Regularized Principal Manifolds*

Alex J. Smola[1], Robert C. Williamson[2],
Sebastian Mika[1], Bernhard Schölkopf[1]

[1] GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[2] Department of Engineering, Australian National University, Canberra, Australia

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 settings. Moreover, this setting is very closely related to both principal curves and the generative topographic map.
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. Experimental results demonstrate the feasibility of the approach. 2) We derive uniform convergence bounds and hence bounds on the learning rates of the algorithm. In particular, we give good bounds on the covering numbers which allows us to obtain a nearly optimal learning rate of order $O(m^{-\frac{1}{2}+\alpha})$ for certain types of regularization operators, where m is the sample size and α an arbitrary positive constant.

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 hypthesis can be compared with training data. Instead, one has to make assumptions on the data, with respect to which questions may be asked.

A possible goal would be to look for reliable feature extractors, a setting that can be shown to lead to Kernel Principal Component Analysis [8]. Another option is to look for properties that represent the data best. This means leads to a *descriptive* model of the data (and possibly also a quite crude model of the underlying probability distribution). Principal Curves [6], the Generative Topographic Mapping [2], several linear Gaussian models, or also simple vector quantizers [1] are examples thereof.

We will study this type of models in the present paper. As many problems of unsupervised learning can be formalized in a quantization functional setting,

* This work was supported in part by grants of the ARC and the DFG (Ja 379/71 and Ja 379/51). Moreover we thank Balazs Kégl and Adam Krzyżak for helpful comments and discussions.

P. Fischer and H.U. Simon (Eds.): EuroCOLT'99, LNAI 1572, pp. 214–229, 1999.

this will allow to use techniques from regularization theory. In particular this leads to a natural generalization (to higher dimensionality and different criteria of regularity) of the principal curves algorithm with a length constraint [7]. See also [4] for an overview and background on principal curves. Experimental results demonstrate the feasibility of this approach.

In the second part we use the quantization functional approach to give uniform convergence bounds. In particular we derive a bound on the covering/entropy number, using functional analytic tools with respect to the $L_\infty(\ell_2^d)$ metric. This allows us to give a bound on the rate of convergence by $O(m^{-\frac{1}{2}+\alpha})$ for arbitrary positive α where m is the number of examples seen. For specific kernels this improves on the rate in [7] which is $O(m^{-\frac{1}{3}})$. Curiously, using our approach and a regularization operator equivalent to that implicitly used in [7] results in a weaker bound of $O(m^{-\frac{1}{4}})$. We suggest a possible reason for this in the penultimate section of the paper.

2 The Quantization Error Functional

Denote by $\mathcal{X}$ a vector space and $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$ a dataset drawn iid from an 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 such f that the so–called quantization error

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

is minimized. Unfortunately, this is unsolvable, as μ is generally unknown. Hence one replaces μ by the empirical density $\mu_m(x) := \frac{1}{m} \sum_{i=1}^m \delta(x - x_i)$ and instead of (1) analyzes the empirical quantization error

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

Many problems of unsupervised learning can be cast in the form of finding a minimizer of (1) or (2). Let us consider some practical examples.

Example 1 (Sample Mean). Define $\mathcal{Z} := \{1\}$, $f : 1 \to f_1$ with $f_1 \in \mathcal{X}$, and $\mathcal{F}$ to be the set of all such functions. Then the minimum of

$$R[f] := \int_{\mathcal{X}} \|x - f_1\|^2 d\mu(x) \tag{3}$$

denotes the variance of the data and the minimizers of the quantization functionals can be determined analytically by

$$\operatorname*{argmin}_{f \in \mathcal{F}} R[f] = \int_{\mathcal{X}} x d\mu(x) \text{ and } \operatorname*{argmin}_{f \in \mathcal{F}} R_{\mathrm{emp}}[f] = \frac{1}{m} \sum_{i=1}^m x_i. \tag{4}$$

This is the (empirical) sample mean. Via the law of large numbers $R_{\text{emp}}[f]$ and its minimizer converge to $R[f]$ and the corresponding minimizer (which is already a uniform convergence statement).

Example 2 (k–Vectors Quantization). Define $\mathcal{Z} := \{1, \ldots, k\}$, $f : i \to f_i$ with $f_i \in \mathcal{X}$, and $\mathcal{F}$ to be the set of all such functions. Then

$$R[f] := \int_{\mathcal{X}} \min_{z \in \{1,\ldots,k\}} \|x - f_z\|^2 d\mu(x) \tag{5}$$

denotes the canonical distortion error of a vector quantizer. In practice one uses the k–means algorithm to find a set of vectors $\{f_1, \ldots, f_k\}$ minimizing $R_{\text{emp}}[f]$. Also here [1], one can prove convergence properties of (the minimizer) of $R_{\text{emp}}[f]$ to (the one of) $R[f]$.

Instead of discrete quantization one can also consider a mapping onto a manifold of lower dimensionality than the input space. PCA can also be viewed in this way [6]:

Example 3 (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. Then the minimizer of

$$R[f] := \int_{\mathcal{X}} \min_{z \in [0,1]} \|x - f_0 - z \cdot f_1\|^2 d\mu(x) \tag{6}$$

yields a line parallel to the direction of largest variance in $\mu(x)$ [6].

Based on the properties of the current example, Hastie & Stuetzle [6] carried this idea further by also allowing other than linear functions $f(z)$.

Example 4 (Principal Curves and Surfaces). Denote $\mathcal{Z} := [0,1]^D$ (with $D > 1$ 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). The minimizer of

$$R[f] := \int_{\mathcal{X}} \min_{z \in [0,1]^D} \|x - f(z)\|^2 d\mu(x) \tag{7}$$

is not well defined, unless $\mathcal{F}$ is a compact set. Moreover, even the minimizer of $R_{\text{emp}}[f]$ is not well defined either, in general. In fact, it is an ill posed problem in the sense of Arsenin and Tikhonow [10]. Until recently [7], no convergence properties of $R_{\text{emp}}[f]$ to $R[f]$ could be stated.

Kégl et al. [7] modified the original "principal–curves" algorithm, in order to prove bounds on $R[f]$ in terms of $R_{\text{emp}}[f]$ 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.[1] Instead of the

[1] 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 rates on the training time. For the uniform convergence part, however, the length constraint is used.

latter we now consider smoothness constraints on the estimated curve $f(x)$. This is done via a regularization operator. As well as allowing greater freedom in the choice of regularizer (which, as we show, can lead to faster convergence), the regularization operator framework can be extended to situations where $D > 1$. Thus we can provide theoretical insight into principal manifold algorithms.

3 Invariant Regularizers

As a first step we will show that the class of admissible operators can be restricted to scalar ones, provided some basic assumption about scaling behavior 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 the group $\mathcal{G}$ on $\mathcal{X}$, i.e. satisfies*

$$Q[f] \geq 0 \text{ for all } f \in \mathcal{F} \tag{8}$$

$$Q[af] = a^2 Q[f] \text{ for all scalars } a \tag{9}$$

$$Q[\rho(g)f] = Q[f] \text{ for all } \rho(g) \in \mathcal{G} \tag{10}$$

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

Proof. It follows directly from (9) 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 as it has to be positive (cf. (8)).

Finally from $\langle Pf, Pf \rangle = \langle P\rho(g)f, P\rho(g)f \rangle$ and the polarization equation it follows that $P^*P\rho(g) = \rho(g)P^*P$ has to hold for any $\rho(g) \in \mathcal{G}$. Thus, by virtue of Schur's lemma (cf. e.g. [5]) P^*P only may be a scalar operator. Without loss of generality, also P may be assumed to be scalar.

A consequence is that there exists no "vector valued" regularization operator satisfying the invariance conditions. Hence it is useless to look for other operators P in the presence of a sufficiently strong invariance.

Under the assumptions of proposition 1 both the canonical representation of the permutation group in a finite dimensional vector space $\mathcal{X}$ and the group of orthogonal transformations on $\mathcal{X}$ enforce scalar operators P. This follows immediately from the fact that these groups are unitary and irreducible on $\mathcal{X}$ by construction. Thus in the following we will only consider scalar operators P.

4 A Regularized Quantization Functional

We now propose a variant to minimizing the empirical quantization functional which leads to an algorithm that is more amenable to implementation. Moreover,

uniform convergence bounds can be obtained for the classes of smooth curves induced by this approach. For this purpose, a regularized version of the empirical quantization functional is needed.

$$R_{\mathrm{reg}}[f] := R_{\mathrm{emp}}[f] + \frac{\lambda}{2}\|Pf\|^2 = \sum_{i=1}^{m} \min_{z\in\mathcal{Z}} \|x_i - f(z)\|^2 + \frac{\lambda}{2}\|Pf\|^2. \quad (11)$$

Here P is a *scalar* regularization operator in the sense of Arsenin and Tikhonov, penalizing unsmooth functions f (see [9] for details). (See some examples in section 4.2.) For the sake of finding principal manifolds, utilizing a scalar regularization operator simply means all curves or surfaces which can be transformed into each other by rotations should be penalized equally.

Using the results of [9] 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. $\langle Pk(x_i,\cdot), Pk(x_j,\cdot)\rangle = k(x_i,x_j)$. Finally assume $P^*Pf_0 = 0$, i.e. constant functions are not regularized. For an expansion like

$$f(z) = f_0 + \sum_{i=1}^{M} \alpha_i k(z_i, z) \text{ with } z_i \in \mathcal{Z},\ \alpha_i \in \mathcal{X},\ \text{and } k : \mathcal{Z}^2 \to \mathbb{R} \quad (12)$$

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

$$\|Pf\|^2 = \sum_{i,j=1}^{M} \langle \alpha_i, \alpha_j\rangle k(z_i, z_j). \quad (13)$$

What remains is to find an algorithm that minimizes R_{reg}. This is achieved by coordinate descent. In the following we will assume the data to be centered and therefore drop the term f_0. This greatly simplifies the notation.

4.1 An Algorithm for minimizing $R_{\mathrm{reg}}[f]$

Minimizing the regularized quantization functional for a given kernel expansion is equivalent to solving

$$\min_{\substack{\{\alpha_1,\ldots,\alpha_M\}\subset\mathcal{X}\\ \{\zeta_1,\ldots,\zeta_m\}\subset\mathcal{Z}}} \left[\sum_{i=1}^{m} \left\| x_i - \sum_{j=1}^{M} \alpha_j k(\zeta_i, z_j) \right\|^2 + \frac{\lambda}{2} \sum_{i,j=1}^{M} \langle \alpha_i, \alpha_j\rangle k(z_i, z_j) \right]. \quad (14)$$

This is achieved in an iterative fashion analogously to how the EM algorithm operates. One iterates over minimizing (14) with respect to $\{\zeta_1, \ldots, \zeta_m\}$, equivalent to the projection step, and $\{\alpha_1, \ldots, \alpha_M\}$, which corresponds to the expectation step. This is repeated until convergence, in practice, until the regularized quantization functional does not decrease significantly any further. One obtains: Projection: For each $i \in \{1, \ldots, m\}$ choose $\zeta_i := \operatorname{argmin}_{\zeta\in\mathcal{Z}} \|f(\zeta) - x_i\|^2$.

Clearly, for fixed α_i, the so chosen ζ_i minimizes the term in (14), which in turn is equal to $R_{\mathrm{reg}}[f]$ for given α_i and X. Adaptation: Now the parameters ζ_i are fixed and α_i is adapted such that $R_{\mathrm{reg}}[f]$ decreases further. For fixed ζ_i differentiation of (14) with respect to α_i yields

$$\left(\frac{\lambda}{2}K_z + K_\zeta^\top K_\zeta\right)\alpha = K_\zeta^\top X \tag{15}$$

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 $m \times M$. Moreover, with slight abuse of notation, α, and X denote the *matrix* of all parameters, and samples, respectively. The term in (14) keeps on decreasing until the algorithm converges to a (local) minimum. What remains is to find good starting values. Initialization If not dealing, as assumed, with centered data, set f_0 to the sample mean, i.e. $f_0 = \frac{1}{m}\sum_{i=1}^m x_i$. Moreover, choose the coefficients α_i such that 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 as follows, analogously to the initialization in the generative topographic map [2, eq. (2.20)].

$$\min_{\{\alpha_1,\ldots,\alpha_M\}\subset\mathcal{X}} \left[\sum_{i=1}^M \left\| V(z_i - z_0) - \sum_{j=1}^M \alpha_j k(z_i, z_j)\right\|^2 + \frac{\lambda}{2}\sum_{i,j=1}^M \langle\alpha_i, \alpha_j\rangle k(z_i, z_j)\right]. \tag{16}$$

Thus α is given by the solution of $\left(\frac{\lambda}{2}\mathbf{1} + K_z\right)\alpha = V(Z - Z_0)$ where Z denoted the matrix of z_i, z_0 the mean of z_i, and Z_0 the matrix of z_0 correspondingly.

The derivation of this algorithm was quite ad hoc, however, there are similar precursors in the literature. An example are principal curves with a length constraint. We will show below that for a particular choice of a regularizer, minimizing (11) is equivalent to the latter.

4.2 Examples of Regularizers

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 $\int_{[0,1]} \|\partial_z f(z)\| dz$ 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 interval $[0, 1]$. Thus $\|Pf\|^2$ equals the squared length L^2 of the curve at the optimal solution.

One can show that 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 [7].[2]

[2] The reasoning is slightly incorrect — f *cannot* be completely reparameterized to constant speed, as it is an expansion in terms of a *finite* number of nodes. However the basic properties still hold, provided the number of kernels is sufficiently high.

In the experiments we chose a Gaussian RBF kernel $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$. This corresponds to a regularizer penalizing all orders of derivatives simultaneously. In particular [14] show that this kernel corresponds to the pseudodifferential operator defined by

$$\|Pf\|^2 = \int dx \sum_{n=0}^{\infty} \frac{\sigma^{2n}}{n!2^n} (O^n f(x))^2 \tag{17}$$

with $O^{2n} = \Delta^n$ and $O^{2n+1} = \nabla\Delta^n$, 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 very slowly. For more details on regularization operators see e.g. [9].

4.3 The Connection to the GTM

Just considering the basic algorithm of the GTM (without the Bayesian framework), one can observe that it minimizes a rather similar quantity to $R_{\text{reg}}[f]$. It differs in its choice of $\mathcal{Z}$, which is chosen to be a grid, identical with the points z_i in our setting, and the different regularizer (called Gaussian prior in that case) which is of ℓ_2 type. In other words instead of using $\|Pf\|^2$ Bishop et al. [2] choose $\sum_i \|\alpha_i\|^2$ as a regularizer. Finally in the GTM several ζ_i may take on "responsibility" for having generated a data-point x_i (this follows naturally from the generative model setting in the latter case).

Note that unlike in the GTM (cf. [2, sec. 2.3]) the number of nodes (for the kernel expansion) is not a critical parameter. This is due to the fact that there is a *coupling* between the single centers of the basis functions $k(z_i, z_j)$ via the regularization operator. If needed, one could also see the proposed algorithm in a Gaussian Process context (see [12]) — the data X then should be interpreted as created by a homogeneous process mapping from $\mathcal{Z}$ to $\mathcal{X}$. Finally the use of periodical kernels (cf. [9]) allows one to model circular structures in $\mathcal{X}$.

5 Experiments

In order to show that the basic idea of the proposed algorithm is sound, we ran several toy experiments (cf. figure 1). In all cases Gaussian rbf kernels, as discussed in section 4.2, were used. 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. The number and width of the basis functions had of course an effect on the solution, too. But their influence on the basic characteristics is quite small.

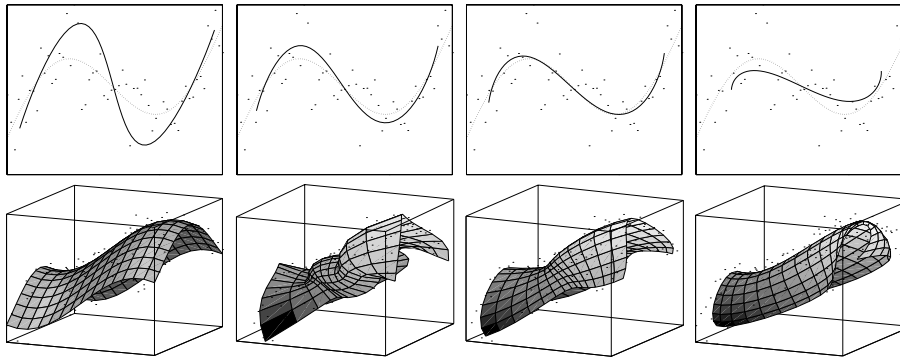


Fig. 1. Upper 4 images. We generated a dataset (small dots) by adding noise to a distribution indicated by the dotted line. The resulting manifold generated by our approach is given by the solid line (over a parameter range of $\mathcal{Z} = [-1, 1]$). From left to right we used different values for 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 constant (1 and 36).

Finally, 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 finds a (local) minimum of $R_{\text{reg}}[f]$.

6 Uniform Convergence Bounds

We now proceed to an analysis of the rate of convergence of the above algorithm. To avoid several technicalities (like boundedness of some moments of the distribution $\mu(x)$ [11]) we will assume that there exists a ball of radius r such that $\Pr\{\|x\| \leq r\} = 1$ for all x. Kégl et al. [7] showed that under these assumptions also the prinicipal manifold f is contained in the ball U_r of radius r, hence the quantization error will be no larger than $(2r)^2$ for all x. In order to derive uniform convergence bounds let us introduce the $L_\infty(\ell_2^d)$ norm on $\mathcal{F}$ (assumed continuous)

$$\|f\|_{L_\infty(\ell_2^d)} := \sup_{z \in \mathcal{Z}} \|f(z)\|_{\ell_2^d} \tag{18}$$

where the $\|\cdot\|_{\ell_2^d}$ denotes the Euclidean norm in d dimensions. The metric is induced by the norm in the usual fashion. Given a metric ρ and a set $\mathcal{F}$, the

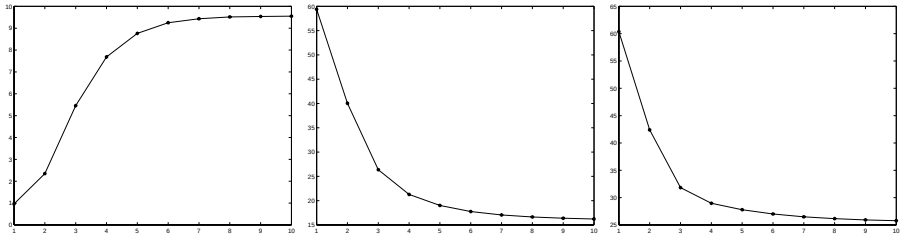


Fig. 2. Left: regularization term, middle: empirical quantization error, right: regularized quantization error vs. number of iterations.

ϵ covering number of $\mathcal{F}$, denoted $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ (also $\mathcal{N}_\epsilon$ when the dependency is obvious), is the smallest number of ρ-balls of radius the union of which contains $\mathcal{F}$.

The next two results are similar in their style to the bounds obtained in [7], however slightly streamlined, as they are independent of some technical conditions on $\mathcal{F}$ as needed in [7].

Proposition 2 ($L_\infty(\ell_2^d)$ 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 i.i.d. from μ, then for all $\eta > 0, \epsilon \in (0, \eta/2)$

$$\Pr\left\{\sup_{f\in\mathcal{F}} \left|R^m_{\text{emp}}[f] - R[f]\right| > \eta\right\} \leq 2\mathcal{N}\left(\tfrac{\epsilon}{8r}, \mathcal{F}, L_\infty(\ell_2^d)\right) e^{-m(\eta-\epsilon)^2/(2r^2)}.$$

Proof. By definition of $R^m_{\text{emp}}[f] = \sum_{i=1}^m \min_z \|f(z) - x_i\|^2$ the empirical quantization functional is a sum of m iid random variables which are each bounded by $4r^2$ due to the fact that x is contained in a ball of radius r. Hence we may apply Hoeffding's inequality to obtain

$$\Pr\left\{\left|R^m_{\text{emp}}[f] - R[f]\right| \geq \eta\right\} \leq 2e^{-m\eta^2/(2r^2)}. \tag{19}$$

By the Lipschitz property of the ℓ_2^d norm (the 'target' values are bounded by r), a $\frac{\varepsilon}{8r}$ cover of $\mathcal{F}$ is an $\frac{\varepsilon}{2}$ cover of the loss function induced class: For every $f \in \mathcal{F}$ there exists some $f_i \in \mathcal{N}_{\varepsilon/8r}$ such that $\|f_i - f\|^2_{L_\infty^{2m}(\ell_2^d)} \leq \frac{\varepsilon}{2}$. Hence also $|R^m_{\text{emp}}[f] - R^m_{\text{emp}}[f_i]| \leq \frac{\varepsilon}{2}$ and $|R[f] - R[f_i]| \leq \frac{\varepsilon}{2}$. Consequently

$$\Pr\left\{\left|R^m_{\text{emp}}[f] - R[f]\right| \geq \eta\right\} \leq \Pr\left\{\left|R^m_{\text{emp}}[f_i] - R[f_i]\right| \geq \eta - \varepsilon\right\} \tag{20}$$

Substituting (20) into (19) and taking the union bound over $\mathcal{N}_{\varepsilon/8r}$ gives the desired result.

This result is useful to assess the quality of an *empirically* found 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 3 (Rates of Convergence for Optimal Estimates). *With the definitions of Proposition 2 and the definition of f^*_{emp} and f^* one has*

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|R[f^*_{\text{emp}}]-R[f^*]\right|>\eta\right\}\leq 2\left(\mathcal{N}\left(\tfrac{\epsilon}{4r},\mathcal{F},L_\infty(\ell_2^d)\right)+1\right)e^{-\frac{m(\eta-\epsilon)^2}{8r^2}}.$$

Proof. The proof is similar to the one of proposition 2, however uses $\mathcal{N}_{\varepsilon/4r}$ and $\eta/2$ to bound $R[f^*_{\text{emp}}]$

$$\begin{aligned}R[f^*_{\text{emp}}]-R[f^*] &= R[f^*_{\text{emp}}]-R_{\text{emp}}[f^*_{\text{emp}}]+R_{\text{emp}}[f^*_{\text{emp}}]-R[f^*] &(21)\\ &\leq \varepsilon+R[f_i]-R_{\text{emp}}[f_i]+R_{\text{emp}}[f^*_{\text{emp}}]-R[f^*] &(22)\\ &\leq \varepsilon+2\max_{f\in N_\varepsilon\cup\{f^*\}}|R[f]-R_{\text{emp}}[f]| &(23)\end{aligned}$$

where $f_i\in\mathcal{N}_\varepsilon$ and clearly $R_{\text{emp}}[f^*_{\text{emp}}]\leq R_{\text{emp}}[f^*]$. Now apply Hoeffding's inequality, the union bound and change $\eta+\varepsilon$ into η to prove the claim.

After that we provided a number of uniform convergence bounds it is now necessary to bound $\mathcal{N}$ in a suitable way.

7 Covering and Entropy Numbers

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. [7] essentially show that

$$\log\mathcal{N}(\varepsilon,\mathcal{F})=O(\tfrac{1}{\varepsilon}) \tag{24}$$

under the following assumptions: They consider polygonal curves $f(\cdot)$ of length L in a sphere U_r of radius r in $\mathcal{X}$. 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 [13] one can obtain more general results, which then, in turn, can replace (24) to obtain better bounds on the expected quantization error by using the properties of the regularization operator.

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

$$\epsilon_n(M):=\inf\{\epsilon:\mathcal{N}(\varepsilon,M,\rho)\leq n\}$$

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

$$\epsilon_n(T):=\epsilon_n(T(U_E)). \tag{25}$$

Note that $\epsilon_1(T) = \|T\|$, and that $\epsilon_n(T)$ certainly is well defined for all $n \in \mathbb{N}$ if T is a *compact operator*, i.e. if $T(U_E)$ is compact.

What will be done in the following is to bound the entropy number of parametrized curves in $L_\infty(\ell_2^d)$ satisfying the constraint $\|Pf(\cdot)\|^2 \leq \Lambda$ by viewing

$$\mathcal{F}_\Lambda := \{f: \mathcal{Z} \ni z \mapsto f(z) \in \mathbb{R}^d : f \text{ is continuous, } \|Pf\| \leq \Lambda\}$$

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.

Lemma 1 (Carl and Stephani [3, p. 11]). *Let E, F, G be Banach spaces, $R \in \mathfrak{L}(F, G)$, and $S \in \mathfrak{L}(E, F)$. Then, for $n, t \in \mathbb{N}$,*

$$\epsilon_{nt}(RS) \leq \epsilon_n(R)\epsilon_t(S),\ \epsilon_n(RS) \leq \epsilon_n(R)\|S\|,\ \epsilon_n(RS) \leq \epsilon_n(S)\|R\|. \tag{26}$$

As 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 in feature space, i.e. $f(z) = W\Phi(z) = (\langle w_1, \Phi(z)\rangle, \ldots, \langle w_d, \Phi(z)\rangle)$ with $\|W\|^2 := \sum_{i=1}^d \|w\|^2$. Here the inner product $\langle \cdot, \cdot \rangle$ is given by the regularization operator P as

$$\langle f, g\rangle := \langle Pf, Pg\rangle_{L_2} = \int (Pf)(x)dx \tag{27}$$

where the latter was described in section 3. In practice w is expanded in terms of kernel functions $k(x_i, \cdot)$. The latter can be shown to represent the map from $\mathcal{Z}$ into the associated Reproducing Kernel Hilbert Space (RKHS) [9] (sometimes called feature space). Hence $\Phi(x) = k(x_i, \cdot)$, where the dot product is given by (27). These techniques may be used to give uniform convergence bounds, which are stated in terms of the eigenvalues λ_i of the RKHS.

Proposition 4 (Williamson, Smola, and Schölkopf [13]). *Let $\Phi(\cdot)$ be the map onto the eigensystem introduced by a Mercer kernel k with eigenvalues λ_i, C_k a constant of the kernel, and A be the diagonal map*

$$A : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}, \qquad A : (x_j)_j \mapsto A(x_j)_j = (a_j x_j)_j. \tag{28}$$

Then A^{-1} maps $\Phi(\mathcal{X})$ into a ball of finite radius $R_A = C_k\|(\sqrt{\lambda_j}a_j)_j\|_{\ell_2}$, 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 to deal with entire classes of functions (instead of just a single $f(\cdot)$). It is defined as

$$S_{\Phi(\mathcal{Z})} : (\ell_2)^d \to L_\infty(\ell_2^d) \text{ and } S_{\Phi(\mathcal{Z})} : W \mapsto (\langle w_1, \Phi(\mathcal{Z})\rangle, \ldots, \langle w_d, \Phi(\mathcal{Z})\rangle)\,. \tag{29}$$

By a technical argument one can see that it is possible to replace $(\ell_2)^d$ by ℓ_2 without further worry — simply reindex the coefficients by

$$\begin{aligned} &I_d : (\ell_2)^d \to \ell_2 \\ &I_d : \begin{array}{c}((w_{11}, w_{12}, \ldots), (w_{21}, w_{22}, \ldots), \ldots, (w_{d1}, w_{d2}, \ldots)) \to \\ (w_{11}, w_{21}, \ldots, w_{d1}, w_{12}, w_{22}, \ldots, w_{d2}, w_{13}, \ldots)\end{array} \end{aligned} \tag{30}$$

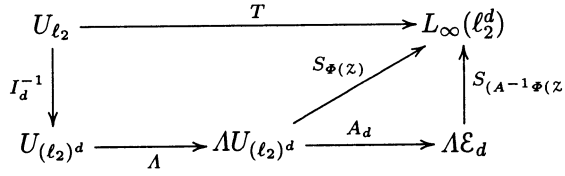
By construction $I_d U_{(\ell_2)^d} = U_{\ell_2}$ and vice versa, thus $\|I_d\| = \|I_d^{-1}\| = 1$. Before proceeding to the actual theorem one has to define a scaling operator A_d for the multi output case as the d times tensor product of A, i.e.

$$A_d : (\ell_2)^d \to (\ell_2)^d \text{ and } A_d := \underbrace{A \times A \times \ldots \times A}_{d\text{–times}} \tag{31}$$

Theorem 1 (Bounds for Principal Curves Classes). *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 (29) and $\Lambda \in \mathbb{R}^+$. Let A be defined by (28) and A_d by (31). Then the entropy numbers of T satisfy the following inequality:*

$$\epsilon_n(T) \leq \Lambda \epsilon_n(A_d) \tag{32}$$

Proof. As pointed out before one has to use a factorization argument. In particular one uses the following property.



(33)

In other words one exploits

$$\epsilon_n\left(S_{\Phi(\mathcal{Z})}\left(\Lambda U_{(\ell_2)^d}\right)\right) = \epsilon_n\left(S_{(A^{-1}\Phi(\mathcal{Z}))} A_d \Lambda I_d^{-1}\right) \tag{34}$$

$$\leq \left\|S_{(A^{-1}\Phi(\mathcal{Z}))}\right\| \epsilon_n(A_d)\Lambda \left\|I_d^{-1}\right\| \leq \Lambda \epsilon_n(A_d). \tag{35}$$

Here we have relied on Proposition 4 which says $A^{-1}\Phi(\mathcal{Z}) \subset U$ and thus by Cauchy-Schwarz, $\|S_{(A^{-1}\Phi(\mathcal{Z}))}\| \leq 1$.

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 [13] one immediately obtains the following corollary.

Corollary 1 (Entropy numbers for the vector valued case). *Let k be a Mercer kernel, let A be defined by (28) and A_d by (31). Then*

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

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

[13] of the integral operator induced by k. d appears twice — once 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 the bounds on entropy numbers in [13] can also be applied here. As this is rather technical, we only sketch a similar result for the case of principal manifolds, for $\dim \mathcal{Z} = 1$ and exponential polynomial decay of the eigenvalues λ_i of the kernel k.

Proposition 5 (Exponential–Polynomial decay). *Suppose k is a Mercer kernel with $\lambda_j = \beta^2 e^{-\alpha j^p}$ for some $\alpha, \beta, p > 0$. Then*

$$\ln \epsilon_n^{-1}(A_d \colon \ell_2 \to \ell_2) = O(\ln^{\frac{p}{p+1}} n) \tag{36}$$

Proof. We use a series $(a_j)_j = e^{-\tau/2 j^p}$. Then we bound

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

and $(a_1 a_2 \dots a_j)^{\frac{1}{j}} = e^{-\frac{1}{2j}\tau \sum_{s=1}^{j} s^p} \leq e^{-\tau \phi j^p}$ 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 $j = \phi' \ln^{\frac{1}{p+1}} n$ and some $\phi' > 0$. Resubstitution yields the claimed rate of convergence for any $\tau \in (0, \alpha)$ which proves the theorem.[3]

Possible kernels for which proposition 5 applies are Gaussian rbf, 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 [13]. Finally one has to invert (36) to obtain a bound on $\mathcal{N}(\epsilon, \mathcal{F}_\Lambda)$. We have:

$$\ln \mathcal{N}\left(\frac{\epsilon}{\Lambda}, \mathcal{F}_\Lambda, L_\infty(\ell_2^d) \right) = O(-\ln^{\frac{p+1}{p}} \epsilon) \tag{37}$$

A similar result may be obtained for the case of polynomial decay in the eigenvalues of the Mercer kernel. Following [13] one gets:

Proposition 6 (Polynomial decay). *Suppose k is a Mercer kernel with $\lambda_j = \beta^2 j^{-\alpha-1}$ for some $\alpha, \beta > 0$. Then $\epsilon_n^{-1}(A_d : \ell_2 \to \ell_2) = O(\ln^{\frac{\alpha}{2}} n)$.*

8 Rates of Convergence

It is of theoretical interest how well Principal Manifolds can be learned. Kégl et al. [7] have show a $O(m^{-1/3})$ result for principal curves ($D = 1$) with length

[3] See [13] how exact constants can be obtained instead of solely asymptotical rates.

contraint regularizer. We show that if one utilizes a more powerful regularizer (as one can do using our algorithm) one can 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 $O(m^{-1/2+\alpha})$ for exponential rates of decay (α is an arbitrary positive constant). The latter is nearly optimal, as supervised learning rates are no better than $O(m^{-1/2})$.

Proposition 7 (Learning Rates for Principal Manifolds). *For any fixed $\mathcal{F}_\Lambda$ the learning rate of principal manifolds can be lower bounded by $O(m^{-1/2+\alpha})$ where α is an arbitrary positive constant, i.e.*

$$R[f^*_{\text{emp}}] - R[f^*] \leq O(m^{-1/2+\alpha}) \text{ for } f^*_{\text{emp}}, f^* \in \mathcal{F}_\Lambda \tag{38}$$

if the eigenvalues of k decay exponentially. Moreover the learning rate can be bounded by $O(m^{-\frac{\alpha}{2(\alpha+1)}})$ in the case of polynomially decaying eigenvalues with rate $\alpha + 1$. We obtain

$$R[f^*_{\text{emp}}] - R[f^*] \leq O(m^{-\frac{\alpha}{2(\alpha+1)}}) \text{ for } f^*_{\text{emp}}, f^* \in \mathcal{F}_\Lambda \tag{39}$$

Proof. We use a clever trick from [7], however without the difficulty of also having to bound the approximation error. Proposition 3 will be useful.

$$\begin{aligned} R[f^*_{\text{emp}}] - R[f^*] &= \int_0^\infty \Pr\left\{R[f^*_{\text{emp}}] - R[f^*] > \eta\right\} d\eta \\ &\leq u + \epsilon + 2(\mathcal{N}(\epsilon/4r) + 1) \int_{u+\varepsilon}^\infty e^{-\frac{m(\eta-\epsilon)^2}{8r^2}} d\eta \\ &\leq u + \epsilon + \frac{8r^2}{um}(\mathcal{N}(\epsilon/4r) + 1) e^{-\frac{mu^2}{8r^2}} \\ &\leq \sqrt{\frac{8r^2 \ln(\mathcal{N}(\epsilon/4r)+1)}{m}} + \epsilon + \sqrt{\frac{8r^2}{m \ln(\mathcal{N}(\epsilon/4r)+1)}} \end{aligned} \tag{40}$$

Here we used $\int_x^\infty \exp(-t^2/2)dt \leq \exp(-x^2/2)/x$ in the second step. The third inequality was derived by substituting $u^2 = \frac{8r^2}{m} \log(\mathcal{N}(\varepsilon/4r) + 1)$.

Setting $\epsilon = \sqrt{1/m}$ and exploiting (37) yields

$$R[f^*_{\text{emp}}] - R[f^*] = O\left(\sqrt{\ln^{\frac{p+1}{p}} m/m}\right) + O(m^{-\frac{1}{2}}). \tag{41}$$

As $\ln^{\frac{p+1}{p}} m$ can be bounded by any $c_\alpha m^\alpha$ for suitably large c_α and $\alpha > 0$ one obtains the desired result. For polynomially decaying eigenvalues one obtains from proposition 6 that for a sufficiently large constant c $\ln \mathcal{N}\left(\epsilon/4r, \mathcal{F}, L_\infty(\ell_2^d)\right) \leq c\varepsilon^{-\frac{2}{\alpha}}$. Substituting this into (40) yields

$$R[f^*_{\text{emp}}] - R[f^*] \leq \sqrt{\frac{2^{3-\frac{4}{\alpha}} r^{2-\frac{2}{\alpha}} c}{m}} \varepsilon^{-\frac{1}{\alpha}} + 2\epsilon + O(m^{-\frac{1}{2}}). \tag{42}$$

The minimum is obtained for $\epsilon = c' m^{-\frac{\alpha}{2(\alpha+1)}}$ for some $c' > 0$. Hence $m^{-\frac{1}{2}} \epsilon^{-\frac{1}{\alpha}}$ is of order $O(m^{-\frac{\alpha}{2(\alpha+1)}})$, which proves the theorem.

Interestingly the above result is slightly weaker than the result in [7] for the case of length constraints, as the latter corresponds to the differentiation operator, thus polynomial eigenvalue decay of order 2, i.e. $\alpha = 1$ and therefore to a rate $\frac{\alpha}{2(\alpha+1)} = \frac{1}{4}$ (Kégl et al. [7] obtain $\frac{1}{3}$). It is unclear, whether this is due to a (possibly) not optimal 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 yet to be fully understood weakness should not detract from the fact that we *can* get better rates by using stronger regularizers, *and* our algorithm can utilize such regularizers.

9 Summing Up

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 to deal with 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.

Sample size dependent bounds for principal manifolds were given which depend on the underlying distribution μ in a very mild way. 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 largely rest on a connection between functional analysis and entropy numbers [13]. This fact also allowed us to give good bounds on the learning rate.

References

1. 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.
2. C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.
3. B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.
4. R. Der, U. Steinmetz, B. Balzuweit, and G. Schüürmann. Nonlinear principal component analysis. University of Leipzig, Preprint, http://www.informatik.uni-leipzig.de/der/Veroeff/npcafin.ps.gz, 1998.
5. M. Hamermesh. *Group theory and its applications to physical problems.* Addison Wesley, Reading, MA, 2 edition, 1962. Reprint by Dover, New York, NY.
6. T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

7. B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999. http://magenta.mast.queensu.ca/˜linder/psfiles/KeKrLiZe97.ps.gz.
8. B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299 – 1319, 1998.
9. A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.
10. A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill–Posed Problems*. Winston, Washington, DC, 1977.
11. V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.
12. C.K.I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. *Learning and Inference in Graphical Models*, 1998.
13. R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. NeuroCOLT NC-TR-98-019, Royal Holloway College, 1998.
14. 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.