right\|_\infty \leq C'e^{-c' \lfloor n/N \rfloor} \leq C'e^{-c'(n/N-1)},$$

which gives (13) with $C = C'e^{c'}$ and $c = c'/N$. □

Corollary 3. *If $\kappa > 0$ on a generating neighborhood of G then the process S is exponentially β-mixing.*

Proof. Let B be a Borel subset of $G \times G$. With $\mu_{\{0,\tau\}}$ being the joint distribution of S_0 and S_τ we have

$$\begin{aligned} \left|\mu_{\{0,\tau\}}(B) - m \times m(B)\right| &= \left|\int_B \left(\kappa^{(\tau)}\left(xy^{-1}\right) - 1\right) dm(x)\, dm(y)\right| \\ &\leq \left\|\kappa^{(\tau)} - 1\right\|_\infty \leq Ce^{-c\tau}. \end{aligned}$$

This proves that $\beta_S(\tau) \leq Ce^{-c\tau}$ □

The joint distribution $\mu_{\{0,1\}}$ of S_0 and S_1 is given by the density $g(x, y) = \kappa\left(x, y^{-1}\right)$ w.r.t $m \times m$, so the autoergodic categories are given by those Borel sets $E \subset G$ which satisfy

$$(1/2)\left(\kappa\left(xy^{-1}\right) + \kappa\left(yx^{-1}\right)\right) \geq 1$$

for all $x, y \in E$ with $x \neq y$. If E is autoergodic, then so is the right translate Ex for any $x \in G$, but not necessarily the left translate xE or the inverse E^{-1}. If E is autoergodic and open, then the right translates Ex of E cover G. By compactness there must be a finite subcover. Since every subset of an autoergodic set is autoergodic, removing all overlaps then leads to a partitioning

$$G = \bigcup_{k=1}^{N} E_k$$

of the state space into autoergodic categories, to which Theorem 5 applies.

Let G_0 be a proper closed subgroup such that

$$(1/2)\left(\kappa\left(xy^{-1}\right) + \kappa\left(yx^{-1}\right)\right) > 1 \tag{15}$$

for all $x, y \in G_0$. Using the topological separation properties in G one can show that there is a neighborhood W of the identity in G such that (15) holds for all $x, y \in G_0W$, so that G_0W is an open autoergodic set and left-invariant under the subgroup G_0. Covering G with right translates of G_0W and removing the overlaps then leads to a disjoint partitioning of G as above, with the additional property that all the categories E_k are left-invariant under the subgroup G_0.

5.3 Examples

As a very practical example consider a fixed, but very large image with periodic boundary conditions. The learners perspective on this image will be a small subimage parametrized by a point on the 3-torus $G = [0,1)^3$. Members of G are written as triplets (x, y, r), the group operation is componentwise addition modulo 1, and the Haar measure is just Lebesgue measure on $[0,1)^3$.

The interpretation is that x, y, r are parameters which define, respectively, the horizontal and vertical position of the center, and the orientation of the learners viewing frame. For any $(x, y, r) \in G$ the sensory map ϕ then takes the contents of the viewing window as a vector of fixed dimension and performs any desired preprocessing to return the stimulus vector $\phi(x, y, r) \in H$.

We then choose a diffusion kernel κ, which should be larger than 1 on a neighborhood of the identity to simultaneously ensure β-mixing and allow the existence of nontrivial autoergodic sets, and start the diffusion process.

The properties of the projection P returned by our algorithm depend more on the nature of the diffusion kernel κ than on the precise contents of the image.

Suppose that we can write $\kappa(x, y, r) = 0$ if (x, y) is outside of a small neighborhood V of the identity $(0,0)$ in $[0,1)^2$. The center of the viewing window will then move rather slowly in small steps bounded by V. The autoergodic categories will be subsets of $V \times [0,1)$ and are characterized by smeared positions in the image, with a certain small tolerance of translations. There will be some tolerance to rotation, depending on the behavior of $\kappa(0,0,r)$ which controls the velocity of diffusion in the orientation component. If $\kappa(0,0,r) > 1, \forall r$, then, as shown above, there is an open autoergodic subset of $V \times [0,1)$ which is invariant under the action of the rotation subgroup. The algorithm will then look for a projection P to distinguish translates of this subset, which means distinguishing subimages regardless of their orientation. The output P should thus perform as a rotation invariant preprocessor.

As shown in the next section, this prediction was confirmed experimentally.

This model can easily be extended to accommodate scale invariance, by passing to the 4-torus $G = [0,1)^4$. The fourth parameter s of $(x, y, r, s) \in G$ now defines the scale of the viewing window within an interval of a minimal scale a and a maximal scale b according to the formula

$$\operatorname{scale}(s) = \begin{cases} a + 2s(b-a) & \text{if } s < 1/2 \\ a + 2(1-s)(b-a) & \text{if } 1/2 \leq s \end{cases}, \tag{16}$$

which maps the stationary distribution to the uniform distribution on $[a, b]$. By controlling κ appropriately, rotation invariance can be replaced by scale-invariance (which also worked very well in the experiments), and one can attempt to learn preprocessors for combined rotation and scale invariance.

The scale function above illustrates another important point: The sensory map need not bear any relation to the algebraic structure of the group. It just needs to be continuous. We could therefore include other bounded parameters to control possible linear or nonlinear deformations of the viewing window in order to train corresponding invariances. The group structure is immaterial to the learner, it just gives us a handle on the mixing and autoergodicity properties of the process.

6 Experiments

We describe some experiments in the field of image recognition, where the projection is trained from a simulated process modeled after the processes described in the previous section, and then applied to problems of pattern recognition. In all cases the categories of the test problems were unknown to the learning system at the time when the projection was trained.

6.1 An algorithm with bounded memory

In practice H will be finite-dimensional. If the process X is slowly mixing, the learning time m can be quite large, leading to excessive storage requirements for any kind of batch algorithm. For this reason we used an online algorithm for principal subspace analysis, to which every successive realization of the operator valued variable $A_t = (1-\alpha) Q_{X_t} - \alpha Q_{\dot{X}_t}$ was fed, for $t = 1, ..., m$. This takes us somewhat astray from the results proved in section (3) and would require a different analysis in terms of stochastic approximation theory (see Benveniste et al [1]), but the principal goal of our experiments was to test the value of our objective function L.

If $\mathbf{v} = (v_1, ..., v_d)$ is an orthonormal basis for the range of some $P \in \mathcal{P}_d$, the Oja-Karhunen flow [12], is given by the ordinary differential equation

$$\dot{v}_k = (I - P_{\mathbf{v}}) T v_k,$$

where $P_{\mathbf{v}}$ is the projection onto the span of the v_k. If T is symmetric it has been shown by Yan et al [22] that a solution $\mathbf{v}(t)$ to this differential equation will remain forever on the Stiefel-manifold of orthonormal sets if the initial condition is orthonormal, and that it will converge to a dominant eigenspace of T for almost all initial conditions. Discretizing gives the update rule

$$v_k(t+1) = v_k(t) + \eta(t) \left(I - P_{\mathbf{v}(t)}\right) T v_k(t),$$

where $\eta(t)$ is a learning rate. Unfortunately a careful analysis shows that the Stiefel manifold becomes unstable if T is not positive. The simplest solution to this problem lies in orthonormalization. This is what we do, but there are more elegant techniques and different flows have been proposed (see e.g. [6]) to extract dominant eigenspaces for general symmetric operators. We now replace $T = E[A_t]$ by the process variable A_t to obtain the final rule

$$v_k(t+1) = v_k(t) + \eta(t) \left(I - P_{\mathbf{v}(t)}\right) \left((1-\alpha) Q_{X_t} - \alpha Q_{\dot{X}_t}\right) v_k(t), \qquad (17)$$

which, together with the orthonormalization prescription, gives the algorithm used in our experiments. The update rule (17) can be considered a combination of Hebbian learning of input data with anti-Hebbian learning of input velocity.

6.2 Image data and parametrization

We applied our technique to train preprocessors for image recognition. In all experiments the above algorithm we used the output dimension $d = 10$, the trade-off parameter $\alpha = 0.2$, a dynamic learning rate of $\eta(t) = \frac{10^2}{10^4 + t}$ and $m = 10^6$ observations.

To train the algorithm we were using processes S modeled after the examples described in section 5 to generate sequences of images. For the experiments with character recognition the images had 28×28 pixels and for face recognition they had 92×112 pixels. Correspondingly please substitute either 28×28 or 92×112 for the parameter dim in the sequel. The images were normalized to unity in the euclidean norm of $\mathbb{R}^{\dim}$.

We considered two possible architectures as fixed preprocessors on the pixel vectors: in the linear case we used the pixel vectors directly as inputs to our algorithm, so that $H = \mathbb{R}^{\dim}$.

In the nonlinear case (RBF) we used our algorithm to train the second layer of a two-layered radial-basis-function network. Define a kernel K on $\mathbb{R}^{\dim} \times \mathbb{R}^{\dim}$ by

$$K(\zeta_1, \zeta_2) = \exp\left(-4 \|\zeta_1 - \zeta_2\|_{\dim}^2\right).$$

For n_π prototypes $\pi_i \in \mathbb{R}^{\dim}$ the first network layer implements the nonlinear map $\psi : \mathbb{R}^{\dim} \to \mathbb{R}^{n_\pi}$ given by

$$\psi(\zeta)_k = \sum_{j=1}^{n_\pi} G_{kj}^{-1/2} K(\pi_j, \zeta), \text{ for } \zeta \in \mathbb{R}^{\dim},$$

where G is the Gramian $G_{ij} = K(\pi_i, \pi_j)$, which is generically non-singular. The transformation through $G_{kj}^{-1/2}$ is chosen to ensure that $\langle \psi(\pi_i), \psi(\pi_j) \rangle_{n_\pi} = K(\pi_i, \pi_j)$. We then applied the algorithm to the output of the first layer, so $H = \mathbb{R}^{n_\pi}$.

The number n_π and choice of the prototypes π_i for the first layer depended on the type of process used to generate the image sequence. For the experiments with face recognition, where the process was driven from a supervised learning task as in section 5.1, the available images in the training set were used. For the training of invariances from diffusion processes as in section 5.2 and section 6.3 below, $n_\pi = 2000$ was used throughout and the π_i were chosen directly from the process at time intervals larger than the mixing time and kept fixed afterwards.

Note that this type of preprocessing does not make any use of the known geometric relationships between image pixels, so that the same results would be obtained under any fixed permutation of the pixel indices.

6.3 Experiments with geometric invariants

The processes are modeled as in section 5.2. We took a large image $\mathcal{J}$ (a 1334×1078 gray-scale photography of a double page of the IEEE transactions on neural networks) with periodic boundary conditions. At any time t the 28x28-process image $\mathcal{J}(S_t)$ is a mapped subimage of $\mathcal{J}$ and completely described by the four parameters $S_t = (x_t, y_t, r_t, s_t) \in G = [0,1)^4$, where x_t and y_t specify the (appropriately scaled) position, $2\pi r_t$ the rotation angle and s_t the scale according to formula (16) in the interval $[a = 1/2, b = 3/2]$.

As the diffusion kernel κ to govern the random motion of the subimage we took a product $\kappa(x, y, r, s) = \kappa_x(x)\,\kappa_y(y)\,\kappa_r(r)\,\kappa_s(s)$, where each of the four factors is a centered normal density (or more precisely the pull-forward of a centered normal density under the map $\omega \in \mathbb{R} \mapsto \omega \bmod 1 \in [0,1)$), so κ is completely specified by the four variances σ_x^2, σ_y^2, σ_r^2 and σ_s^2. We set $\sigma_x = 1/(2 \times 1334)$ and $\sigma_y = 1/(2 \times 1078)$, so that $\sigma_x = \sigma_y = 1/2$ in pixel units. The center of the subimage therefore moves by about half a pixel on each time step.

For the experiments with rotation invariance we set $\sigma_r = 1$ and $\sigma_s = 1/50$. The distribution of κ_r is then nearly uniform on $[0,1) \bmod 1$, while the distribution of κ_s is very concentrated. This leads to rapidly changing orientation accompanied by rather small changes in scale.

For the experiments with scale invariance we set $\sigma_r = 1/50$ and $\sigma_s = 1$, causing small changes in orientation and large changes in scale.

The group-valued process S then gives rise the $\mathbb{R}^{\dim}$ valued process $\mathcal{J}(S_t)$, which is either directly fed into the algorithm (*Linear*) or further processed by the RBF-layer ψ as described above (*RBF*). In the linear case the sensory map is $\phi = \mathcal{J}$, in the RBF-case it is $\phi = \psi \circ \mathcal{J}$.

The performance of the resulting preprocessors is tested on a real life problem, the rotation- (scale-)-invariant recognition of characters. To this end two test-sets were prepared containing images of the digits 0-8 (0-9) in 100 randomly chosen states of orientation (scaling between 1/2 and 3/2).

An important criterion for the quality of a preprocessor is the ability of the distance between preprocessed and projected examples to serve as a detector for class-equality. Figure 1 shows corresponding receiver-operating-characteristics. The area under these curves then estimates the probability that for four independently drawn examples $\|a_1 - b_1\| \leq \|a_2 - b_2\|$, given that a_1 and b_1 belong to the same, and a_2 and b_2 to different classes, where the a_i, b_i are either unprocessed inputs (*Raw*), the projected inputs (*Linear*) or the projected outputs of the RBF layer (*RBF*). We also give a practical measure by recording the error rate of a *single-example-per-class* nearest-neighbor classifier, trained on a randomly selected example for each pattern class and averaged over 10 random selections of the training examples, *Error* in table 1.

In the case of rotation invariance, the linear preprocessor architecture even slightly outperformed the RBF network. The latter showed stable good performance in both cases.

It is important to realize that the classes to be recognized (the digits) were disclosed only after the projection had already been learnt, and that similar results are to be expected for any set of categories with the same invariance properties, such as rotation or scale-invariant capital letters or greek symbols.

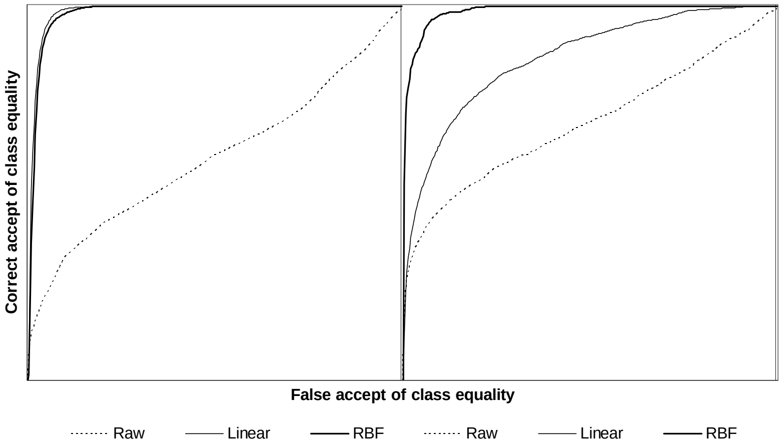

Fig. 1. ROC curves for the metric as a detector of class-equality for (left) rotation- and (right) scale-invariant character recognition. Every point on each curve corresponds to a distance threshold. The axes are the fractions of pairs from distinct categories (horizontal axis) and equal categories (vertical axis) which are mapped closer than this threshold.

6.4 Experiments with faces

For an experiment with face recognition we used the corresponding ATT-face dataset, which contains 10 facial images each of 40 persons. We used subject 1-35 to train the projection and 36-40 for the test. The process selected the training images as described in section 5.1: At time t an image was chosen with probability $(1-\lambda)$ from the uniform distribution of images representing the same subject as the previous image, and with probability λ from the uniform distribution in the entire training set. The value $\lambda = 1/20$ was used throughout.

On the test set we measured the ROC curve for the distance of represented examples as a detector for class-equality, and the error rates of nearest-neighbor classifiers trained from single randomly chosen examples for every subject in the test set, similar to the other experiments reported above. The results are reported in table 2. As there is no overlap of subjects between training and test set, we are effectively testing the algorithms capabilities of meta-generalization.

7 Conclusion

We presented a technique where an unsupervised learner can exploit short-time dependencies in stationary processes to learn a low dimensional data representation. Some of the theoretical questions related to our approach are settled and first experiments are very encouraging, but there are many open problems.

Table 1. Comparison of representation quality without training (Raw Data), and with training for unprocessed data (Linear) and preprocessed data (RBF)

Invariance Type	Method used	ROC-Area	Error
Rotation	Raw Data	0.597	0.716
	Linear	0.987	0.126
	RBF	0.983	0.138
Scaling	Raw Data	0.690	0.508
	Linear	0.866	0.421
	RBF	0.989	0.100

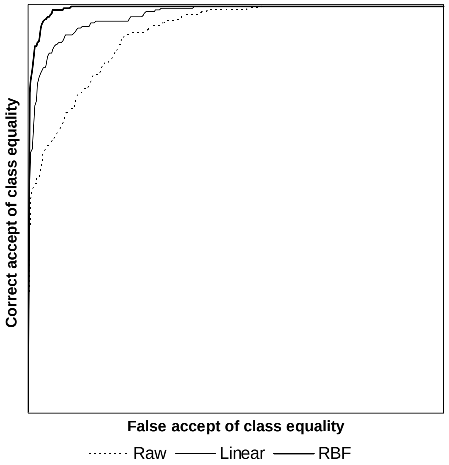

Fig. 2. ROC curves for the metric as a detector of class-equality for face recognition

One of the most important theoretical ones concerns a possible weakening of the autoergodicity requirement in Theorem 5, perhaps at the expense of stronger constraints of the sensory map ϕ. Another interesting direction is extending the technique from projections to more general Hilbert-Schmidt operators. The essential message of Theorem 5 is independent of the type of distortion function used, so that a variety of different methods to learn such functions can be tried.

The design of efficient, memory-bounded implementations is an important practical problem

Finally, it will be interesting to see the results of experiments conducted with real-world processes, perhaps similar to the models in section 5.

Acknowledgements: The author is grateful to the reviewers for their careful examination of the paper and their encouraging remarks.

Table 2. Face recognition results

Method used	ROC-Area	Error
Raw Data	0.934	0.113
Linear	0.977	0.044
RBF	0.996	0.017

References

1. A. Benveniste, M. Métevier, Pierre Priouret. *Adaptive algorithms and stochastic approximations.* Springer, 1987.
2. H. Bauer. *Measure and integration theory.* De Gruyter, Berlin, 2001.
3. R. C. Bradley. Basic properties of strong mixing conditions. A survey and open questions. *Probability surveys*, 2: 107-144, 2005.
4. P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3: 194-200, 1991.
5. A. C. Lozano, S. R. Kulkarni, R. E. Shapire. Convergence and consistency of regularized boosting algorithms with stationary, β-mixing observations. *Advances in Neural Information Processing Systems* 18, 2006.
6. J.H. Manton, U. Helmke, I.M.Y. Mareels. A dual purpose principal and minor component flow. *Systems & Control Letters* 54: 759-769, 2005.
7. A. Maurer, Bounds for linear multi-task learning. *JMLR*, 7:117–139, 2006.
8. A. Maurer, Unsupervised slow subspace learning from stationary processes. J.L. Balcázar, P.M. Long, F. Stephan (Eds.): *ALT 2006*, LNAI 4264: 363-377, 2006.
9. Colin McDiarmid, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.
10. R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39, 5-34, 2000.
11. A. Nobel, A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters* 17: 169-172, 1993.
12. E. Oja. Principal component analysis. *The Handbook of Brain Theory and Neural Networks.* M. A. Arbib ed. MIT Press, 910-913, 2002.
13. W. Rudin. *Functional analysis.* McGraw-Hill, 1973.
14. 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.
15. J. Shawe-Taylor, N. Christianini, Estimating the moments of a random vector, *Proceedings of GRETSI 2003 Conference*, I: 47–52, 2003.
16. M. Reed, B. Simon. *Functional Analysis*, part I of *Methods of Mathematical Physics, Academic Press*, 1980.
17. E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants.* Springer 2000.
18. B. Simon. *Trace Ideals and Their Applications.* Cambridge University Press, London, 1979
19. J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, J.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.
20. M. Vidyasagar, *Learning and generalization with applications to neural networks.* Springer, London, 2003.

21. L. Wiskott, T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14: 715-770, 2003.
22. W. Yan, U. Helmke, J.B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Trans. on Neural Networks* 5,5: 674-683, 1994.
23. B. Yu. Rate of convergence for empirical processes of stationary mixing sequences. Annals of Probability 22, 94-116, 1994.
24. Laurent Zwald, Olivier Bousquet and Gilles Blanchard. Statistical Properties of Kernel Principal Component Analysis, *COLT* 2004: 594-608, 2004.

Appendix: Frequently used notation

Notation	Short Description	Section
Ω	state space	1, 2.1
Σ	σ-algebra on Ω, events	2.1
$\Sigma_1 \otimes \Sigma_2$	smallest σ-algebra containing $\{E_1 \times E_2 : E_i \in \Sigma_i\}$	
Σ^I for $I \subseteq \mathbb{Z}$	product σ-algebra $\otimes_{i \in I} \Sigma$	2.1
$S = (S_t)_{t \in \mathbb{Z}}$	stationary process with values in Ω	1, 2.1
μ_I for $I \subseteq \mathbb{Z}$	joint distribution of $(S_t)_{t \in I}$ on $\left(\Omega^{\lvert I \rvert}, \Sigma^I\right)$	2.1
1_E	function $= 1$ on E, zero outside	
$\beta_S(\tau)$	mixing coefficient for process S and time τ	2.1
H	real, separable Hilbert space	1, 2.2
$\langle .,. \rangle$ and $\lVert . \rVert$	inner product and norm on H	2.2
ϕ	sensory map $\phi : \Omega \to H$	1, 2.2
$X = (X_t)_{t \in \mathbb{Z}}$	the process $X_t = \phi(S_t)$	1, 2.2
$\dot{X} = \left(\dot{X}_t\right)_{t \in \mathbb{Z}}$	the velocity process $\dot{X}_t = X_t - X_{t-1}$	1, 2.2
H_2	Hilbert-Schmidt operators on H	2.2
$\langle .,. \rangle_2$ and $\lVert . \rVert_2$	inner product and norm on H_2	2.2
$\mathcal{P}_d$	d-dimensional orthogonal projections in H	1, 2.2
Q_x, for $x \in H$	operator $Q_x z = \langle z, x \rangle x$, $\forall z \in H$	2.2
L_α or L	true objective functional	1
$\hat{L}_\alpha$ or $\hat{L}$	empirical objective functional	1
T	operator for true objective	1 (1)
$\hat{T}$	operator for empirical objective	1 (2)
$\lVert V \rVert$ for $V \subset H$	$\lVert V \rVert = \sup_{v \in V} \lVert v \rVert$	3
$\lvert \langle V, W \rangle \rvert$	$\lvert \langle V, W \rangle \rvert = \sup_{v \in V, w \in W} \lvert \langle v, w \rangle \rvert$	3
A_t	operator-valued process	3 (7)