Journal of Machine Learning Research 1 (2001) 179-209 Submitted 01/01; Published 6/01

Regularized Principal Manifolds

Alexander J. Smola Alex.Smola@anu.edu.au
Department of Engineering and RSISE
Australian National University
Canberra, ACT 0200, Australia

Sebastian Mika mika@first.gmd.de
GMD FIRST
Kekuléstr. 7
12489 Berlin, Germany

Bernhard Schölkopf* bs@conclu.de
Department of Engineering
Australian National University
Canberra, ACT 0200, Australia

Robert C. Williamson Bob.Williamson@anu.edu.au
Department of Engineering
Australian National University
Canberra, ACT 0200, Australia

Editor: Douglas H. Fisher (joint publication with www.kernel-machines.org)

Abstract

Many settings of unsupervised learning can be viewed as quantization problems - the minimization of the expected quantization error subject to some restrictions. This allows the use of tools such as regularization from the theory of (supervised) risk minimization for unsupervised learning. This setting turns out to be closely related to principal curves, the generative topographic map, and robust coding.

We explore this connection in two ways: (1) we propose an algorithm for finding principal manifolds that can be regularized in a variety of ways; and (2) we derive uniform convergence bounds and hence bounds on the learning rates of the algorithm. In particular, we give bounds on the covering numbers which allows us to obtain nearly optimal learning rates for certain types of regularization operators. Experimental results demonstrate the feasibility of the approach.

Keywords: Regularization, Uniform Convergence, Kernels, Entropy Numbers, Principal Curves, Clustering, generative topographic map, Support Vector Machines, Kernel PCA

1. Introduction

The problems of unsupervised learning are much less precisely defined than those of supervised learning. Usually no explicit cost function exists by which the hypothesis can be compared with training data. Instead, one has to make assumptions on the data, with respect to which questions may be asked.

*. Current address: Barnhill Technologies, Blumenstr. 21, 70794 Filderstadt, Germany

- A possible problem is: "Which properties of the data can be extracted with high confidence?" Or, in other words, which feature-extracting functions can be found among a given class with, say, unit variance and zero mean, and moreover whose properties will not change too much on unseen data. This leads to a *feature extracting* approach of unsupervised learning. Kernel principal component analysis (Schölkopf et al., 1998) is such an algorithm.

- Another question is: "Which properties describe the data best?" This means that one is looking for a *descriptive* model of the data, thus also a (possibly quite crude) model of the underlying probability distribution. Generative models like principal curves (Hastie and Stuetzle, 1989), the generative topographic map (Bishop et al., 1998), several linear Gaussian models (Roweis and Ghahramani, 1999), or also simple vector quantizers (Bartlett et al., 1998) are examples thereof.

We will study the second type of model in the present paper. Since many problems of unsupervised learning can be formalized in a quantization-functional setting (see section 2), this will allow us to use techniques from regularization theory. In particular, it leads to a natural generalization (to higher dimensionality and different criteria of regularity) of the principal curves algorithm with a length constraint (Kégl et al., 2000), presented in section 3 together with an efficient algorithm (section 5).

We also show that regularized quantization functionals can be seen in the context of robust coding, i.e., optimal coding in the presence of a noisy channel. This is achieved by using an idea of Bishop (1995), who explored this connection in the context of supervised learning. Another connection can be drawn to the generative topographic map (GTM) (Bishop et al., 1998), which essentially differs in the choice of a regularizer and the Bayesian probabilistic underpinning of the algorithm (section 6).

The quantization-functional approach also provides a versatile tool for stating uniform convergence bounds. In section 7 we derive bounds on the quantization error in terms of L_∞ covering numbers for the corresponding classes of functions. By using functional analytic tools (the details are relegated to Appendix A.1) we are able to bound the rate of convergence by $O(m^{-\frac{1}{2}+\alpha})$ for arbitrary positive α, where m is the number of examples seen (section 7.4). For some kernels this improves on the rate of Kégl et al. (2000).

We finish with section 8, giving some experimental results demonstrating the feasibility of our approach and a discussion.

2. The Quantization Error Functional

The idea of the quantization error functional approach is that one tries to obtain interesting information about the data at hand by encoding it in a compressed, yet meaningful form. The quality of this code is assessed by the reconstruction error (the quantization error) it causes, i.e., how close the reconstruction comes to the initial data, and the simplicity of the device having generated the code. The latter is important, since the coding device will then contain the information we seek to extract. Contrary to most engineering applications, we will also allow for *continuous* codes. This reflects our emphasis on information extraction by learning the coding device itself.

Denote by $\mathcal{X}$ a (compact subset of a) vector space and $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$ a dataset drawn iid (independent identically distributed) from an unknown underlying probability distribution $\mu(x)$. Moreover consider index sets $\mathcal{Z}$, maps $f : \mathcal{Z} \to \mathcal{X}$, and classes $\mathcal{F}$ of such maps (with $f \in \mathcal{F}$).

Here the map f is supposed to describe some basic properties of $\mu(x)$. In particular, one seeks f such that the so-called quantization error

$$R[f] := \int_{\mathcal{X}} \min_{z \in \mathcal{Z}} c(x, f(z)) \mathrm{d}\mu(x) \tag{1}$$

is minimized. In this setting $c(x, f(z))$ is the cost function determining the error of reconstruction. Very often one sets $c(x, f(z)) = \|x - f(z)\|^2$, where $\|\cdot\|$ denotes the Euclidean distance. Unfortunately, the problem of minimizing $R[f]$ is unsolvable, as μ is generally unknown. Hence one replaces μ by the empirical measure

$$\mu_m(x) := \frac{1}{m} \sum_{i=1}^{m} \delta(x - x_i) \tag{2}$$

and instead of (1) analyzes the empirical quantization error

$$R_{\mathrm{emp}}^m[f] := R_{\mathrm{emp}}[f] := \int_{\mathcal{X}} \min_{z \in \mathcal{Z}} c(x, f(z)) \mathrm{d}\mu_m(x) = \frac{1}{m} \sum_{i=1}^{m} \min_{z \in \mathcal{Z}} c(x_i, f(z)). \tag{3}$$

The general problem of minimizing (3) is ill-posed (Tikhonov and Arsenin, 1977, Morozov, 1984). Even worse - with no further restrictions on $\mathcal{F}$, small values of $R_{\mathrm{emp}}[f]$ do not guarantee small values of $R[f]$ either. Many problems of unsupervised learning can be cast in the form of finding a minimizer of (1) or (3). Let us consider some practical examples.

Example 1 (Sample Mean) *Define $\mathcal{Z} := \{1\}$, $f \in \mathcal{X}$, and $\mathcal{F}$ to be the set of all constant functions. Moreover set $c(x, f(z)) = \|x - f(z)\|^2$. Then the minimum of*

$$R[f] := \int_{\mathcal{X}} \|x - f\|^2 \mathrm{d}\mu(x) \tag{4}$$

yields the variance of the data and the minimizers of the quantization functionals can be determined analytically:

$$\operatorname*{argmin}_{f \in \mathcal{F}} R[f] = \int_{\mathcal{X}} x \mathrm{d}\mu(x) \text{ and } \operatorname*{argmin}_{f \in \mathcal{F}} R_{\mathrm{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} x_i. \tag{5}$$

Example 2 (k-Means Vector Quantization) *Define $\mathcal{Z} := \{1, \ldots, k\}$, $f : i \to f_i$ with $f_i \in \mathcal{X}$, $\mathcal{F}$ to be the set of all such functions, and again $c(x, f(z)) = \|x - f(z)\|^2$. Then*

$$R[f] := \int_{\mathcal{X}} \min_{z \in \{1, \ldots, k\}} \|x - f_z\|^2 \mathrm{d}\mu(x) \tag{6}$$

denotes the canonical distortion error of a vector quantizer. In practice one can use the k-means algorithm (Lloyd, 1982) to find a set of vectors $\{f_1, \ldots, f_k\}$ minimizing $R_{\mathrm{emp}}[f]$. Also here (Bartlett et al., 1998), one can prove convergence properties of (the minimizer) of $R_{\mathrm{emp}}[f]$ to (one of) the minimizer(s) of $R[f]$.

Note that, in this case, minimization of the empirical quantization error leads to local minima, a problem quite common in this type of setting. A different choice of cost function c leads to a clustering algorithm proposed by Bradley et al. (1997).

Example 3 (k-Median and Robust Vector Quantization) *With the definitions of the previous example and $c(x, f(z)) = \|x - f(z)\|_1$ one obtains the k-median problem ($\| \cdot \|_1$ is the city-block metric). In this case,*

$$R[f] := \int_{\mathcal{X}} \min_{z \in \{1,\ldots,k\}} \|x - f_z\|_1 \mathrm{d}\mu(x). \tag{7}$$

This setting is robust against outliers, since the maximum influence of each pattern is bounded. An intermediate setting can be derived from Huber's robust cost function (Huber, 1981). Here we have

$$c(x, f(z)) = \begin{cases} \frac{1}{2\sigma}\|x - f(z)\|^2 & \textit{for } \|x - f(z)\| \leq \sigma \\ \|x - f(z)\| - \frac{\sigma}{2} & \textit{otherwise,} \end{cases} \tag{8}$$

for suitably chosen σ. Eq. (8) behaves like a k-means vector quantizer for small x_i, however with the built-in safeguard of limiting the influence of each single pattern.

Instead of discrete quantization one can also consider a mapping onto a manifold of lower dimensionality than the input space. PCA can be viewed in this way (Hastie and Stuetzle, 1989):

Example 4 (Principal Components) *Define $\mathcal{Z} := \mathbb{R}$, $f : z \to f_0 + z \cdot f_1$ with $f_0, f_1 \in \mathcal{X}$, $\|f_1\| = 1$, and $\mathcal{F}$ to be the set of all such line segments. Moreover let $c(x, f(z)) := \|x - f(z)\|^2$. Then the minimizer of*

$$R[f] := \int_{\mathcal{X}} \min_{z \in [0,1]} \|x - f_0 - z \cdot f_1\|^2 \mathrm{d}\mu(x) \tag{9}$$

over $f \in \mathcal{F}$ yields a line parallel to the direction of largest variance in $\mu(x)$ (Hastie and Stuetzle, 1989).

A slight modification results in simultaneous diagonalization of the covariance matrix with respect to a different metric tensor.

Example 5 (Transformed Cost Metrics) *Denote by D a symmetric positive definite matrix. With the definitions as above and the cost function*

$$c(x, f(z)) := (x - f(z))^\top D^{-1} (x - f(z)) \tag{10}$$

the minimizer of the empirical quantization can be found by simultaneous diagonalization of D and the covariance matrix $\mathrm{cov}(x)$.

This can be seen as follows. Replace x by $\tilde{x} := D^{-\frac{1}{2}}x$ and f by $\tilde{f} := D^{-\frac{1}{2}}f$. Now $c(x, f(z)) = \|\tilde{x} - \tilde{f}(z)\|^2$, hence we reduced the problem to the one of finding principal components for the covariance matrix $D^{-\frac{1}{2}}\mathrm{cov}(x)D^{-\frac{1}{2}}$. This, however, is equivalent to simultaneous diagonalization of D and $\mathrm{cov}(x)$, which proves the above remark.

Further choices of c based on either the $\|\cdot\|_1$ metric or Huber's robust loss function lead to solutions that are less prone to instabilities caused by outliers than standard PCA.

A combination of the k-means clustering and principal components immediately recovers the k-planes clustering algorithm proposed by Bradley and Mangasarian (1998), also known as *Local PCA* by Kambhatla and Leen (1994, 1997).[1] There, clustering is carried out with respect to k planes instead of simply k cluster points. After an assignment of the data points to the planes, the latter are re-estimated by using PCA (i.e., the directions with smallest variance are eliminated). Both Kambhatla and Leen (1997) and Bradley & Mangasarian show that this can improve results on certain datasets.

Hastie and Stuetzle (1989) extended PCA in a different direction by allowing other than linear functions $f(z)$:

Example 6 (principal curves and Surfaces) *Denote by $\mathcal{Z} := [0,1]^d$ (with $d \in \mathbb{N}$ and $d > 1$ for principal surfaces), $f : z \to f(z)$ with $f \in \mathcal{F}$ be a class of continuous $\mathbb{R}^d$-valued continuous functions (possibly with further restrictions), and again $c(x, f(z)) := \|x - f(z)\|^2$. The minimizer of*

$$R[f] := \int_{\mathcal{X}} \min_{z \in [0,1]^d} \|x - f(z)\|^2 \mathrm{d}\mu(x) \tag{11}$$

is not well defined, unless $\mathcal{F}$ is a compact set. Moreover, even the minimizer of $R_{\mathrm{emp}}[f]$ is not well defined either, in general. In fact, it is an ill-posed problem in the sense of Tikhonov and Arsenin (1977). Until recently (Kégl et al., 2000), no uniform convergence properties of $R_{\mathrm{emp}}[f]$ to $R[f]$ could be stated.

Kégl et al. (2000) modified the original "principal-curves" algorithm in order to prove bounds on $R[f]$ in terms of $R_{\mathrm{emp}}[f]$ and to show that the resulting estimate is well defined. The changes imply a restriction of $\mathcal{F}$ to polygonal lines with a fixed number of knots and, most importantly, *fixed* length L.[2]

This is essentially equivalent to using a regularization operator. Instead of a length constraint, which, as we will show in section 3.2, corresponds to a particular regularization operator, we now consider more general smoothness constraints on the estimated curve $f(x)$.

3. A Regularized Quantization Functional

The overall requirement is for estimates that not only yield small expected quantization error but are also smooth curves (or manifolds) where the "smoothness" is *independent* of

[1] While Kambhatla and Leen (1997) introduces the problem by considering local linear versions of principal component analysis and takes a neural networks perspective, Bradley et al. (1997) treat the task mainly as an optimization problem for which convergence to a local minimum in a finite number of steps is proven. While the resulting algorithm is identical, the motivation in the two cases differs significantly. In particular, the ansatz of Bradley et al. (1997) makes it easier for us to formulate the problem as one of minimizing a quantization functional.

The original local linear vector quantization formulation put forward by Kambhatla and Leen (1994) would allow us to give a quantization formulation for local PCA as well. There we would simply consider linear subspaces together with their enclosing Voronoi cells.

[2] In practice Kegl et al. use a constraint on the angles of a polygonal curve rather than the actual length constraint to achieve sample complexity rate bounds on the training time of the algorithm. For the uniform convergence part, however, the length constraint is used.

the parameterization of the curve. In general this is difficult to achieve. An easier task is to work with a measure of smoothness of f *depending* on the parameterization of $f(z)$. A wide range of regularizers from supervised learning can be readily used for this purpose. As a side-effect we obtain a smooth parameterization.

We now propose a variant of minimizing the empirical quantization functional which seeks hypotheses from certain classes of smooth curves, leads to an algorithm that is readily implemented, and is amenable to the analysis of sample complexity via uniform convergence techniques. We will make use of a regularized version of the empirical quantization functional. Let

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \lambda Q[f], \tag{12}$$

where $Q[f]$ is a convex nonnegative regularization term and $\lambda > 0$ is a trade-off constant determining how much *simple* functions f should be favoured over functions with low empirical quantization error. We now consider some possible choices of Q.

3.1 Quadratic Regularizers

A common choice of regularizers are quadratic functionals as proposed by Tikhonov and Arsenin (1977), i.e.,

$$Q[f] = \frac{1}{2}\|Pf\|^2. \tag{13}$$

Here P is a regularization operator penalizing unsmooth functions f via a mapping into a dot product space (e.g., a reproducing kernel Hilbert space (Kimeldorf and Wahba, 1971, Wahba, 1979, 1990)). In this case one obtains

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \frac{\lambda}{2}\|Pf\|^2 = \frac{1}{m}\sum_{i=1}^{m} \min_{z \in \mathcal{Z}} \|x_i - f(z)\|^2 + \frac{\lambda}{2}\|Pf\|^2. \tag{14}$$

As we will show in section 4, if one requires certain invariances regarding the regularizer to hold, one need only consider a special class of operators P (scalar ones).

Using the results of Smola et al. (1998) regarding the connection between regularization operators and kernels, it appears suitable to choose a kernel expansion of f matching the regularization operator P; i.e., for any $x_i, x_j \in \mathcal{X}$,

$$\langle Pk(x_i, \cdot), Pk(x_j, \cdot)\rangle = k(x_i, x_j). \tag{15}$$

Such functions k can be found as the Greens functions of P^*P (see Girosi et al. (1995), Smola et al. (1998), Girosi (1998)). Finally, assume that if f_0 is a constant function then $(P^*P)(f_0) = 0$. This assumption leads to translation invariance of the problem - any shifting operation can be counterbalanced by a constant offset. For an expansion like

$$f_{(\alpha_1,\ldots,\alpha_m)}(z) = f(z) = f_0 + \sum_{i=1}^{M} \alpha_i k(z_i, z) \text{ where } z_i \in \mathcal{Z},\ \alpha_i \in \mathcal{X}, k : \mathcal{Z}^2 \to \mathbb{R} \tag{16}$$

with some previously chosen nodes $z_1, \ldots, z_M$ (of which one takes as many as one may afford in terms of computational cost) the regularization term can be written as

$$\|Pf\|^2 = \sum_{i,j=1}^{M} \langle \alpha_i, \alpha_j \rangle k(z_i, z_j). \tag{17}$$

This is the functional form of $\|Pf\|^2$ we need (and will use) to derive efficient algorithms.

3.2 Examples of Regularization Operators

The first example considers the equivalence between principal curves with a length constraint and minimizing the regularized quantization functional.

Example 7 (Regularizers with a Length Constraint) *By choosing $P := \partial_z$, i.e., the differentiation operator, $\|Pf\|^2$ becomes an integral over the squared "speed" of the curve. Reparameterizing f to constant speed leaves the empirical quantization error unchanged, whereas the regularization term is minimized. This can be seen as follows: by construction $\int_{[0,1]} \|\partial_z f(z)\| \mathrm{d}z$ does not depend on the (re)parameterization. The variance, however, is minimal for a constant function, hence $\|\partial_z f(z)\|$ has to be constant over the interval $[0, 1]$. Thus $\|Pf\|^2$ equals the squared length L^2 of the curve at the optimal solution.*

However, minimizing the empirical quantization error plus a regularizer is equivalent to minimizing the empirical quantization error for a fixed value of the regularization term (for λ adjusted suitably). Hence the proposed algorithm is equivalent to finding the optimal curve with a length constraint, i.e., it is equivalent to the algorithm proposed by Kégl et al. (2000).[3]

As experimental and theoretical evidence from regression indicates, it may be beneficial to choose a kernel enforcing higher degrees of smoothness in higher derivatives of the estimate as well.

Example 8 (Gaussian Kernels) *Here one has*

$$k(x, x') = \exp\left(-\tfrac{\|x-x'\|^2}{2\sigma^2}\right). \tag{18}$$

This corresponds to a regularizer penalizing all orders of derivatives simultaneously. Yuille and Grzywacz (1988) show that this kernel corresponds to the pseudo-differential operator defined by

$$\|Pf\|^2 = \int \mathrm{d}x \sum_{n=0}^{\infty} \frac{\sigma^{2n}}{n! 2^n} (D^n f(x))^2 \tag{19}$$

with $D^{2n} = \Delta^n$ and $D^{2n+1} = \nabla\Delta^n$, Δ being the Laplacian and ∇ the gradient operator. This means that one is looking not only for smooth functions but also curves whose curvature and other higher-order properties change slowly.

(We used a Gaussian kernel in the experiments reported below).

Finally the use of periodical kernels (see Smola et al. (1998)) allows one to model circular structures in $\mathcal{X}$.

Example 9 (Periodical Kernels) *Possible kernels would be of the type*

$$k(x, x') = \sum_{j=1}^{\infty} c_j \cos\left(\tfrac{2\pi j}{\tau} x\right) \cos\left(\tfrac{2\pi j}{\tau} x'\right) + s_j \sin\left(\tfrac{2\pi j}{\tau} x\right) \sin\left(\tfrac{2\pi j}{\tau} x'\right) \tag{20}$$

[3] The reasoning is slightly incorrect for the case of a finite number of basis functions - there f *cannot* be completely reparameterized to constant speed. However the basic properties still hold, provided the number of kernels is sufficiently high.

where τ is the periodicity and c_j, s_j are positive and absolutely summable coefficients, or

$$k(x, x') = \sum_{j=-\infty}^{\infty} \tilde{k}(x - x' - j\tau) \tag{21}$$

where $\tilde{k}$ is some non-periodic kernel. The latter formulation may be numerically more advantageous if $\tilde{k}$ is a translation invariant kernel with compact support or rapid decay (e.g., Gaussian RBF) as the series then can be truncated after a few terms.

For more details on regularization operators see e.g., (Girosi et al., 1995, Smola et al., 1998, Girosi, 1998). Essentially one may use any kernel introduced in support vector machines (Vapnik, 1998), Gaussian processes (Williams, 1998), or reproducing kernel Hilbert spaces (Wahba, 1990) in the expansions described above.

The appealing property of this formulation is that it is completely independent of the dimensionality and particular structure of $\mathcal{Z}$.

3.3 Linear Programming Regularizers

It may not always be desirable to find expansions of $f = \sum_{i=1}^{M} \alpha_i k(z_i, \cdot)$ in terms of *many* basis functions $k(z_i, \cdot)$. Instead it would be better if one obtained a (nearly as good) estimate of f with just a few basis functions. This can be achieved via a regularizer enforcing sparsity, e.g., by setting

$$Q[f] := \sum_{i=1}^{M} |\alpha_i|. \tag{22}$$

(For $\alpha_i \in \mathbb{R}^d$ use $\|\alpha_i\|_1 = \sum_{j=1}^{d} |\alpha_{ij}|$ instead of $|\alpha_i|$.) Such approaches have been studied by Mangasarian (1969), Chen et al. (1999), Weston et al. (1999), Girosi (1998), Bennett (1999), Frieß and Harrison (1998), Smola (1998) in various settings such as wavelet expansions, mathematical programming, or support vector machines. We will show (section A.2) that by using an argument similar to the one of Smola et al. (2000) this setting allows efficient capacity control, too.

4. Invariant Regularizers

In the previous section it was claimed that in many cases one could restrict oneself to the class of scalar regularization operators, i.e., operators that act on each component of f (independent of the choice of the basis) separately and identically. This is the case, provided some basic assumptions about scaling behaviour and permutation symmetry are imposed.

Proposition 1 (Homogeneous Invariant Regularization) *Any regularizer $Q[f]$ that is both homogeneous quadratic and invariant under an irreducible orthogonal representation ρ of a group $\mathcal{G}$ on $\mathcal{X}$, i.e., satisfies*

$$\begin{aligned} Q[f] &\geq 0 \text{ for all } f \in \mathcal{F} && (23)\\ Q[af] &= a^2 Q[f] \text{ for all scalars } a && (24)\\ Q[\rho(g)f] &= Q[f] \text{ for all } \rho(g) \in \mathcal{G} && (25) \end{aligned}$$

is of the form $Q[f] = \langle Pf, Pf \rangle$ where P is a scalar operator.

Proof It follows directly from (24) and Euler's "homogeneity property", that $Q[f]$ is a quadratic form, thus $Q[f] = \langle f, Mf \rangle$ for some operator M. Moreover M can be written as P^*P since it has to be a positive operator (see (23)). Finally from

$$\langle Pf, Pf \rangle = \langle P\rho(g)f, P\rho(g)f \rangle \tag{26}$$

and the polarization equation (i.e., exploiting (26) for $f + f'$, $f - f'$, subtracting both terms, and making use of the symmetry of $\langle \cdot, \cdot \rangle$) it follows that

$$\langle Pf, Pf' \rangle = \langle P\rho(g)f, P\rho(g)f' \rangle \text{ for all } f, f' \in \mathcal{F}. \tag{27}$$

Hence by virtue of the Fischer-Riesz theorem one obtains $P^*Pf = \rho(g)^*P^*P\rho(g)f$. However, $\rho(g)\rho(g)^* = 1$ since ρ is a unitary representation of $\mathcal{G}$ and therefore

$$P^*P\rho(g) = \rho(g)P^*P \tag{28}$$

has to hold for any $\rho(g) \in \mathcal{G}$. Finally, by virtue of Schur's lemma (see e.g., (Hamermesh, 1962)) P^*P may only be a scalar operator. Therefore, without loss of generality, also P may be assumed to be scalar. ■

The requirement (24) may seem artificial, however, it is not. In particular when stating uniform convergence bounds in terms of entropy numbers (see Appendix A) a regularizer with these properties is desirable: entropy numbers scale linearly when f is multiplied by a scalar. Therefore one wants homogeneous scaling behaviour of some degree, say 2, as in (24). A consequence of the proposition above is that there exists no vector-valued regularization operator satisfying the invariance conditions. Hence there is no need to look for other operators P in the presence of a sufficiently strong invariance. Now for a practical application of proposition 1.

Corollary 2 (Permutation and Rotation Symmetries) *Under the assumptions of Proposition 1 both the canonical representation of the permutation group (by permutation matrices) in a finite dimensional vector space $\mathcal{Y}$ and the group of orthogonal transformations on $\mathcal{Y}$ enforce scalar operators P.*

This follows immediately from the fact that the representation of these groups is unitary and irreducible on $\mathcal{Y}$ by construction.

In other words, every time the nature of the data does not change when it undergoes a rotation or a permutation, i.e., there exists no particular ordering of the data in terms of features, one should use scalar operators P. Of course, this reasoning only applies to quadratic regularizers since for other types of regularization the operator P may not even be well defined.

5. An Algorithm for minimizing $R_{\text{reg}}[f]$

In this section we present an algorithm that approximately minimizes $R_{\text{reg}}[f]$ via coordinate descent. We certainly do not claim it is the best algorithm for this task - our modest goals were to find an algorithm consistent with our framework (which is amenable to sample complexity theory) and which seems to work in practice; the following algorithm meets these goals.

In the following we will assume the data to be centered and therefore drop the term f_0 in the expansion (16). This greatly simplifies the notation (the extension is straightforward). Moreover, for the sake of practicality, we will assume that the ansatz for f can be written in terms of a finite number of parameters $\alpha_1, \dots \alpha_M$ and that likewise the regularizer $Q[f]$ can also be expressed as a function of $\alpha_1, \dots, \alpha_M$. This allows us to rephrase the problem of minimizing the regularized quantization functional in the following form.

$$\min_{\substack{\{\alpha_1,\dots,\alpha_M\}\subset\mathcal{X}\\ \{\zeta_1,\dots,\zeta_m\}\subset\mathcal{Z}}} \left[\frac{1}{m}\sum_{i=1}^{m} c(x_i, f_{(\alpha_1,\dots,\alpha_M)}(\zeta_i)) + \lambda Q(\alpha_1,\dots,\alpha_M)\right]. \tag{29}$$

The minimization here is achieved in an iterative fashion by coordinate descent over ζ and α. It operates analogously to the EM (expectation maximization) algorithm (Dempster et al., 1977): there the aim is to find (the parameters θ of) a distribution $p_\theta(x, l)$ where x are observations and l are latent variables. Keeping θ fixed one proceeds by maximizing $p_\theta(x, l)$ with respect to l. The M-step consists of maximizing $p_\theta(x, l)$ with respect to θ. These two steps are repeated until no further improvement can be achieved.

Likewise one iterates over minimizing (29) with respect to $\{\zeta_1, \dots, \zeta_m\}$, equivalent to the E-step (projection), and then with respect to $\{\alpha_1, \dots, \alpha_M\}$, corresponding to the M-step (adaptation). This is repeated until convergence, or, in practice, until the regularized quantization functional does not decrease significantly further. Let us have a closer look at the individual phases of the algorithm.

5.1 Projection

For each $i \in \{1, \dots, m\}$ choose

$$\zeta_i := \operatorname*{argmin}_{\zeta\in\mathcal{Z}} c(x_i, f(\zeta)); \tag{30}$$

e.g., for squared loss $\zeta_i := \operatorname{argmin}_{\zeta\in\mathcal{Z}} \|x_i - f(\zeta)\|^2$. Clearly, for fixed α_i, the so chosen ζ_i minimizes the loss term in (29), which in turn is equal to $R_{\text{reg}}[f]$ for given α_i and X. Hence $R_{\text{reg}}[f]$ is decreased while keeping $Q[f]$ fixed (the latter is the case since the variables α_i do not change). In practice one uses standard low-dimensional nonlinear function minimization algorithms (see Press et al. (1992) for details and references) to achieve this goal.

The computational complexity is $O(m \cdot M)$ since the minimization step has to be carried out for each sample separately (m). Moreover each function evaluation (whose number we assumed to be approximately constant per minimization) scales linearly with the number of basis functions (M).

5.2 Adaptation

Now the parameters ζ_i are fixed and α_i is adapted such that $R_{\text{reg}}[f]$ decreases further. The design of practical algorithms to decrease $R_{\text{reg}}[f]$ is closely connected with the particular forms both the cost function $c(x, f(z))$ and the regularizer $Q[f]$ take. We will restrict ourselves to squared loss in this section (i.e., $c(x, f(z)) = \|x - f(z)\|^2$) and to the quadratic or linear regularization terms as described in section 3. We thus assume that

$$f(z) = \sum_{i=1}^{M} \alpha_i k(x_i, x) \tag{31}$$

for some kernel k, in the quadratic case, matching the regularization operator P.

Quadratic Regularizers The problem to be solved in this case is to minimize

$$\frac{1}{m} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{M} \alpha_j k(z_j, \zeta_i) \right\|^2 + \frac{\lambda}{2} \sum_{i,j=1}^{M} \langle \alpha_i, \alpha_j \rangle k(z_i, z_j) \tag{32}$$

with respect to α. Here, α and X denote the *matrices* of all parameters and samples, respectively. Differentiation of (32) with respect to α_i yields

$$\left(\frac{\lambda m}{2} K_z + K_\zeta^\top K_\zeta \right) \alpha = K_\zeta^\top X \text{ and hence } \alpha = \left(\frac{\lambda m}{2} K_z + K_\zeta^\top K_\zeta \right)^{-1} K_\zeta^\top X \tag{33}$$

where $(K_z)_{ij} := k(z_i, z_j)$ is an $M \times M$ matrix and $(K_\zeta)_{ij} := k(\zeta_i, z_j)$ is $m \times M$.

The computational complexity of the adaptation step is $O(M^2 \cdot m)$ for the matrix computation and $O(M^3)$ for the computation of the parameters α_i. Assuming termination of the overall algorithm in a finite number of steps (independent of M and m) we showed that the overall complexity of the proposed algorithm is $O(M^3) + O(M^2 \cdot m)$; i.e., it scales only linearly in the number of samples (but cubic in the number of parameters).[4]

Linear Regularizers Here the adaptation step can be solved via a quadratic optimization problem. The trick is to break up the ℓ_1 norms of the coefficient vectors α_i into pairs of nonnegative variables $\alpha_i - \alpha_i^*$, thus replacing $\|\alpha_i\|_1$ by $\langle \alpha_i, \vec{1} \rangle + \langle \alpha_i^*, \vec{1} \rangle$ where $\vec{1}$ denotes the vector of d ones. Consequently one has to minimize

$$\frac{1}{m} \sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{M} (\alpha_j - \alpha_j^*) k(z_j, \zeta_i) \right\|^2 + \lambda \sum_{i=1}^{M} \langle \alpha_i + \alpha_i^*, \vec{1} \rangle \tag{34}$$

under the constraint that α_i, α_i^* live in the positive orthant in $\mathcal{X}$. Optimization is carried out by standard quadratic programming codes (e.g., Murtagh and Saunders (1983), IBM Corporation (1992), Vanderbei (1997)). This has (depending on the particular implementation of the algorithm) a similar order of complexity as a matrix inversion, i.e., the calculations to solve the unconstrained quadratic optimization problem described previously.

An algorithm iterating between the projection and adaptation step as described above will generally decrease the regularized risk term and eventually reach a local minimum of the optimization problem.[5] What remains is to find good starting values.

[4] Note that also the memory requirements are at least $O(M \cdot m)$ and that for optimal performance M should increase with increasing m.

[5] $R_{\text{reg}}[f]$ is bounded from below by 0, hence any decreasing series of $R_{\text{reg}}[f_i]$ where f_i denotes the estimate at step i has a limit which then will be a global, or most likely a local minimum. Note that this does not guarantee that we will reach the minimum in a *finite* number of steps.

5.3 Initialization

The idea is to choose the coefficients α_i such that the initial guess of f approximately points into the directions of the first D principal components given by the matrix $V := (v_1, \ldots, v_D)$. This is done analogously to the initialization in the generative topographic map (Bishop et al., 1998, eq. (2.20)). Choose

$$\alpha^{(0)} = \operatorname*{argmin}_{\alpha=(\alpha_1,\ldots,\alpha_M)\subset\mathcal{X}} \frac{1}{M}\sum_{i=1}^{M} c\left(V(z_i - z_0) - f_{(\alpha_1,\ldots,\alpha_M)}(z_i)\right) + \lambda Q[f]. \tag{35}$$

Hence for squared loss and quadratic regularizers, $\alpha^{(0)}$ is given by the solution of the linear system $\left(\frac{\lambda}{2}\mathbf{1} + K_z\right)\alpha = V(Z - Z_0)$ where Z denotes the matrix of z_i, z_0 the mean of z_i, and Z_0 the matrix of m identical copies of z_0 correspondingly. If not dealing, as assumed, with centered data, set f_0 to the sample mean; i.e., $f_0 = \frac{1}{m}\sum_{i=1}^{m} x_i$.

6. Relations to Other Algorithms

6.1 The Connection to the GTM

Just considering the basic algorithm of the GTM (without the Bayesian framework or its interpretation in terms of generative models), one can observe that its goal is to minimize a quantity similar to $R_{\text{reg}}[f]$. More precisely, it maximizes the posterior probability of the data having been generated by a lower-dimensional discrete grid $\mathcal{Z} := \{z_1, \ldots z_M\} \subset \mathbb{R}^D$, mapped into $\mathcal{X}$, and corrupted by additive Gaussian noise (this is where the squared loss enters).

The difference lies in its choice of $\mathcal{Z}$, set to be identical with the points z_i in our setting (no distinction is made between z_i and the points generating the basis functions k) and the *probabilistic* assignment of x_i to $\mathcal{Z}$, compared to the deterministic assignment in the projection step of section 5.1: several "nodes" may be "responsible" for having generated a particular datapoint x_i. The latter is computationally tractable in the GTM setting, since the cardinality of $\mathcal{Z}$ is finite (and small). For uncountable $\mathcal{Z}$, such an assignment could be approximated by sampling from the resulting distribution or variational calculations, which might render the algorithm more efficient in finding a good local minimum (cf. simulated annealing).

A further difference can be found in the choice of a regularizer which is of ℓ_2 type. In other words, instead of using $\frac{1}{2}\|Pf\|^2$ (Bishop et al., 1998), choose $\frac{1}{2}\sum_{i=1}^{M}\|\alpha_i\|_2^2$ as a regularizer. This may not always be favourable (Smola et al., 2000), since by increasing the number of basis functions, uniform convergence bounds for these classes of functions become less tight. In fact, it has been observed that in the GTM (Bishop et al., 1998, Sec. 2.3) the number of nodes M (for the kernel expansion) is a critical parameter.

A quadratic regularizer as proposed in section 3.2 does not exhibit this weakness since it takes a *coupling* between the single centers of the basis functions $k(z_i, z_j)$ into account, which helps to avoid overfitting. It is worthwhile noticing that such a modification could also be applied to the original GTM algorithm. This would correspond to a Gaussian Process (Williams, 1998) having created the manifold (where the prior over all such manifolds f is determined by the covariance function k). A detailed description of such a modification is beyond the scope of the current work and is thus omitted.

6.2 Robust Coding and Regularized Quantization

From a mere coding point of view it might not seem too obvious at first glance to seek very smooth curves. In fact, one could construct a space-filling curve (e.g., a Peano curve). This ensures one could achieve zero empirical and expected quantization error, by exploiting the fact that codewords may be specified to arbitrary precision. However, the codebook in this setting would have to be *exact* and the resulting estimate f would be quite useless for any practical purposes.

The subsequent reasoning explains why such a solution f would not be desirable from a learning-theory point of view either. Let us modify the situation slightly and introduce a noisy channel, i.e., the reconstruction would not occur for

$$\zeta(x) = \operatorname*{argmin}_{\zeta \in \mathcal{Z}} c(x, f(\zeta)) \tag{36}$$

but for the random variable $\hat{\zeta}(x)$ with

$$\hat{\zeta}(x) := \operatorname*{argmin}_{\zeta \in \mathcal{Z}} c(x, f(\zeta)) + \xi. \tag{37}$$

Here ξ is another random variable which is symmetrically distributed with zero mean according to $p(\xi)$ and has finite variance σ^2. Consider the minimization of a slightly different risk functional

$$R_{\text{noise}}[f] := \int_{\mathcal{X} \times \mathbb{R}} c\left(x, f\left(\operatorname*{argmin}_{z \in \mathcal{Z}} c(x, f(z)) + \xi\right)\right) \mathrm{d}\mu(x) \mathrm{d}p(\xi). \tag{38}$$

This modified setting rules out space-filling curves such as the Peano curve. Equation (38) is inspired by the problem of robust vector quantization (see Gersho and Gray (1991)) and the proof of Bishop (1995) that in supervised learning training with noise is equivalent to Tikhonov regularization. It is an adaptation of these techniques that we will use to derive a similar result in unsupervised learning.

Assume now that $c(\cdot, \cdot)$ is the squared loss. If the overall influence of ξ is small, the moments of higher order than two are essentially negligible (for small ξ), and if f is twice differentiable, one may expand f in a Taylor expansion with $f(\zeta + \xi) \approx f(\zeta) + \xi f'(\zeta) + \frac{\xi^2}{2} f''(\zeta)$. Using the reasoning of Bishop (1995) one arrives at

$$\begin{aligned} R_{\text{noise}}[f] &\approx R[f] + 2 \int_{\mathbb{R}} \xi^2 dp(\xi) \int_{\mathcal{X}} \left\| f'(\zeta) \right\|^2 + \frac{1}{2} \langle f(\zeta) - x, f''(\zeta) \rangle \mathrm{d}\mu(x) \\ &= R[f] + 2\sigma^2 \int_{\mathcal{X}} \left\| f'(\zeta) \right\|^2 + \frac{1}{2} \langle f(\zeta) - x, f''(\zeta) \rangle \mathrm{d}\mu(x) \end{aligned} \tag{39}$$

where ζ is defined as in (36). Finally we expand f at the unbiased solution f_0 (where $\sigma = 0$) in terms of σ^2. Consequently the second term in (39) inside the integral is only $O(\sigma^2)$, hence its overall contribution is only $O(\sigma^4)$ and can be neglected. What remains is

$$R_{\text{noise}}[f] \approx R[f] + 2\sigma^2 \int_{\mathcal{X}} \| f'(\zeta) \|^2 \mathrm{d}\mu(x) \text{ with } \zeta = \zeta(x) = \operatorname*{argmin}_{\zeta \in \mathcal{Z}} \| x - f(\zeta) \|^2. \tag{40}$$

Modulo the fact that the integral is with respect to x (and hence with respect to some complicated measure with respect to ζ), the second term is a regularizer enforcing smooth functions by penalizing the first derivative as discussed in section 3.2. Hence we recovered principal curves with a length constraint as a by-product of robust coding.

We chose not to use the discrete sample-size setting as done by Bishop (1995) since it appears not very practicable to use a training-with-input-noise scheme like in supervised learning to the problem of principal manifolds. The discretization of $R[f]$, i.e., its approximation by the empirical risk functional, is independent of this reasoning. It might be of practical interest, though, to use a probabilistic projection of samples onto the curve for algorithmic stability (as done for instance in simulated annealing for the k-means algorithm).

7. Uniform Convergence Bounds

We now determine bounds on the sample size sufficient to ensure that the above algorithm can find an f close to the best possible. We do this using methods which are very similar to those of Kégl et al. (2000) and are based on uniform (over a class of functions) convergence of empirical risk functionals to their expected value. The basic probabilistic tools we need are given in section 7.2. In section 7.3 we will state bounds on the relevant covering numbers for the classes of functions induced by our regularization operators:

$$\mathcal{F}_\Lambda := \{f : \mathcal{Z} \to \mathcal{X} : \ Q[f] \leq \Lambda\}. \tag{41}$$

Recall $Q[f] = \frac{1}{2}\|Pf\|^2$ and $\|Pf\|^2$ is given by (17). Since bounding covering numbers can be technically intricate, we will only state the results and basic techniques in the main body and relegate their proof and more detailed considerations to the appendix. Section 7.4 gives overall sample complexity rates.

In order to avoid technical complications arising from unbounded cost functions (like boundedness of some moments of the distribution $\mu(x)$ (Vapnik, 1982)) we will assume that there exists some $r > 0$ such that the probability measure of a ball of radius r is 1, i.e., $\mu(U_r) = 1$. Kégl et al. (2000) showed that under these assumptions also the prinicipal manifold f is contained in U_r, hence the quantization error will be no larger than $e_c := \max_{x,x' \in U_r} c(x, x')$ for all x. For squared loss we have $e_c = 4r^2$.

7.1 Preliminaries

We wish to derive bounds on the deviation between the empirical quantization error $R_{\text{emp}}[f]$ and the expected quantization error $R[f]$. In order to do this we will use uniform convergence bounds and to that end we utilize the ε-cover of the loss-function-induced class

$$\mathcal{F}^c_\Lambda := \mathcal{F}^c := \{(x, z) \mapsto c(x, f(z)) \colon f \in \mathcal{F}_\Lambda\} \tag{42}$$

on U_r. Given a metric ρ and a set $\mathcal{F}$, the ε covering number of $\mathcal{F}$, denoted by $\mathcal{N}(\varepsilon, \mathcal{F}, \rho)$ is the smallest number of ρ-balls of radius ε the union of which contains $\mathcal{F}$. A metric on $\mathcal{F}^c_\Lambda$ is defined by letting

$$d(f_c, f'_c) := \sup_{z \in \mathcal{Z}, x \in U_r} |c(x, f(z)) - c(x, f'(z))| \tag{43}$$

where $f, f' \in \mathcal{F}_\Lambda$.

Whilst d is the metric we are interested in, it is quite hard to compute covering numbers with respect to it directly. However, by an argument of Williamson et al. (1998), Anthony and Bartlett (1999), it is possible to upper-bound these quantities in terms of corresponding entropy numbers of the class of functions $\mathcal{F}_\Lambda$ itself if c is Lipschitz continuous. Denote by $l_c > 0$ a constant for which $|c(x, x') - c(x, x'')| \leq l_c \|x' - x''\|_2$ for all $x, x', x'' \in U_r$. In this case

$$d(f_c, f'_c) \leq l_c \sup_{z \in \mathcal{Z}} \|f(z) - f'(z)\|_2, \tag{44}$$

hence all we have to do is compute the $L_\infty(\ell_2^d)$ covering numbers of $\mathcal{F}$ to obtain the corresponding covering numbers of $\mathcal{F}_\Lambda^c$, with the definition of the norm on $\mathcal{F}$ as

$$\|f\|_{L_\infty(\ell_2^d)} := \sup_{z \in \mathcal{Z}} \|f(z)\|_{\ell_2^d}. \tag{45}$$

The metric is induced by the norm in the usual fashion. For the polynomial loss $c(x, f(z)) := \|x - f(z)\|_2^p$, one obtains $l_c = p(2r)^{p-1}$. With the definitions from above one can see immediately that $\mathcal{N}(\varepsilon, \mathcal{F}_\Lambda^c, d) \leq \mathcal{N}\left(\varepsilon / l_c, \mathcal{F}_\Lambda, L_\infty(\ell_2^d)\right)$.

7.2 Upper and Lower Bounds

The next two results are similar in their flavour to the bounds obtained by Kégl et al. (2000). They are slightly streamlined since they are independent of some technical conditions on $\mathcal{F}$ used by Kégl et al. (2000).

Proposition 3 ($L_\infty(\ell_2^d)$ bounds for Principal Manifolds)
Denote by $\mathcal{F}$ a class of continuous functions from $\mathcal{Z}$ into $\mathcal{X} \subseteq U_r$ and let μ be a distribution over $\mathcal{X}$. If m points are drawn iid from μ, then for all $\eta > 0, \varepsilon \in (0, \eta/2)$

$$\Pr\left\{\sup_{f \in \mathcal{F}} \left|R_{\text{emp}}^m[f] - R[f]\right| > \eta\right\} \leq 2\mathcal{N}\left(\tfrac{\varepsilon}{2l_c}, \mathcal{F}, L_\infty(\ell_2^d)\right) e^{-2m(\eta-\varepsilon)^2/e_c}. \tag{46}$$

Proof By definition of $R_{\text{emp}}^m[f] = \frac{1}{m} \sum_{i=1}^m \min_z \|f(z) - x_i\|^2$ the empirical quantization functional is an average over m iid random variables which are each bounded by e_c. Hence we may apply Hoeffding's inequality to obtain

$$\Pr\left\{\left|R_{\text{emp}}^m[f] - R[f]\right| > \eta\right\} \leq 2e^{-2m\eta^2/e_c}. \tag{47}$$

The next step is to discretize $\mathcal{F}_\Lambda^c$ by a $\frac{\varepsilon}{2}$ cover (i.e., $\mathcal{F}_\Lambda$ by a $\frac{\varepsilon}{2l_c}$ cover) with respect to the metric d: for every $f_c \in \mathcal{F}_\Lambda^c$ there exists some f_i in the cover such that $|R[f] - R[f_i]| \leq \frac{\varepsilon}{2}$ and $|R_{\text{emp}}^m[f] - R_{\text{emp}}^m[f_i]| \leq \frac{\varepsilon}{2}$. Consequently

$$\Pr\left\{\left|R_{\text{emp}}^m[f] - R[f]\right| > \eta\right\} \leq \Pr\left\{\left|R_{\text{emp}}^m[f_i] - R[f_i]\right| > \eta - \varepsilon\right\}. \tag{48}$$

Substituting (48) into (47) and taking the union bound over the $\frac{\varepsilon}{2}$ cover of $\mathcal{F}_\Lambda^c$ gives the desired result. ∎

This result is useful to assess the quality of an *empirically* determined manifold. In order to obtain rates of convergence we also need a result connecting the expected quantization error of the principal manifold f^*_{emp} minimizing $R^m_{\text{emp}}[f]$ and the manifold f^* with minimal quantization error $R[f^*]$.

Proposition 4 (Rates of Convergence for Optimal Estimates)
Suppose $\mathcal{F}$ is compact. Let $f^{*,m}_{\text{emp}} := \operatorname{argmin}_{f \in \mathcal{F}^c_\Lambda} R_{\text{emp}}[f]$ and $f^* := \operatorname{argmin}_{f \in \mathcal{F}^c_\Lambda} R[f]$. With the definitions and conditions of Proposition 3,

$$\Pr\left\{\sup_{f \in \mathcal{F}} \left|R[f^{*,m}_{\text{emp}}] - R[f^*]\right| > \eta\right\} \leq 2\left(\mathcal{N}\left(\tfrac{\varepsilon}{l_c}, \mathcal{F}_\Lambda, L_\infty(\ell_2^d)\right) + 1\right) e^{-\frac{m(\eta-\varepsilon)^2}{2e_c}}. \tag{49}$$

The proof is similar to that of proposition 3 and can be found in Appendix B.1.

7.3 Bounding Covering Numbers

After propositions 3 and 4, the missing ingredient to state uniform convergence bounds is a bound on the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, L_\infty(\ell_2^d))$. (For the remainder of this section we will simply write $\mathcal{N}(\varepsilon, \mathcal{F})$.)

Before going into details let us briefly review what already exists in terms of bounds on the covering number $\mathcal{N}$ for $L_\infty(\ell_2^d)$ metrics. Kégl et al. (2000) essentially show that

$$\log \mathcal{N}(\varepsilon, \mathcal{F}) = O(\tfrac{1}{\varepsilon}) \tag{50}$$

under the following assumptions: they consider polygonal curves $f(\cdot)$ of length L in the ball $U_r \subset \mathcal{X}$. The distance measure (no metric!) for $\mathcal{N}(\varepsilon)$ is defined as $\sup_{x \in U_r} |\Delta(x, f) - \Delta(x, f')| \leq \varepsilon$. Here $\Delta(x, f)$ is the minimum distance between a curve $f(\cdot)$ and $x \in U_r$.

By using functional analytic tools developed by Williamson et al. (1998) one can obtain results for more general regularization operators, which can then be used in place of (50) to obtain bounds on the expected quantization error. The technical details are in Appendix A. The key point is to characterize the simplicity (as measured by covering numbers) of the class of functions via the regularization term under consideration.

It turns out that a feature space representation of kernels k is useful in this regard. In particular, like (15) we can write any kernel $k(x, x')$ satisfying Mercer's condition (Mercer, 1909) as a dot product in some feature space (see Appendix A.2 for details) by

$$k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x'). \tag{51}$$

Here (λ_i, ϕ_i) is the eigensystem of the operator $T_k f := \int_{\mathcal{Z}} f(x') k(x', x) \mathrm{d}x'$. It is this notion of linear functionals introduced by (51) that allows us to treat nonlinear functions with ease.

Roughly speaking, if the λ_i decay rapidly, the possibly infinite expansion in (51) can be approximated with high precision by a low-dimensional space which means that we are effectively dealing only with simple function classes. This becomes more obvious in the following theorem:

Proposition 5 (Eigenvalues and Covering Numbers) *Suppose k is a Mercer kernel with eigenvalues (sorted in decreasing order) satisfying $\lambda_j = O(e^{-\alpha j^p})$ for some $\alpha, p > 0$. Then*

$$\log \mathcal{N}(\varepsilon, \mathcal{F}) = O\left(\log^{\frac{p+1}{p}} \varepsilon^{-1}\right). \tag{52}$$

Suppose k is a Mercer kernel with eigenvalues satisfying $\lambda_j = O(j^{-\alpha-1})$ for some $\alpha > 0$. Then

$$\log \mathcal{N}(\varepsilon, \mathcal{F}) = O\left(\varepsilon^{-\frac{\alpha}{2}+\delta}\right) \tag{53}$$

for any $\delta \in (0, \alpha/2)$.

Proof The rates follow immediately from propositions 12 and 13 and that $\mathcal{F}$ can be described by a linear operator. See Appendix A for details. ■

The rates obtained in proposition 5 are quite strong. In particular recall that for compact sets in finite dimensional spaces of dimension d the covering number is $\mathcal{N}(\varepsilon, \mathcal{F}) = O(\varepsilon^{-d})$ (Carl and Stephani, 1990). In view of (52) this means that even though we are dealing with a nonparametric estimator, it behaves almost as if it was a finite dimensional one.

All that is left is to substitute (52) and (53) into the uniform convergence results to obtain bounds on the performance of our learning algorithm. The slow growth in $\mathcal{N}(\varepsilon, \mathcal{F})$ is the reason why we will be able to prove fast rates of convergence below.

7.4 Rates of Convergence

Another property of interest is the sample complexity of learning principal manifolds. Kégl et al. (2000) have shown a $O(m^{-1/3})$ rate of convergence for principal curves ($d = 1$) with a length constraint regularizer. We prove that by using a more powerful regularizer (as one can do using our algorithm) one may obtain a bound of the form $O(m^{-\frac{\alpha}{2(\alpha+1)}})$ for polynomial rates of decay of the eigenvalues of k ($\alpha+1$ is the rate of decay); or $O(m^{-1/2+\beta})$ for exponential rates of decay (β is an arbitrary positive constant). It would be surprising if we could do any better given that supervised learning rates are typically no better than $O(m^{-1/2})$ (Anthony and Bartlett, 1999, Chapter 19). In the following we assume that $\mathcal{F}_\Lambda$ is compact; this is true of all the specific $\mathcal{F}_\Lambda$ considered above.

Proposition 6 (Learning Rates for Principal Manifolds) *Suppose $\mathcal{F}_\Lambda$ defined by (41) is compact. Define $f^{*,m}_{\text{emp}}, f^* \in \mathcal{F}_\Lambda$ as in Proposition 4.*
1. *If $\log \mathcal{N}(\varepsilon, \mathcal{F}^c_\Lambda) = O(\log^\alpha \frac{1}{\varepsilon})$ for some $\alpha > 0$ then*

$$R[f^{*,m}_{\text{emp}}] - R[f^*] = O(m^{-1/2} \log^{\alpha/2} m) = O(m^{-1/2+\beta}) \tag{54}$$

for any $\beta > 0$.
2. *If $\log \mathcal{N}(\varepsilon, \mathcal{F}^c_\Lambda) = O(\varepsilon^{-\alpha})$, for some $\alpha > 0$ then*

$$R[f^{*,m}_{\text{emp}}] - R[f^*] \leq O(m^{-\frac{1}{\alpha+2}}). \tag{55}$$

The proof can be found in Appendix B.2. A restatement of the optimal learning rates in terms of the eigenspectrum of the kernel leads to the following corollary.

Corollary 7 (Learning Rates for given Eigenspectra) *Suppose $\mathcal{F}_\Lambda$ is compact, $f^{*,m}_{\text{emp}}$, $f^* \in \mathcal{F}_\Lambda$ are as before, and λ_j are the eigenvalues of the kernel k inducing $\mathcal{F}_\Lambda$ (sorted in decreasing order). If there is a $c > 0$ such that for all $j \in \mathbb{N}$, $\lambda_j \leq e^{-cj^\alpha}$, then*

$$R[f^*_{\text{emp}}] - R[f^*] \leq O(m^{-1/2} \log^{\frac{\alpha+1}{2\alpha}} m). \tag{56}$$

If $\lambda_j = O(j^{-\alpha})$ for quadratic regularizers, or $\lambda_j = O(j^{-\alpha/2})$ for linear regularizers, then

$$R[f^{*,m}_{\text{emp}}] - R[f^*] \leq O(m^{-\frac{\alpha-1}{2\alpha}}). \tag{57}$$

Interestingly the above result is slightly weaker than the result by Kégl et al. (2000) for the case of length constraints, as the latter corresponds to the differentiation operator, thus polynomial eigenvalue decay of order 2, and therefore to a rate of $\frac{1}{4}$ (Kégl et al. (2000) obtain $\frac{1}{3}$). For a linear regularizer, though, we obtain a rate of $\frac{3}{8}$. It is unclear, whether this is due to a (possibly) suboptimal bound on the entropy numbers induced by k, or the fact that our results were stated in terms of the (stronger) $L_\infty(\ell_2^d)$ metric. This weakness, which is yet to be fully understood, should not detract from the fact that we *can* get better rates by using stronger regularizers, *and* our algorithm can utilize such regularizers.

8. Experiments

We now illustrate that the basic idea of the algorithm proposed in section 5 is sound by reporting the results of several experiments (figures 1, 3). In all cases Gaussian RBF kernels, as discussed in section 3.2, were used. First, we generated different data sets in 2 and 3 dimensions from 1 or 2 dimensional parameterizations. Then we applied our algorithm using the prior knowledge about the original parameterization dimension of the data set in choosing the latent variable space to have the appropriate size. For almost any parameter setting (λ, M, and width of basis functions) we obtained reasonable results.

We found that for a suitable choice of the regularization factor λ, a very close match to the original distribution can be achieved. Of course, the number and width of the basis functions had an effect on the solution, too. But their influence on the basic characteristics was quite small. Figure 2 shows the convergence properties of the algorithm. One can clearly observe that the overall regularized quantization error decreases for each step, while both the regularization term and the quantization error term are free to vary. This experimentally shows that the algorithm strictly decreases $R_{\text{reg}}[f]$ at every step and will eventually converge to a (local) minimum.

Given the close relationship to the GTM, we also applied our algorithm to the oil flow data set used by Bishop et al. (1998). The data set consists of 1000 samples from $\mathbb{R}^{12}$, organized into 3 classes. The goal is to visualize these data, so we chose the latent space to be $\mathcal{Z} = [-1, 1]^2$. We then generated the principal manifold and plotted the distribution of the latent variables for each sample (see figure 3). For comparison we did the same with principal component analysis (PCA). It can be seen, that the result achieved with principal manifolds reveals much more of the structure intrinsic to the data set than simply to search for directions with high variance. Comparing to (Bishop et al., 1998) we achieved a competitive result.

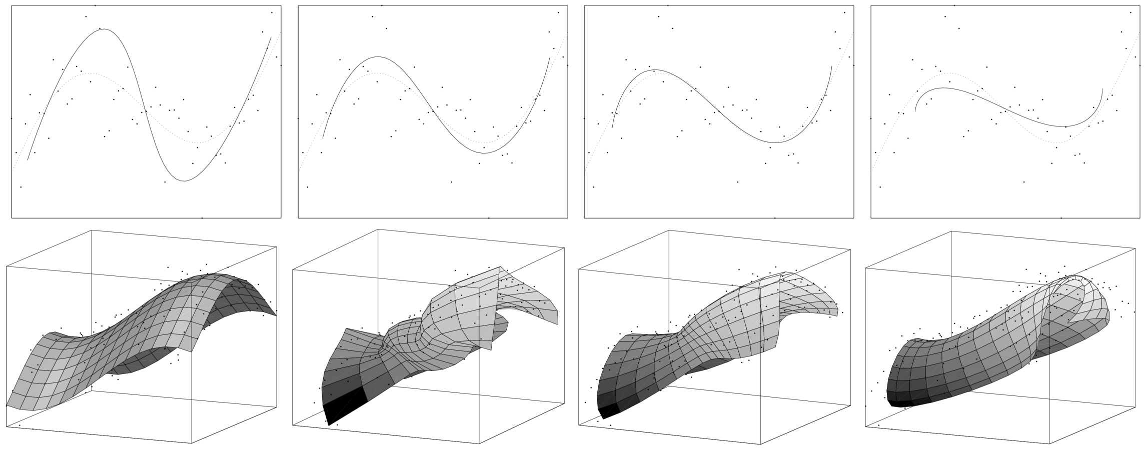

Figure 1: Upper 4 images: we generated a dataset (small dots) by adding noise to a distribution indicated by the dotted line. The resulting manifold generated by our approach is given by the solid line (over a parameter range of $\mathcal{Z} = [-1, 1]$). From left to right we used different values for the regularization parameter $\lambda = 0.1, 0.5, 1, 4$. The width and number of basis function was constant 1, and 10 respectively. Lower 4 images: here we generated a dataset by sampling (with noise) from a distribution depicted in the left-most image (small dots are the sampled data). The remaining three images show the manifold yielded by our approach over the parameter space $\mathcal{Z} = [-1, 1]^2$ for $\lambda = 0.001, 0.1, 1$. The width and number of basis functions was again constant (1 and 36).

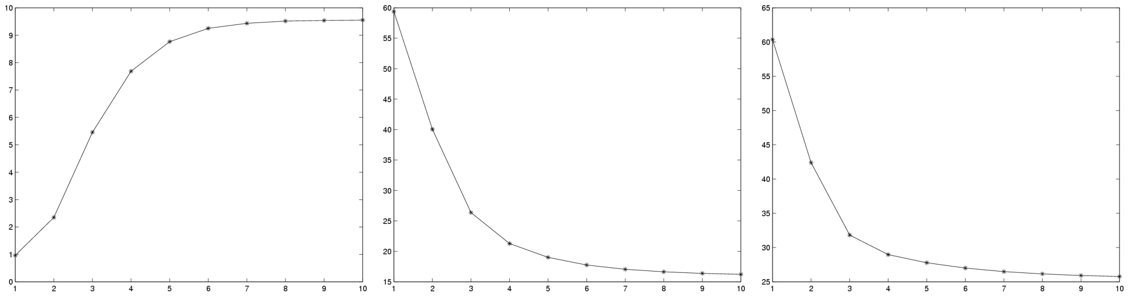

Figure 2: Left: regularization term, middle: empirical quantization error, right: regularized quantization error vs. number of iterations.

Finally, in order to demonstrate the fact that RPM can be applied to construct higher dimensional manifolds as well, we constructed a 3 dimensional latent variable space for the oil flow dataset (see fig. 4). In this setting the different flow regimes are much more apparent. Moreover it suggests a further subdivision of one class into 5 more regimes.

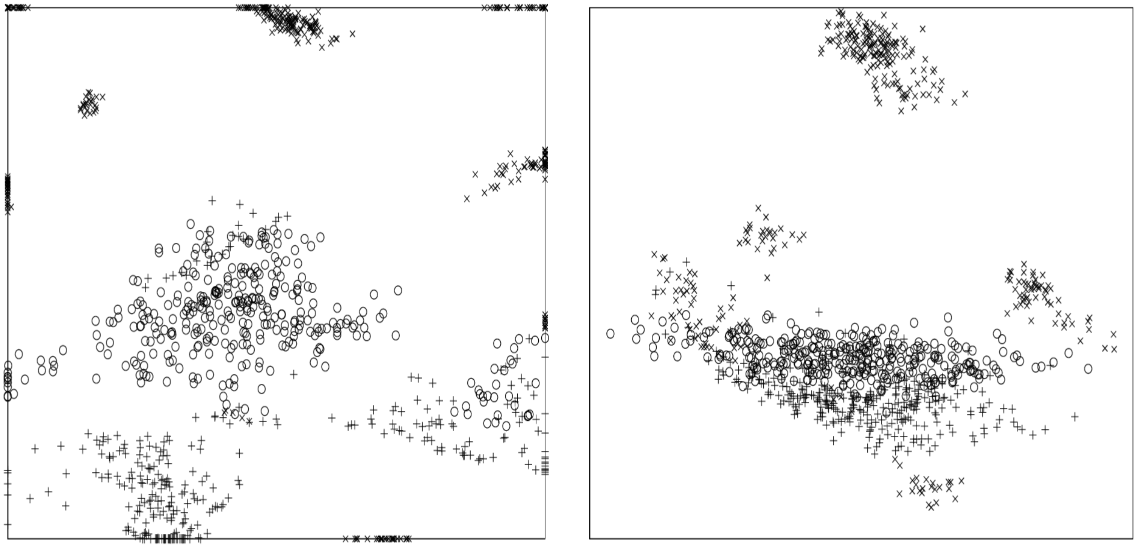

Figure 3: Organization of the latent variable space for the oil flow data set using principal manifolds (left, 49 nodes, kernel width 1, regularization 0.01) and principal component analysis (right). The lower-dimensional representation found by principal manifolds nicely reveals the class structure, comparably to the GTM. Linear PCA fails completely.

9. Conclusion

We proposed a framework for unsupervised learning that can draw on the techniques available in minimization of risk functionals in supervised learning. This yielded an algorithm suitable for obtaining principal manifolds. The expansion in terms of kernel functions and the treatment by regularization operators made it easier to decouple the algorithmic part (of finding a suitable manifold) from the part of specifying a class of manifolds with desirable properties. In particular, our algorithm does not crucially depend on the number of nodes used.

Bounds on the sample complexity of learning principal manifolds were given. These may be used to perform capacity control more effectively. Moreover our calculations have shown that regularized principal manifolds are a feasible way to perform unsupervised learning. The proofs relied on the function-analytic tools developed by Williamson et al. (1998).

There are several directions to take this research further; we mention the most obvious three. The algorithm could well be improved. In contrast to successful kernel algorithms such as SV machines, our algorithm here is not guaranteed to find a global minimum. Is it possible to develop an efficient algorithm that does? Furthermore the algorithm is related to methods that carry out a probabilistic assignment of the observed data to the manifold. The latter often exhibit improved numerical properties and the assignments themselves can be interpreted statistically. It would be interesting to exploit this fact in the present context. Finally, the theoretical bounds could be improved - hopefully achieving the same

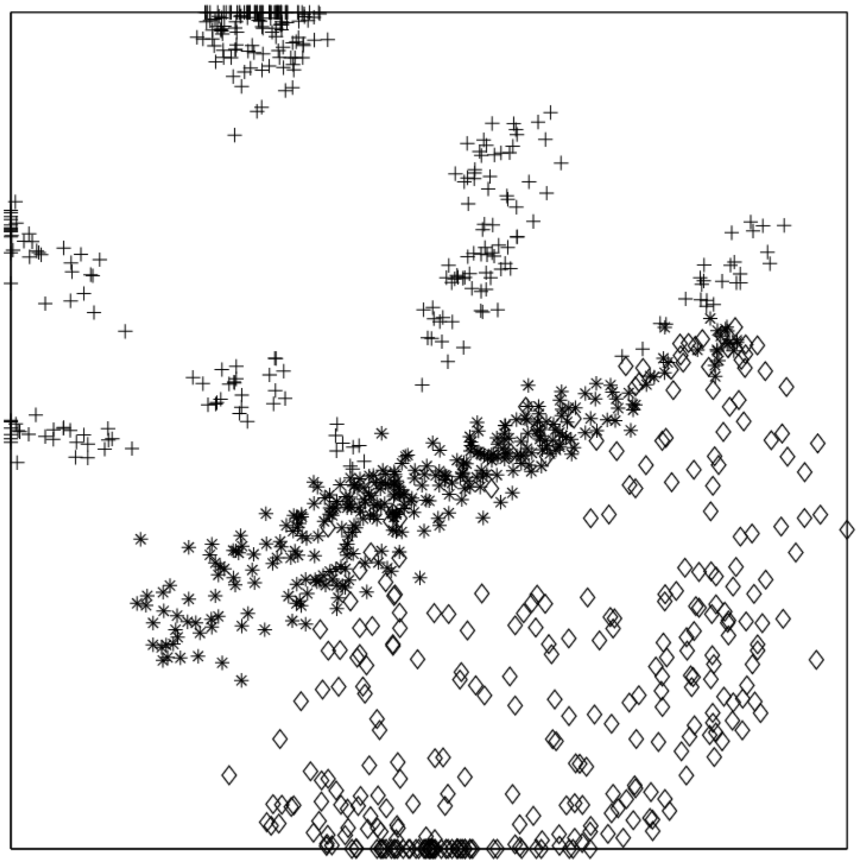

Figure 4: Organization of the latent variable space for the oil flow data set using principal manifolds in 3 dimensions with $6^3 = 216$ nodes, kernel width 1 and regularization 0.01. The three-dimensional latent variable space was projected onto 2 dimensions for display reasons. Observe the good separation between the different flow regimes. The map furthermore suggests that there exist 5 subdivisions of the regime denoted by '+'.

rate as Kégl et al. (2000) for their special case, while still keeping the better rates for more powerful regularizers.

Acknowledgments

This work was supported in part by grants of the ARC and the DFG (Ja 379/71,51,91,5-2, Sm 62/1-1). Parts of the work were done while AS and BS were still with GMD FIRST, Berlin. We thank Balázs Kégl and Adam Krzyżak for helpful comments and discussions.

Appendix A. Covering and Entropy Numbers

A.1 Entropy Numbers

Denote by $\mathfrak{L}(E, F)$ the set of all bounded linear operators T between two normed spaces $(E, \|\cdot\|_E)$, $(F, \|\cdot\|_F)$. The nth *entropy number of a set* $M \subset E$ relative to a metric ρ, for $n \in \mathbb{N}$, is

$$\varepsilon(M) := \inf\{\varepsilon : \mathcal{N}(\varepsilon, M, \rho) \leq n\}. \tag{58}$$

Similarly, the *entropy numbers of an operator* $T \in \mathfrak{L}(E, F)$ are defined as

$$\varepsilon(T) := \varepsilon(T(U_E)) \quad \text{where} \quad U_E = \{x \in E : \|x\|_E \leq 1\}. \tag{59}$$

Note that $\varepsilon(T) = \|T\|$, and that $\varepsilon(T)$ certainly is well defined for all $n \in \mathbb{N}$ if T is a *compact operator*, i.e., if $T(U_E)$ is precompact.[6]

The key idea in the following is to bound the entropy number of parameterized curves in $L_\infty(\ell_2^d)$ satisfying the constraint $Q[f] \leq \Lambda$ by viewing $\mathcal{F}_\Lambda$ as the image of the unit ball under an operator T. A key tool in bounding the relevant entropy number is the following factorization result.

Proposition 8 (Carl and Stephani (1990), p. 11) *Let E, F, G be Banach spaces, $R \in \mathfrak{L}(F, G)$, and $S \in \mathfrak{L}(E, F)$. Then, for $n, t \in \mathbb{N}$,*

$$\begin{aligned} \varepsilon nt(RS) &\leq \varepsilon(R)\varepsilon(S) & (60)\\ \varepsilon(RS) &\leq \varepsilon(R)\|S\| & (61)\\ \varepsilon(RS) &\leq \varepsilon(S)\|R\|. & (62) \end{aligned}$$

Since one is dealing with vector-valued functions $\mathcal{F}_\Lambda$, it handy to view $f(\cdot)$ as generated by a linear $d = \dim \mathcal{X}$ dimensional operator, i.e.,

$$f(z) = w\Phi(z) := (\langle w_1, \Phi(z)\rangle, \ldots, \langle w_d, \Phi(z)\rangle). \tag{63}$$

Here the inner product $\langle\cdot,\cdot\rangle$ is given by the regularization operator P as

$$\langle f, g\rangle := \langle Pf, Pg\rangle_{L_2} = \int_{\mathcal{Z}} \overline{(Pg)(x)}(Pf)(x)\mathrm{d}x \tag{64}$$

where the latter was described in section 4.

A.2 Using the Shape of the Feature Space

It is convenient to view (64) also the other way around. For any kernel k corresponding to a positive integral operator

$$(T_k f)(x) := \int_{\mathcal{Z}} f(x')k(x', x)\mathrm{d}x', \tag{65}$$

[6] Precompactness of a set X means that for any $\varepsilon > 0$ there exists a finite number of open balls of radius ε that cover X.

e.g., for any *Mercer kernel* (one satisfying Mercer's condition (Mercer, 1909)), one can write k as

$$k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x'). \tag{66}$$

Here (λ_i, ϕ_i) is the eigensystem of the integral operator T_k. Hence the map into feature space may be written as

$$\Phi(x) = \left(\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), \ldots\right). \tag{67}$$

This property has been used to develop and understand learning algorithms for RBF networks (Aizerman et al., 1964), support vector machines (Boser et al., 1992), and kernel PCA (Schölkopf et al., 1998). In the current context we will use the geometrical viewpoint to provide bounds on the entropy numbers of the classes of functions $\mathcal{F}_\Lambda$ generated by kernels.

One can prove (Williamson et al., 1998) by using Mercer's theorem that Φ corresponding to k (and matching regularization operators P) maps x into a box $\mathcal{B}$ in a Hilbert space with side-lengths $C_k\sqrt{\lambda_i}$. Here C_k is a constant depending on the kernel and the eigenvalues λ_i are those of the integral operator T_k.

Finally we need to introduce the mixed spaces $\ell_p^d(\ell_2)$ to describe the geometric properties of the setting:

$$\ell_p^d(\ell_2) := \left\{ x = (x_1, \ldots, x_d) \colon x_i \in \ell_2, \text{ and } \|x\|_{\ell_p^d(\ell_2)} := \left(\textstyle\sum_{i=1}^d \|x_i\|_{\ell_2}^p\right)^{1/p} < \infty \right\}. \tag{68}$$

Quadratic Regularizers Here $\mathcal{F}_\Lambda$ is the class of all functions with $\frac{1}{2}\|Pf\|^2 \leq \Lambda$. Consequently $\Phi(x) \in \mathcal{B}$ and moreover $\frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_{i=1}^d \|w_i\|^2 \leq \Lambda$. Thus $w := (w_1, \ldots, w_d) \in \sqrt{2\Lambda}U_{\ell_2^d(\ell_2)}$.

Linear Regularizers Here

$$\mathcal{F}_\Lambda = \left\{ f = \sum_{i=1}^M \alpha_i K(z_i, \cdot) \colon \sum_{i=1}^M \|\alpha_i\|_1 \leq \Lambda \right\}. \tag{69}$$

Hence we have $\Phi(x) \in \mathcal{B}$ and moreover $w_i \in \Lambda_i \mathcal{B}$ with $\sum_i \Lambda_i \leq \Lambda$. This is the case since $w_i = \sum_{j=1}^M \alpha_{ij}\Phi(z_i)$, hence $w_i \in \left(\sum_{j=1}^M |\alpha_{ij}|\right)\mathcal{B}$ which satisfies the above restrictions by construction of $\mathcal{F}_\Lambda$.

Our strategy will be (as in (Williamson et al., 1998, Smola et al., 2000)) to find operators A mapping $\mathcal{B}$ (or their d-times replication) into balls of some radius R_A and use the Cauchy-Schwartz inequality to obtain overall covering numbers for the class of functions under consideration.

Proposition 9 (Williamson, Smola, and Schölkopf (1998)) *Let $\Phi(\cdot)$ be the map onto the eigensystem introduced by a Mercer kernel k with eigenvalues $(\lambda_i)_i$. Denote by C_k a constant depending on the kernel given by $C_k := \sup_i \|\psi_i\|_{L_\infty}$, where ψ_i is the eigenfunction corresponding to λ_i and normalised such that $\|\psi_i\|_{L_2} = 1$. Let A be the diagonal map*

$$\begin{array}{rcl} A : \mathbb{R}^{\mathbb{N}} & \rightarrow & \mathbb{R}^{\mathbb{N}} \\ A : (x_j)_j & \mapsto & A(x_j)_j = R_A(a_j x_j)_j \text{ where } a_j \in \mathbb{R} \end{array} \tag{70}$$

and $R_A := C_k \|(\sqrt{\lambda_j} a_j)_j\|_{\ell_2}$. Then by construction A^{-1} maps $\Phi(\mathcal{X})$ into the unit ball, centered at the origin if and only if $(\sqrt{\lambda_j} a_j)_j \in \ell_2$.

The evaluation operator S plays a crucial role in dealing with entire classes of functions (instead of just a single $f(\cdot)$). It is defined as

$$\begin{array}{lcl} S_{\Phi(\mathcal{Z})} : \ell_p^d(\ell_2) & \to & L_\infty(\ell_2^d) \\ S_{\Phi(\mathcal{Z})} : w & \mapsto & (\langle w_1, \Phi(\mathcal{Z})\rangle, \ldots, \langle w_d, \Phi(\mathcal{Z})\rangle) \,. \end{array} \tag{71}$$

Furthermore we will need a bound on the operator norm of $\|S_{A^{-1}\Phi(\mathcal{Z})}\|$ in order to provide bounds on the entropy numbers of a concatenated operator constructed from it. By application of the Cauchy-Schwartz inequality we obtain

$$\begin{array}{lcl} \|S_{A^{-1}\Phi(\mathcal{Z})}\| & = & \sup_{z \in \mathcal{Z}} \|(\langle w_1, A^{-1}\Phi(z)\rangle, \ldots, \langle w_d, A^{-1}\Phi(z)\rangle)\|_{\ell_2^d} \\ & \leq & \left(\sup_{z \in \mathcal{Z}} \|A^{-1}\Phi(z)\|\right) \left(\sum_{i=1}^d \|w_i\|^2\right)^{1/2} \\ & \leq & \max\left(1, d^{\frac{1}{2} - \frac{1}{p}}\right) \end{array} \tag{72}$$

since we assumed $w = (w_1, \ldots, w_d)$ to be constrained to the ball $U_{\ell_p^d(\ell_2)}$, (this means that $(\|w_1\|_{\ell_2}, \ldots, \|w_d\|_{\ell_2}) \in U_{\ell_p^d})$. Before we proceed to the actual bounds for different classes $\mathcal{F}_\Lambda$, we define a scaling operator A_d for the multi-output case as the d times tensor product of A, i.e.

$$A_d : \ell_p^d(\ell_2) \to \ell_p^d(\ell_2) \text{ and } A_d := \underbrace{A \times A \times \cdots \times A}_{d\text{-times}}. \tag{73}$$

A.3 Quadratic Regularizers

The final step, the computation of $\varepsilon(\mathcal{F}_\Lambda)$, is achieved by computing the entropy numbers of an operator mapping $U_{\ell_p^d(\ell_2)}$ or similarly restricted sets into $L_\infty(\ell_2^d)$.

Proposition 10 (Bounds for Quadratic Regularizers) *Let k be a Mercer kernel, be Φ the corresponding map into feature space, and let $T := S_{\Phi(\mathcal{Z})}\Lambda$ where $S_{\Phi(\mathcal{Z})}$ is given by (71) and $\Lambda \in \mathbb{R}^+$. Let A be defined by (70) and A_d by (73). Then the entropy numbers of T satisfy*

$$\varepsilon(T) \leq \Lambda \varepsilon(A_d). \tag{74}$$

Proof The proof relies on the fact that the following diagram commutes.

$$\begin{tikzcd} U_{\ell_2^d(\ell_2)} \subset \ell_2^d(\ell_2) \arrow[r, "T"] \arrow[d, "\Lambda"'] & L_\infty(\ell_2^d) \\ \Lambda U_{\ell_2^d(\ell_2)} \subset \ell_2^d(\ell_2) \arrow[r, "A_d"'] \arrow[ur, "S_{\Phi(\mathcal{Z})}"] & \ell_2^d(\ell_2) \arrow[u, "S_{A^{-1}\Phi(\mathcal{Z})}"'] \end{tikzcd} \tag{75}$$

That this is so can be see as follows:

$$\begin{aligned} \varepsilon(T) &= \varepsilon\left(\Lambda S_{\Phi(\mathcal{Z})} U_{\ell_2^d(\ell_2)}\right) && (76)\\ &= \varepsilon\left(\Lambda S_{A^{-1}\Phi(\mathcal{Z})} A_d U_{\ell_2^d(\ell_2)}\right) && (77)\\ &\leq \Lambda \left\| S_{(A^{-1}\Phi(\mathcal{Z}))} \right\| \varepsilon(A_d) && (78)\\ &\leq \Lambda \varepsilon(A_d) && (79) \end{aligned}$$

Here we relied on the fact that $\|A\Phi(z)\| \leq 1$, the factorization property of entropy numbers (proposition 8) and on the fact that by construction

$$(\langle w_1, \Phi(z)\rangle, \ldots, \langle w_d, \Phi(z)\rangle) = \left(\langle Aw_1, A^{-1}\Phi(z)\rangle, \ldots, \langle Aw_d, A^{-1}\Phi(z)\rangle\right), \tag{80}$$

which is just the explicit notation of A_d. ■

The price for dealing with vector-valued functions is a degeneracy in the eigenvalues of A_d - scaling factors appear d times, instead of only once in the single output situation. From a theorem for degenerate eigenvalues of scaling operators (Williamson et al., 1998) one immediately obtains the following corollary.

Corollary 11 (Entropy numbers for the vector valued case) *Let k be a Mercer kernel, let A be defined by (70) and A_d by (73). Then there exists an operator $\hat{A}_d$ such that*

$$\varepsilon(\hat{A}_d\colon \ell_2 \to \ell_2) \leq \inf_{(a_s)_s : \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} 6 C_k \sqrt{d} \left\| \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \right\|_{\ell_2} n^{-\frac{1}{j \cdot d}} (a_1 a_2 \cdots a_j)^{\frac{1}{j}}. \tag{81}$$

Note that the dimensionality of $\mathcal{Z}$ does not affect these considerations directly, however it has to be taken into account implicitly by the decay of the eigenvalues (Williamson et al., 1998) of the integral operator induced by k. The output dimensionality d, however, affects the bound in two ways - firstly due to the increased operator norm (the $\sqrt{d}$ term) for the scaling operator A_d, and secondly due to the slower decay properties (each scaling factor a_i appears d times).

The same techniques that led to explicit bounds on entropy numbers of Williamson et al. (1998) can also be applied here. As this is rather technical, we will only briefly sketch a similar result for the case of principal manifolds.

Proposition 12 (Exponential-Polynomial decay) *Suppose k is a Mercer kernel with $\lambda_j = O(e^{-\alpha j^p})$ for some $\alpha, p > 0$. Then*

$$|\log \varepsilon_n(A_d\colon \ell_2 \to \ell_2)| = O(\log^{\frac{p}{p+1}} n). \tag{82}$$

Proof We use a series $(a_j)_j = e^{-\tau/2j^p}$. Moreover there exists some $\beta \in \mathbb{R}^+$ such that $\lambda_j \leq \beta^2 e^{-\alpha j^p}$. Now we may bound

$$\begin{aligned} \sqrt{d} \left\| \left(\tfrac{\sqrt{\lambda_j}}{a_j}\right)_j \right\|_{\ell_2} &= \sqrt{d}\beta \left(\textstyle\sum_{j=0}^{\infty} e^{(\tau-\alpha)j^p}\right)^{\frac{1}{2}} \leq \sqrt{d}\beta \sqrt{1 + \textstyle\int_0^\infty e^{(\tau-\alpha)t^p} \mathrm{d}t}\\ &= \sqrt{d}\beta \sqrt{1 + \tfrac{\Gamma(1/p)}{p(\alpha-\tau)^{1/p}}} \end{aligned} \tag{83}$$

and $(a_1 a_2 \dots a_j)^{\frac{1}{j}} = \exp\left(-\frac{1}{2j}\tau \sum_{s=1}^{j} s^p\right) \leq e^{-\tau\phi j^p}$ for some positive number ϕ. For the purpose of finding an upper bound, $\sup_{j \in \mathbb{N}}$ can be replaced by $\sup_{j \in [1,\infty]}$. One computes $\sup_{j \in [1,\infty]} n^{-\frac{1}{dj}} e^{-\tau\phi j^p}$ which is obtained for some $j = \phi' \log^{\frac{1}{p+1}} n$ and some $\phi' > 0$. Resubstitution yields the claimed rate of convergence for any $\tau \in (0, \alpha)$ which proves the theorem.[7] ■

Possible kernels for which proposition 12 applies are Gaussian radial basis functions, i.e., $k(x, x') = \exp(-\|x - x'\|^2)$ $(p = 2)$ and the "Damped Harmonic Oscillator", i.e., $k(x, x') = \frac{1}{1+\|x-x'\|^2}$ with $p = 1$. For more details on this issue see (Williamson et al., 1998). Finally one has to invert (82) to obtain a bound on $\mathcal{N}(\varepsilon, \mathcal{F}_\Lambda, L_\infty(\ell_2^d))$. We have:

$$\log \mathcal{N}\left(\varepsilon, \mathcal{F}_\Lambda, L_\infty(\ell_2^d)\right) = O(\log^{\frac{p+1}{p}}(\frac{1}{\varepsilon})). \tag{84}$$

A similar result may be obtained for the case of polynomial decay in the eigenvalues of the Mercer kernel. Following (Williamson et al., 1998) one gets:

Proposition 13 (Polynomial Decay (Williamson et al., 1998)) *Let k be a Mercer kernel with eigenvalues $\lambda_j = O(j^{-(\alpha+1)})$ for some $\alpha > 0$. Then for any $\delta \in (0, \alpha/2)$ we have*

$$\begin{aligned} \varepsilon_n(A\colon \ell_2 \to \ell_2) &= O\left(\log^{-\frac{\alpha}{2}+\delta} n\right). \\ \varepsilon_n(A\colon \ell_2 \to \ell_2) &= \Omega\left(\log^{-\frac{\alpha}{2}} n\right). \end{aligned}$$

thus

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_\Lambda, L_\infty(\ell_2^d)) = O(\varepsilon^{-2/\alpha+\delta}). \tag{85}$$

A.4 Linear Regularizers

Analogous application of the techniques described in the previous section will lead to bounds on the entropy numbers of $\mathcal{F}_\Lambda$ for linear regularizers. The additional complication in the setting arises from the fact that now also the separate "components" $w_1, \dots, w_d$ of w are contained inside a scaled version of the box $\mathcal{B}$ rather than a scaled version of the unit ball.

However, by construction of A and A_d one obtains that $A_d w \in \Lambda U_{\ell_1^d(\ell_2)}$, since $A\mathcal{B} \in U_{\ell_2}$. Hence we have

$$\varepsilon(\mathcal{F}_\Lambda) \leq \Lambda \varepsilon(S_{\Phi(\mathcal{Z})} A_d U_{\ell_1^d(\ell_2)}). \tag{86}$$

Here we assumed $S_{\Phi(\mathcal{Z})}$ to have $U_{\ell_1^d(\ell_2)}$ as its domain of validity instead of $U_{\ell_2^d(\ell_2)}$ as in the previous section. All techniques, in particular the factorization of $S_{\Phi(\mathcal{Z})}$ carries over and we obtain

$$\varepsilon(\mathcal{F}_\Lambda) \leq \Lambda \varepsilon((A_d)^2). \tag{87}$$

Hence $\mathcal{F}_\Lambda$ for the linear regularizers behaves as if the rate of decay in the eigenvalues of k was twice as fast as in the quadratic regularizer setting.

[7] See (Williamson et al., 1998) for how explicit bounds can be obtained instead of just asymptotic rates.

A.5 Linear Regularizers for Kernels with Lipschitz Properties

In the above we have only made use of information concerning the eigenvalues of the kernel used. Interestingly we can do better if in addition the kernels satisfy a Lipschitz property, i.e.,

$$|k(z,z') - k(z,z'')| \leq c_k \|z' - z''\| \text{ for all } z, z', z'' \in \mathcal{Z}. \tag{88}$$

In the latter case, also the resulting function $f \in \mathcal{F}_\Lambda$ satisfies a Lipschitz property with Lipschitz constant Λc_k. To see this, note that by construction $\sum_{i,j} |\alpha_{ij}| \leq \Lambda$.

Now suppose that $z_1, \ldots, z_a$ ($a \in \mathbb{N}$) form an ε-cover of $\mathcal{Z}$ with respect to the standard metric. Suppose, moreover, that $f_1, \ldots, f_n$ ($n \in \mathbb{N}$) are n elements of $\mathcal{F}_\Lambda$, whose restrictions to $\{z_1, \ldots, z_a\}$ form an ε'-cover of $\mathcal{F}_\Lambda|_{\{z_1,\ldots,z_a\}}$ in the ℓ_∞^a metric. Then, due to the Lipschitz property, we have an $\varepsilon' + \Lambda c_k \varepsilon$ cover, consisting of $f_1, \ldots, f_n$, in terms of the $L_\infty(\ell_2^d)$ metric on $\mathcal{F}_\Lambda$. In terms of entropy numbers, we thus arrive at

$$\varepsilon_n(\mathcal{F}_\Lambda, L_\infty(\ell_2^d)) \leq \Lambda c_k \varepsilon_a(\mathcal{Z}) + \varepsilon_n\left(\mathcal{F}_\Lambda, \ell_\infty^a\right). \tag{89}$$

This result holds for arbitrary $a \in \mathbb{N}$. The rest of the proof strategy is as follows: bound the entropy numbers of $\mathcal{Z}$ and $\mathcal{F}_\Lambda$ with respect to the corresponding metrics. The first part is straightforward via volume considerations, $\mathcal{Z}$ being a bounded subset of a finite-dimensional space. The second part takes into account the entropy number properties of the kernel; it is technically more demanding, but can be done in analogy to (Williamson et al., 1998). Finally, one can exploit the freedom to choose a by optimizing over it (potentially by numerical means) to obtain the tightest possible form of the bound.

Appendix B. Proofs

B.1 Proof of Proposition 4

Proof By the compactness assumption, $f_{\text{emp}}^{*,m}$, and f^* as defined exist. We proceed similarly to the proof of proposition 3, however use $\mathcal{N}(\varepsilon, \mathcal{F}_\Lambda^c, d)$ and $\frac{\eta}{2}$ to bound $R[f_{\text{emp}}^*]$

$$\begin{aligned}
R[f_{\text{emp}}^{*,m}] - R[f^*] &= R[f_{\text{emp}}^{*,m}] - R_{\text{emp}}[f_{\text{emp}}^{*,m}] + R_{\text{emp}}[f_{\text{emp}}^{*,m}] - R[f^*] && (90)\\
&\leq \varepsilon + R[f_i] - R_{\text{emp}}[f_i] + R_{\text{emp}}[f_{\text{emp}}^{*,m}] - R[f^*] && (91)\\
&\leq \varepsilon + 2 \max_{f \in V_\varepsilon \cup \{f^*\}} |R[f] - R_{\text{emp}}[f]| && (92)
\end{aligned}$$

where V_ε is the ε-cover of $\mathcal{F}_\Lambda$ of size $\mathcal{N}(\varepsilon, \mathcal{F}, L_\infty(\ell_2^d))$, $f_i \in V_\varepsilon$ and clearly $R_{\text{emp}}[f_{\text{emp}}^{*,m}] \leq R_{\text{emp}}[f^*]$. Now apply Hoeffding's inequality, the union bound and change $\eta + \varepsilon$ into η to prove the claim. ■

B.2 Proof of Proposition 6

Proof The proof uses a clever trick from (Kégl et al., 2000), however without the difficulty of also having to bound the approximation error. Since by hypothesis $\mathcal{F}_\Lambda$ is compact, we can use Proposition 4. We have

$$\begin{aligned}
R[f^*_{\mathrm{emp}}] - R[f^*] &= \int\limits_0^\infty \Pr\left\{R[f^{*,m}_{\mathrm{emp}}] - R[f^*] > \eta\right\} \mathrm{d}\eta \\
&\leq u + \varepsilon + 2\left(\mathcal{N}\left(\varepsilon, \mathcal{F}^c_\Lambda, d\right) + 1\right) \int\limits_{u+\varepsilon}^\infty e^{-\frac{m(\eta-\varepsilon)^2}{2e_c}} \mathrm{d}\eta \\
&\leq u + \varepsilon + \frac{2e_c}{um}\left(\mathcal{N}\left(\varepsilon, \mathcal{F}^c_\Lambda, d\right) + 1\right) e^{-\frac{mu^2}{2e_c}} \\
&\leq \sqrt{\tfrac{2e_c \log\left(\mathcal{N}\left(\varepsilon, \mathcal{F}^c_\Lambda, d\right)+1\right)}{m}} + \varepsilon + \sqrt{\tfrac{2e_c}{m \log\left(\mathcal{N}\left(\varepsilon, \mathcal{F}^c_\Lambda, d\right)+1\right)}}. \qquad (93)
\end{aligned}$$

Here we used $\int_x^\infty \exp(-t^2/2)\mathrm{d}t \leq \exp(-x^2/2)/x$ in the second step. The third inequality was derived by substituting

$$u = \sqrt{\frac{2e_c}{m} \log\left(\mathcal{N}\left(\varepsilon, \mathcal{F}^c_\Lambda, d\right) + 1\right)}. \qquad (94)$$

For part **1**, set $\varepsilon = m^{-1/2}$ and we obtain

$$R[f^{*,m}_{\mathrm{emp}}] - R[f^*] = O\left(m^{-1/2} \log^{\alpha/2} m\right). \qquad (95)$$

For part **2**, (93) implies (for some constants $c, c' > 0$)

$$R[f^{*,m}_{\mathrm{emp}}] - R[f^*] \leq c\varepsilon^{-\alpha/2} m^{-1/2} + \varepsilon + c'\varepsilon^{\alpha/2} m^{-1/2}. \qquad (96)$$

The minimum is obtained for $\varepsilon = c'' m^{-1/(\alpha+2)}$ for some $c'' > 0$. Hence the overall term is of order $O(m^{-\frac{1}{\alpha+2}})$, as required. ■

References

M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

M. Anthony and P. Bartlett. *A Theory of Learning in Artificial Neural Networks.* Cambridge University Press, 1999.

P. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813, 1998.

K. Bennett. Combining support vector and mathematical programming methods for induction. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - SV Learning*, pages 307–326, Cambridge, MA, 1999. MIT Press.

C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7:108–116, 1995.

C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

P. Bradley and O. Mangasarian. Massive data discrimination via linear suppport vector machines. Mathematical Programming Technical Report 98-05, University of Wisconsin Madison, 1998.

P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. In *Advances in Neural Information Processing Systems*, volume 9, pages 368–374, Cambridge, MA, 1997. MIT Press.

B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *Siam Journal of Scientific Computing*, 20(1):33–61, 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–22, 1977.

T. T. Frieß and R. F. Harrison. Linear programming support vector machiens for pattern classification and regression estimation and the set reduction algorithm. TR RR-706, University of Sheffield, Sheffield, UK, 1998.

A. Gersho and R. M. Gray. *Vector quantization and signal compression.* Kluwer Academic Publishers, Boston, 1991.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

M. Hamermesh. *Group theory and its applications to physical problems.* Addison Wesley, Reading, MA, 2 edition, 1962. Reprint by Dover, New York, NY.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

P. J. Huber. *Robust Statistics.* John Wiley and Sons, New York, 1981.

IBM Corporation. IBM optimization subroutine library guide and reference. *IBM Systems Journal*, 31, 1992. SC23-0519.

N. Kambhatla and T. K. Leen. Fast non-linear dimension reduction. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6. Proceedings of the 1993 Conference*, pages 152–159, San Francisco, CA, 1994. Morgan Kaufmann.

N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, 1997.

B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 281–297, 2000.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28: 129–137, 1982.

O. L. Mangasarian. *Nonlinear Programming.* McGraw-Hill, New York, NY, 1969.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, A 209: 415–446, 1909.

V. A. Morozov. *Methods for Solving Incorrectly Posed Problems.* Springer Verlag, 1984.

B. A. Murtagh and M. A. Saunders. MINOS 5.1 user's guide. Technical Report SOL 83–20R, Stanford University, CA, USA, 1983. Revised 1987.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.).* Cambridge University Press, Cambridge, 1992. ISBN 0-521-43108-5.

S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11(2), 1999.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

A. J. Smola. *Learning with Kernels.* PhD thesis, Technische Universität Berlin, 1998. GMD Research Series No. 25.

A. J. Smola, A. Elisseeff, B. Schölkopf, and R. C. Williamson. Entropy numbers for convex combinations and MLPs. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 369–387, Cambridge, MA, 2000. MIT Press.

A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill–Posed Problems.* Winston, Washington, DC, 1977.

R. J. Vanderbei. LOQO user's manual — version 3.10. Technical Report SOR-97-08, Princeton University, Statistics and Operations Research, 1997. Code available at http://www.princeton.edu/˜rvdb/.

V. Vapnik. *Statistical Learning Theory.* Wiley, NY, 1998.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

G. Wahba. Smoothing and ill-posed problems. In M. Golberg, editor, *Solutions methods for integral equations and applications*, pages 183–194. Plenum Press, New York, 1979.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, Philadelphia, 1990.

J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, and C. Watkins. Support vector density estimation. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 293–306, Cambridge, MA, 1999. MIT Press.

C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models.* Kluwer, 1998.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report 19, NeuroCOLT, http://www.neurocolt.com, 1998. Accepted for publication in IEEE Transactions on Information Theory.

A. Yuille and N. Grzywacz. The motion coherence theory. In *Proceedings of the International Conference on Computer Vision*, pages 344–354, Washington, D.C., December 1988. IEEE Computer Society Press.