Learning Bounds for Kernel Regression using Effective Data Dimensionality

Tong Zhang
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
tzhang@watson.ibm.com

Abstract

Kernel methods can embed finite dimensional data into infinite dimensional feature spaces. In spite of the large underlying feature dimensionality, kernel methods can achieve good generalization ability. This observation is often wrongly interpreted, and it has been used to argue that kernel learning can magically avoid the "curse-of-dimensionality" phenomenon encountered in statistical estimation problems. The purpose of this paper is to show that although using kernel representation, one can embed data into an infinite-dimensional feature space, the effective dimensionality of this embedding, which determines the learning complexity of the underlying kernel machine, is usually small. In particular, we introduce an algebraic definition of a scale-sensitive effective dimension associated with a kernel representation. Based on this quantity, we derive upper bounds on the generalization performance of some kernel regression methods. Moreover we show that the resulting convergent rates are optimal under various circumstances.

1 Introduction

Kernel methods have attracted significant attention recently since they can embed data into much larger nonlinear features spaces. Although these methods can utilize infinite dimensional features in the corresponding reproducing kernel Hilbert spaces (RKHS), the kernel representation makes the computation feasible.

An important property of kernel methods is their good generalization abilities in spite of the large underlying feature spaces. This is achieved through regularization, which restrict the representation power of kernel methods into a much smaller subset of the infinite dimensional feature space. The purpose of this paper is to show that the complexity of regularized kernel learning methods can be characterized by a scale-sensitive effective dimension of the data represented in the feature space.

The effective dimension of the data can be measured through eigenvalues in the kernel principle component analysis (PCA) in the feature space. For kernel methods, although the feature space can be very large, the eigenvalues of the principle components decrease rapidly. Therefore given an approximation scale, we can truncate the dimensions corresponding to the small eigenvalues, as long as they produce a combined contribution that is smaller than the scale. The remaining eigenvectors, corresponding to the large eigenvalues, are finite. The number of these eigenvectors give the effective data dimension at the selected approximation scale. This suggests that at any desirable approximation scale, we can obtain generalization bounds using the corresponding effective dimension. A precise algebraic definition of effective dimensionality, motivated through direct computation for the signal reconstruction problem, is introduced in Section 3.

We shall note that learning bounds for kernel methods can also be obtained through covering numbers. For example, for the least squares regression problem, bounds were obtained in Lee et al. (1998); Chucker and Smale (2002). These results were improved in Mendelson (2002), where the chaining technique was utilized to obtain tighter bounds. The resulting bounds in Mendelson (2002) are similar to those of van de Geer (2000). However, bounds obtained in Mendelson (2002) are presented in a form that is more suitable for learning problems typically formulated in the machine learning literature. Covering number estimates have also been obtained in various papers (for example, Guo et al. (2002); Williamson et al. (2001); Zhang (2002)). By combining the covering number results with learning bounds in Mendelson (2002), we can obtain generalization bounds for least squares regression.

However, it is not clear that existing bounds for covering numbers are tight. In addition, generalization bounds that are based on covering numbers can be improved in many cases. In order to obtain simplified and good convergence rates, one often imposes specific assumptions on the form of covering numbers. For example, for finite dimensional problems, we will obtain the correct $O(1/n)$ rate of convergence (n is the sample size), while results in Mendelson (2002) will give a suboptimal rate which is $\log n$ factor worse. In addition, bounds obtained using empirical process theory (such as Mendelson (2002) and van de Geer (2000)) often contain large un-specified constants.

A more recent development along this line, which has remedied some problems in the covering number approach, is through the use of concentration inequalities and localized Rademacher complexity measures by Bartlett et al. (2002). This approach simplifies more traditional analysis, but the resulting bounds still contain large constants. Also, to obtain an estimate on the localized Rademacher complexity, one often still has to rely on bounds for covering numbers. Only for very specific problems such as kernel least squares, it is possible to estimate localized Rademacher complexity more directly. In particular, the algebraic method used in Mendelson (2003) is related to the approach employed here. As shown in this paper, for kernel regression, using a similar algebraic approach, we can directly obtain generalization bounds without going through the extra-step of localized Rademacher complexity analysis. It is actually easier to do so than to estimate the localized Rademacher complexity itself (as in Mendelson (2003)).

Our approach presents a novel alternative to the complicated empirical process machineries, such as those employed in the traditional covering number or localized Rademacher complexity analysis. In fact, the analysis given here is completely elementary and self-contained (besides a Bernstein inequality for random vectors, which can also be avoided if we do not state our main result as an exponential probability inequality). In the mean time, bounds obtained in this paper are slightly tighter than those from earlier analysis.

We note that the standard empirical process machinery is more generally applicable; however it is also an indirect analysis (through covering numbers) which can often be improved. The technique used in this paper is specific to kernel regression, and hence it gives a more direct analysis and leads to a clean and tight generalization bound. Moreover, the effective dimension is defined algebraically, and thus can be explicitly calculated given the data. Therefore we can avoid the complicated problem of estimating covering numbers. Since the concept of effective dimension employed in this paper is derived through exact computation in the signal recovering problem, we can naturally expect the resulting bounds for the standard learning formulation to be quite tight as well. In fact, we will show that by applying our analysis to some well-studied problems for which the optimal minimax rates are known, we are able to obtain the optimal rates.

Our analysis also sheds useful insights into the complexity of kernel learning methods. The effective dimension depends on the decay of eigenvalues in the principle component analysis of the kernel data representation in the feature space. However, this effective dimension, which is scale-sensitive, is controlled through regularization (which determines the scale). In this sense, the role of regularization in kernel learning can be regarded as an implicit method of dimensionality reduction (or feature selection in machine learning) which selects the first few principle component directions. However, the regularization approach may have certain computational advantages since it is arguably easier to solve a regularized kernel regression problem than an kernel eigenvalue problem which is required in a PCA based dimensionality reduction scheme. The analysis in this paper suggests that the two approaches control learning complexity similarly.

The current paper improves some preliminary results in Zhang (2003a). We organize it as follows. In Section 2, we introduce the kernel learning formulation studied in the paper, and establish necessary notations. In Section 3, the concept of effective dimension is motivated from the signal reconstruction problem. Learning bounds for kernel regression problems based on effective dimensionality are obtained in Section 4. Section 5 applies our analysis to some specific kernel learning formulations. We show that the optimal convergence rates can be obtained for various problems. Some final remarks are given in Section 6.

2 Kernel Regression

Consider the problem of predicting a real-valued output y based on its corresponding input vector x. In machine learning, our goal is to estimate a functional relationship $y \approx p(x)$ from a set of training examples. Usually the quality of a predictor $p(x)$ can be measured by a loss function $\phi(p(x), y)$.

In the standard machine learning formulation, we assume that the data (x, y) are drawn from an unknown underlying distribution D. Our goal is to find $p(x)$ so that the expected true loss of p given below is as small as possible:

$$L(p(\cdot)) = E_{x,y}\phi(p(x), y),$$

where we use $E_{x,y}$ to denote the expectation with respect to the true (but unknown) underlying distribution D. Typically, one needs to restrict the hypothesis function family size so that a stable estimate within the function family can be obtained from a finite number of samples. Let the training samples be $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the hypothesis function family that predicts y based on x can be specified with the following kernel method:

$$p(\alpha, x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), \tag{1}$$

where $\alpha = [\alpha_i]_{i=1,\ldots,n}$ is a parameter vector that needs to be estimated from the data. K is a symmetric positive kernel. That is, $K(a, b) = K(b, a)$, and the $n \times n$ Gram matrix $G = [K(x_i, x_j)]_{i,j=1,\ldots,n}$ is always positive semi-definite.

Definition 2.1 *Let $H_0 = \{\sum_{i=1}^{\ell} \alpha_i K(x_i, x) : \ell \in N, \alpha_i \in R\}$. H_0 is an inner product space with norm defined as*

$$\|\sum_i \alpha_i K(x_i, \cdot)\| = (\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j))^{1/2}.$$

Let H be the closure of H_0 under the norm $\|\cdot\|$, which forms a Hilbert space, called the reproducing kernel Hilbert space of K. We denote the corresponding norm as $\|\cdot\|_H$.

It is well-known and not difficult to check that the norm $\|\cdot\|_H$ in Definition 2.1 is well-defined, and it defines an inner product. Further information on reproducing Hilbert spaces can be found in Wahba (1990). For notational purpose, we shall denote $K(x_i, \cdot) \in H$ by $\psi_x \in H$. Definition 2.1 implies that $\forall p \in H$:

$$p(x) = p \cdot \psi_x. \tag{2}$$

Since the reproducing Kernel Hilbert space H can be large, in order to avoid overfitting, it is often necessary to consider models in a bounded convex subset C of H. We would like to find the best model in C defined as:

$$p_C(\cdot) = \arg\inf_{p \in C} L(p) = \arg\inf_{p \in C} E_{x,y}\phi(p(x), y). \tag{3}$$

In supervised learning, we construct an estimator $\hat{p}$ of $p_C(\cdot)$ from a set of n training examples $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Throughout the paper, we use symbol $\hat{}$ to denote empirical quantities based on the n observed training data S. Specifically, we use $\hat{E}_{x,y}$ to denote the empirical expectation with respect to the training samples, and

$$\hat{L}(p) = \hat{E}_{x,y}\phi(p(x), y) = \frac{1}{n}\sum_{i=1}^{n} \phi(p(x_i), y_i).$$

In the standard machine learning analysis, one often assumes that the predictor $\hat{p}$ is taken from a hypothesis function class C that models the relationship of the input x and the output y. A frequently studied learning algorithm is the *empirical risk minimization* (ERM) method, where we find a predictor in C that minimizes the empirical risk:

$$\hat{p} = \arg\inf_{p \in C} \hat{L}(p) = \arg\inf_{p \in C} \frac{1}{n}\sum_{i=1}^{n} \phi(p(x_i), y_i). \tag{4}$$

This formulation is related to a different form of penalized kernel learning formulation used in practical computation. Assume that $C \in H$ is defined by the constraint: $C = \{p \in H : r(p) \leq c_0\}$, where $r : H \rightarrow R$ is a functional called regularization operator. By introducing a Lagrangian multiplier λ_n (≥ 0), we can rewrite (4) as

$$\hat{p} = \arg\inf_{p \in H} \left[\frac{1}{n}\sum_{i=1}^{n} \phi(p(x_i), y_i) + \lambda_n r(p) \right]. \tag{5}$$

For kernel methods with the corresponding reproducing kernel Hilbert space H, one often considers $r(p) = \frac{1}{2}\|p(\cdot)\|^2$. For this choice of $r(\cdot)$, by differentiating (5) and using (2), we obtain the first order condition at the optimal solution $\hat{p}$: $\lambda_n \hat{p} = -\frac{1}{n}\sum_{i=1}^{n} \phi'(p(x_i), y_i)\psi_{x_i}$. That is,

$$\lambda_n \hat{p}(x) = -\frac{1}{n}\sum_{i=1}^{n} \phi'(p(x_i), y_i) K(x_i, x),$$

which is of the form (1). We see that although (5) is formulated as an optimization in a possibly infinite dimensional Hilbert space H, the solution lies in a finite dimensional space, which makes the computation feasible.

Although computationally, (4) and (5) are equivalent when we let $r(p) = \frac{1}{2}\|p(\cdot)\|^2$ with an appropriately selected λ_n, in learning theory, they are analyzed differently since one typically assumes that λ_n is sample-independent when analyzing (5), and assumes that C is sample-independent when analyzing (4).

It is trickier to obtain good generalization bounds for (5), and for technical reasons which we shall not elaborate, the leave-one-out approach in Zhang (2003b) is often more suitable for this purpose. The goal of this paper is to obtain kernel-dependent generalization bounds for the constrained formulation (4), using a concept of effective dimensionality of the underlying kernel representation, which we will introduce in the next section. This kernel dependent effective dimension, denoted as D_λ, is scale-sensitive. The scale parameter λ is closely related to the regularization parameter λ_n in (4). We shall prove (in Theorem 4.1) that with large probability, the generalization error of any training data dependent estimator $\hat{p} \in C$ has a form of:

$$L(\hat{p}) \leq L(p_C) + c(\hat{L}(\hat{p}) - \hat{L}(p_C)) + \inf_{\lambda > 0} [O(\lambda) + O(D_\lambda / n)].$$

where c is a positive constant that only depends on the loss function. This type of bounds are often referred to as *oracle inequalities* in the literature since they compare the risk of an estimator $\hat{p}$ to the best possible (but unknown) predictor p_C. This is also the type of bounds studied in previous works on least squares regression mentioned in the introduction.

Now for the empirical risk minimization method defined in (4), the second term on the right hand side is always non-positive, and hence can be ignored. We are thus mainly interested in the third term, which characterizes the learning complexity. The parameter λ is an arbitrary positive number that can be interpreted as the approximation scale (bias), and D_λ/n can be interpreted as the variance associated with this approximation scale. The complexity in the third term is minimized when we choose the right balance of bias-variance trade-off at $\lambda \approx D_\lambda/n$. If H is a finite dimensional space, then the third term is $O(d/n)$ where $d = \dim(H)$ is the dimension of H. If H is an infinite dimensional space (or when d is large compared to n), one can adjust λ appropriately based on the sample size n to get a bound $O(d_n/n)$. In this case the effective dimension d_n at the optimal scale λ becomes sample-size dependent. However the dimension (at the optimal scale) will never grow faster than $d_n = O(\sqrt{n})$ and hence even in the worse case, the third term converges to zero at a rate no worse than $O(1/\sqrt{n})$.

Note that in some practical kernel learning formulations, a bias term may be included in (1), where the corresponding function space H' has the form $p(\alpha, x) + b = \sum_{i=1}^n \alpha_i K(x_i, x) + b$. It is well-known that H' can be considered as the reproducing kernel Hilbert space with kernel $K'(x_1, x_2) = K(x_1, x_2) + 1$. Therefore we do not consider the formulation with bias in this paper for simplicity. We shall mention that by treating H' as the reproducing kernel Hilbert space of a different kernel, the resulting learning formulations may be slightly different from those in the literature where the bias b is typically not included in the penalization term.

3 Effective dimensionality in the signal reconstruction problem

In this section, we shall motivate the concept of effective dimension using the problem of reconstructing signals from observations that are corrupted with noise. We consider a set of observations x_i and the associated signals $f(x_i)$. We observe corrupted response $y_i = f(x_i) + n_i$, where $\{n_i\}$ are iid noise, drawn from a zero-mean probability distribution with variance σ. Given the set of observed response y_i, the goal is to obtain an estimate $\hat{y}_i \approx f(x_i)$ such that the mean-squared-error

$$\|\hat{y} - f\|^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - f(x_i))^2$$

is as small as possible. Note that we have used $\hat{y}$ and f to denote the vectors $[\hat{y}_i]$ and $[f(x_i)]$ respectively.

Unlike the standard regression formulation in learning problems, in the signal reconstruction problem, the input data points x_i are fixed, and we are only interested in the behavior of $\{f(x_i)\}$ with respect to the random observation $\{y_i\}$. In statistics, this formulation is often referred to as *fixed design* (where each x_i is a design point). The standard learning formulation, where x_i is assumed to be taken from a random distribution, is often referred to as *random design*.

To see the effect of dimensionality with respect to the reconstruction accuracy, we first consider the case that the signal vector $f = [f(x_i)]$ belongs to a known low dimensional space of dimension d, which can be expressed as $f = Pu$, where P is an $n \times d$ projection operator such that $P^T P = I$. We use I to denote the identity matrix. Consider the natural estimator $\hat{f} = PP^T y$, which projects the observation onto the signal subspace, it is easy to see that the expected mean-squared-error is $E\|\hat{f} - f)\|^2 = \frac{d}{n}\sigma^2$. This means that the generalization ability of the estimator is proportional to the dimensionality of the target.

In general, f may not belong to a fixed subspace. However, we can extend the above analysis to any linear estimator $\hat{f} = Sy$, and obtain

$$E\|\hat{f} - f\|^2 = \|Sf - f\|^2 + \frac{\mathrm{tr}(S^T S)}{n}\sigma^2. \tag{6}$$

This gives a bias-variance decomposition. The second term denotes the variance, which is small when $\mathrm{tr}(S^T S)$ is small. This term determines the learning complexity, and is what we are interested in. The first term is the bias term, which can be large when S is not close to the identity matrix I. However if we have good prior knowledge, then we may guess an operator S such that $Sf \approx f$ and $\mathrm{tr}(S^T S)$ is small. Based on the previous paragraph, it is natural to regard $\mathrm{tr}(S^T S)$ as a measure of the effective dimension of the linear estimator S.

The above analysis can be easily applied to kernel methods. A natural way to use kernel representation in signal reconstruction is by solving the following regularized least squares problem:

$$\hat{f} = \arg\min_f \left[\frac{1}{n}(f(x_i) - y_i)^2 + \lambda\|f\|_H\right],$$

where $\|f\|_H$ is the norm of f in the kernel-induced reproducing Hilbert space. Let $G = [K(x_i, x_j)]$ be the kernel Gram matrix, then $\|f\|_H^2 = f^T G^{-1} f$. It follows that the solution can be expressed as

$$\hat{f} = (G + \lambda I)^{-1} G y,$$

where I is the identity operator. Therefore the effective dimension of kernel learning is measured by the quantity

$$\mathrm{tr}((G + \lambda I)^{-2} G^2) \leq \mathrm{tr}((G + \lambda I)^{-1} G).$$

The right hand side is simpler, and is of the same scale as the left hand side when we choose a small λ to balance the bias-variance trade-off. For notation simplicity, in this paper, we shall specify our bounds using the quantity $\mathrm{tr}[(G + \lambda I)^{-1} G]$ on the right-hand side, and regard it as the effective dimension for kernel methods. It can be shown that[1]

$$\mathrm{tr}((G + \lambda I)^{-1} G) = \mathrm{tr}((\tilde{G} + \lambda I)^{-1} \tilde{G}),$$

[1] The proof of this claim is not difficult. However, since it is not important in our analysis except for the purpose of motivating the right effective-dimension formula to use, we shall skip the detailed proof.

where

$$\tilde{G} = \frac{1}{n}\sum_{i=1}^{n} \psi_{x_i}\psi_{x_i}^T, \tag{7}$$

and ψ_x is the representation of x in H as in (2). We have also used the matrix notation $\psi_x\psi_x^T$ to denote the self-adjoint operator $H \to H$ defined as: $(\psi_x\psi_x^T)h = \psi_x(\psi_x \cdot h) = h(x)\psi_x$.

The goal of this paper is to extend the above analysis for fixed design kernel least squares regression to standard learning problems in the random design setting. Moreover, we will derive exponential probability bounds that are often more useful than the expected error bound in (6).

In the setting of learning theory, we consider samples $x_1, \ldots, x_n$ that are taken from a probability distribution. It is natural to replace $\tilde{G}$ in (7) by the expectation over the true distribution: $\tilde{G} = E_x\psi_x\psi_x^T$. Therefore we define the effective dimension associated with kernel representation with a scale (regularization) parameter λ as:

$$D_\lambda = \mathrm{tr}\left[(E_x\psi_x\psi_x^T + \lambda I)^{-1}E_x\psi_x\psi_x^T\right]. \tag{8}$$

In the sequel, we show that this quantity characterizes the learning complexity of kernel regression.

Clearly D_λ is a decreasing function of λ that approaches zero as $\lambda \to \infty$. This means that the effective-dimension becomes small when the approximation scale λ becomes large. Some useful properties of D_λ are listed in Appendix A. In particular, Proposition A.1 gives the relationship of D_λ and eigenvalues in the corresponding kernel PCA.

Properties of D_λ in Appendix A imply that for a finite dimensional space H, D_λ is upper bounded by the dimensionality $\dim(H)$ of the space. Moreover, as $\lambda \to 0$, D_λ converges to the rank of $E_x\psi_x\psi_x^T$, which can be regarded as the physical dimension of the data points ψ_x. As λ increases, the effective dimension D_λ decreases. We always have $\lambda D_\lambda \leq E_x\psi_x^T\psi_x = E_xK(x,x)$. Moreover, the equality holds when $\lambda \to \infty$.

Intuitively, we may interpret the effective dimension D_λ as the number of most significant principle components of ψ_x (through kernel PCA) needed to approximate a function in the unit ball of H at the approximation scale λ.

4 Complexity of Kernel Learning

In this section, we obtain learning bounds that use the effective data dimension D_λ, and establish the main result of the paper in Theorem 4.1.

4.1 Decomposition of loss function

For second-order differentiable convex loss functions that we are interested in, we shall introduce a decomposition that is needed in our analysis. The decomposition approximates convex loss functions using the least squares loss. Since the effective dimension introduced in the previous section was derived from the least squares regression problem, this approximation is quite natural.

Consider a convex subset $C \subset H$, and let $p_C \in C$ be the optimal predictor defined in (3). By differentiating (3) at the optimal solution, and using the convexity of C with respect to p, we obtain the following first order condition:

$$E_{x,y}\phi_1'(p_C(x), y)(p(x) - p_C(x)) \geq 0 \qquad (\forall p \in C), \tag{9}$$

where $\phi_1'(p, y)$ is the derivative of $f(p, y)$ with respect to p.

The following notation becomes convenient later in our analysis.

Definition 4.1 *The Bregman distance of ϕ (with respect to its first variable) is defined as:*

$$d_\phi(p,q;y) = \phi(q,y) - \phi(p,y) - \phi_1'(p,y)(q-p).$$

It is well known (and easy to check) that for a convex function, its Bregman divergence is always non-negative.

We further assume in our analysis that there exist positive constants c_l and c_u such that $0 < c_l \leq \phi_1''(p,y)/2 \leq c_u$, where ϕ_1'' is the second order derivative of f with respect to the first variable. Using Taylor expansion of ϕ, it is easy to see that we have the following inequality for d_ϕ:

$$c_l(p-q)^2 \leq d_\phi(p,q;y) \leq c_u(p-q)^2. \tag{10}$$

Now, $\forall p \in C$, we consider the following decomposition:

$$\phi(p(x),y) - \phi(p_C(x),y) = d_\phi(p_C(x),p(x);y) + \phi_1'(p_C(x),y)(p(x)-p_C(x)).$$

We obtain from (10) the following inequalities which are needed in our analysis:

$$\begin{aligned} & c_l(p(x)-p_C(x))^2 + \phi_1'(p_C(x),y)(p(x)-p_C(x)) \\ \leq & \phi(p(x),y) - \phi(p_C(x),y) \\ \leq & c_u(p(x)-p_C(x))^2 + \phi_1'(p_C(x),y)(p(x)-p_C(x)). \end{aligned} \tag{11}$$

Note that c_l and c_u only depend on the loss function ϕ. For the least squares problem, we can take $c_l = c_u = 1$.

4.2 Empirical ratio inequalities

In this section, we derive some ratio uniform convergence inequalities for kernel methods using the effective dimensionality D_λ. Such uniform convergence inequalities can then be used to obtain generalization bounds for kernel regression methods. Note that similar ratio uniform convergence results can also be obtained using covering numbers (if we have an estimate of such covering numbers), and are also needed in the localized Rademacher complexity analysis. The approach presented here is much more direct and elementary. Consequently, the resulting bounds are cleaner.

Given a positive definite self-adjoint operator $Q : H \to H$, we define an inner product structure on H as:

$$\langle u,v\rangle_Q = u \cdot Qv = u^T Qv.$$

The corresponding norm is $\|u\|_Q = \langle u,u\rangle_Q^{1/2}$.

Given a positive number λ, and let I be the identity operator, we define the following self-adjoint operator on H:

$$Q_\lambda = (E_x \psi_x \psi_x^T + \lambda I)^{-1}.$$

Using this operator, we consider the inner product space T_λ on the set of self-adjoint operators on H, with the inner product defined as

$$\langle A,B\rangle_{T_\lambda} = \mathrm{tr}(AQ_\lambda B),$$

where $\mathrm{tr}(S)$ is the trace of a linear operator S (sum of eigenvalues). The corresponding norm is denoted as $\|\cdot\|_{T_\lambda}$.

We start our analysis with the following simple lemma:

Lemma 4.1 *For all self adjoint operators A, B and Q_λ, we have:*

$$[\mathrm{tr}(AB)]^2 \leq \mathrm{tr}(AQ_\lambda A)\mathrm{tr}(B{Q_\lambda}^{-1}B).$$

Proof Note that $\forall \rho > 0$:

$$\begin{aligned}&\rho^2 \mathrm{tr}(AQ_\lambda A) + \frac{1}{\rho^2}\mathrm{tr}(B{Q_\lambda}^{-1}B)\\ =&\mathrm{tr}[(\rho A{Q_\lambda}^{1/2} - \frac{1}{\rho}B{Q_\lambda}^{-1/2})(\rho A{Q_\lambda}^{1/2} - \frac{1}{\rho}B{Q_\lambda}^{-1/2})^T] + \mathrm{tr}(AB^T) + \mathrm{tr}(BA^T)\\ \geq&\mathrm{tr}(AB^T) + \mathrm{tr}(BA^T) = 2\mathrm{tr}(AB).\end{aligned}$$

Now the lemma can be established by choosing ρ to minimize the left-hand side. □

The following bounds form the foundation of our analysis.

Lemma 4.2 *For any function $a(x, y)$, the following bounds are valid:*

$$\begin{aligned}&\sup_{p\in H} \frac{|\hat{E}_{x,y}a(x,y)p(x) - E_{x,y}a(x,y)p(x)|}{\sqrt{E_x p(x)^2 + \lambda\|p\|_H^2}} \leq \|\hat{E}_{x,y}a(x,y)\psi_x - E_{x,y}a(x,y)\psi_x\|_{Q_\lambda},\\ &\sup_{p\in H} \frac{|\hat{E}_x p(x)^2 - E_x p(x)^2|}{\|p\|_H\sqrt{E_x p(x)^2 + \lambda\|p\|_H^2}} \leq \|\hat{E}_x\psi_x\psi_x^T - E_x\psi_x\psi_x^T\|_{T_\lambda}.\end{aligned}$$

Proof Note that $E_x p(x)^2 + \lambda\|p\|_H^2 = p^T {Q_\lambda}^{-1}p$. Therefore let $v = \hat{E}_{x,y}a(x,y)\psi_x - E_{x,y}a(x,y)\psi_x$, we obtain from Cauchy-Schwartz inequality

$$|\hat{E}_{x,y}a(x,y)p(x) - E_{x,y}a(x,y)p(x)| = |p \cdot v| \leq (p^T{Q_\lambda}^{-1}p)^{1/2}(v^T Q_\lambda v)^{1/2}.$$

This proves the first inequality.

To show the second inequality, we apply Lemma 4.1:

$$\begin{aligned}|\hat{E}_x p(x)^2 - E_x p(x)^2| =&|\mathrm{tr}[(\hat{E}_x\psi_x\psi_x^T - E_x\psi_x\psi_x^T)pp^T]|\\ \leq&\|\hat{E}_x\psi_x\psi_x^T - E_x\psi_x\psi_x^T\|_{T_\lambda}[\mathrm{tr}(pp^T Q_\lambda^{-1}pp^T)]^{1/2}\\ =&\|\hat{E}_x\psi_x\psi_x^T - E_x\psi_x\psi_x^T\|_{T_\lambda}\|p\|_H\sqrt{E_x p(x)^2 + \lambda\|p\|_H^2}.\end{aligned}$$

This proves the second inequality. □

The importance of Lemma 4.2 is that it bounds the behavior of an arbitrary estimator $p \in H$ (which can be sample dependent) in terms of the norm of the empirical mean of n zero-mean Hilbert-space valued random vectors. The convergence rate of the latter can be easily estimated from the variance of the random vectors, and therefore we have significantly simplified the problem.

In order to estimate the variance of the random vectors on the right hand sides of Lemma 4.2, we shall use the notion of effective data dimensionality (at the scale λ) defined in (8), which can now be expressed as:

$$D_\lambda = E_x\psi_x^T Q_\lambda\psi_x = E_x\|\psi_x\|_{Q_\lambda}^2.$$

We also define the following quantities to measure the boundedness of the input data:

$$M_H \geq \sup_x \|\psi_x\|_H, \qquad M_\lambda = \sup_x \|\psi_x\|_{Q_\lambda}. \tag{12}$$

It is easy to see that $M_\lambda \leq M_H/\sqrt{\lambda}$.

Lemma 4.3 *Let $c = \sup_{x,y} a(x, y)$, then we have*

$$E_{x,y} \|a(x, y)\psi_x - E_{x',y'} a(x', y')\psi_{x'}\|_{Q_\lambda}^2 \le c^2 D_\lambda,$$
$$E_x \|\psi_x \psi_x^T - E_{x'} \psi_{x'} \psi_{x'}^T\|_{T_\lambda}^2 \le M_H^2 D_\lambda.$$

Proof Let $\phi = E_{x',y'} a(x', y')\psi_{x'}$, then we have

$$E_{x,y} \|a(x, y)\psi_x - \phi\|_{Q_\lambda}^2 = E_{x,y} \|a(x, y)\psi_x\|_{Q_\lambda}^2 - \|\phi\|_{Q_\lambda}^2 \le c^2 D_\lambda,$$

which gives the first inequality.

Note that $\forall \phi \in H$: $\|\phi\phi^T\|_{T_\lambda} = \|\phi\|_{Q_\lambda} \|\phi\|_H$. Therefore

$$E_x \|\psi_x \psi_x^T - E_{x'} \psi_{x'} \psi_{x'}^T\|_{T_\lambda}^2 = E_x \|\psi_x\|_{Q_\lambda}^2 \|\psi_x\|_H^2 - \|E_x \psi_x \psi_x^T\|_{T_\lambda}^2 \le M_H^2 D_\lambda,$$

leading to the second inequality. □

4.3 Generalization bounds for kernel regression

In order to obtain generalization bounds, we need the following version of Bernstein inequality in Hilbert spaces.

Lemma 4.4 (Yurinsky (1995)) *Let ξ_i be zero-mean independent random vectors in a Hilbert space. If there exist $B, M > 0$ such that for all natural numbers $l \ge 2$: $\frac{1}{n}\sum_{i=1}^n E\|\xi_i\|_H^l \le \frac{B^2}{2} l! M^{l-2}$. Then for all $\delta > 0$: $P(\|\frac{1}{n}\sum_i \xi_i\|_H \ge \delta) \le 2\exp(-\frac{n}{2}\delta^2/(B^2 + \delta M))$.*

In this paper, we shall use the following variant of the above bound for convenience.

$$P\left(\left\|\frac{1}{n}\sum_i \xi_i\right\|_H \ge \frac{2Mt}{n} + \sqrt{\frac{2t}{n}} B\right) \le 2\exp(-t). \tag{13}$$

Lemma 4.5 *Under the assumptions of Lemma 4.3, let $\epsilon_\lambda(t) = \sqrt{\frac{2tD_\lambda}{n} + \frac{2tM_\lambda}{n}}$. Then with probability of at least $1 - 2\exp(-t)$:*

$$\sup_{p \in H} \frac{|\hat{E}_{x,y} a(x, y)p(x) - E_{x,y} a(x, y)p(x)|}{\sqrt{E_x p(x)^2 + \lambda \|p\|_H^2}} \le \epsilon_\lambda(t)c.$$

Similarly, with probability of at least $1 - 2\exp(-t)$, we have:

$$\sup_{p \in H} \frac{|\hat{E}_x p(x)^2 - E_x p(x)^2|}{\|p\|_H \sqrt{E_x p(x)^2 + \lambda \|p\|_H^2}} \le \epsilon_\lambda(t) M_H.$$

Proof For the first inequality, we let random vector ξ_i be $a(x_i, y_i)\psi_{x_i} - E_{x'} a(x', y')\psi_{x'}$ under the Q_λ-norm. Clearly $\|\xi_i\|_{Q_\lambda} \le 2cM_\lambda$, and $E\|\xi_i\|_{Q_\lambda}^2 \le c^2 D_\lambda$ by Lemma 4.3. It thus follows that we can let $B = c{D_\lambda}^{1/2}$ and $M = cM_\lambda$ in Lemma 4.4. The first inequality thus follows from Lemma 4.2 and (13).

Similarly, let $\xi_i = \psi_{x_i} \psi_{x_i}^T - E_{x'} \psi_{x'} \psi_{x'}^T$ under the T_λ-norm. We have $\|\xi_i\|_{T_\lambda} \le 2M_H M_\lambda$, and by Lemma 4.3 $E\|\xi_i\|_{T_\lambda}^2 \le M_H^2 D_\lambda$. It thus follows that we can let $B = M_H {D_\lambda}^{1/2}$ and $M = M_H M_\lambda$ in Lemma 4.4. The second inequality follows from Lemma 4.2 and (13). □

We are now ready to derive the main result of the paper:

Theorem 4.1 *Assume that* $\sup_{x,y} |\phi_1'(p_C(x), y)| \leq b_C c_l$, *where* c_l *and* c_u *satisfy (10). Consider any sample dependent estimator* $\hat{p}$ *such that* $\hat{p} \in C$ *(that is,* $\hat{p} \in C$ *is a function of the training sample* S*). Let* $\epsilon_\lambda(t) = \sqrt{\frac{2tD_\lambda}{n} + \frac{2tM_\lambda}{n}}$. *Then* $\forall \lambda > 0$, *with probability of at least* $1 - 4\exp(-t)$, *the generalization error is bounded as:*

$$L(\hat{p}) \leq L(p_C) + \frac{2c_u}{c_l}[\hat{L}(\hat{p}) - \hat{L}(p_C)] + \lambda c_l \|\hat{p} - p_C\|_H^2 + \frac{c_u^2 \epsilon_\lambda(t)^2}{c_l}[b_C + M_H\|\hat{p} - p_C\|_H]^2.$$

Proof We introduce the following notations for convenience:

$$\begin{aligned}
&\hat{A}(p) = \hat{E}_{x,y}\phi_1'(p_C(x), y)(p(x) - p_C(x)), \quad A(p) = E_{x,y}\phi_1'(p_C(x), y)(p(x) - p_C(x)),\\
&\hat{B}(p) = \hat{E}_x(p(x) - p_C(x))^2, \quad B(p) = E_x(p(x) - p_C(x))^2,\\
&E(p) = B(p) + \lambda\|p - p_C\|_H^2.
\end{aligned}$$

We obtain from Lemma 4.5 that with probability of at least $1 - 4\exp(-t)$:

$$|\hat{A}(\hat{p}) - A(\hat{p})| \leq \epsilon_\lambda(t) b_C c_l E(\hat{p})^{1/2}, \qquad |\hat{B}(\hat{p}) - B(\hat{p})| \leq \epsilon_\lambda(t) M_H \|\hat{p} - p_C\|_H E(\hat{p})^{1/2}.$$

Combining the above two inequalities, we obtain:

$$|\hat{A}(\hat{p}) - A(\hat{p})| + c_l|\hat{B}(\hat{p}) - B(\hat{p})| \leq \epsilon_\lambda(t) E(\hat{p})^{1/2} c_l [b_C + M_H\|\hat{p} - p_C\|_H].$$

Using (11) and recalling (9), we obtain

$$\frac{c_l}{c_u}[L(\hat{p}) - L(p_C)] \leq [\hat{L}(\hat{p}) - \hat{L}(p_C)] + \epsilon_\lambda(t) E(\hat{p})^{1/2} c_l [b_C + M_H\|\hat{p} - p_C\|_H]. \tag{14}$$

Let

$$K_1(p) = [L(p) - L(p_C)] + \lambda c_l\|p - p_C\|_H^2, \quad \hat{K}_2(p) = [\hat{L}(p) - \hat{L}(p_C)] + \lambda\frac{c_l^2}{c_u}\|p - p_C\|_H^2,$$

then (9) and (11) imply that $c_l E(p) \leq K_1(p)$. We can derive from (14)

$$\begin{aligned}
\frac{c_l}{c_u}K_1(\hat{p}) \leq &\hat{K}_2(\hat{p}) + \epsilon_\lambda(t)\sqrt{\frac{K_1(\hat{p})}{c_l}} c_l [b_C + M_H\|\hat{p} - p_C\|_H]\\
\leq &\hat{K}_2(\hat{p}) + \frac{c_l}{2c_u}K_1(\hat{p}) + \epsilon_\lambda(t)^2\frac{c_u}{2}[b_C + M_H\|\hat{p} - p_C\|_H]^2.
\end{aligned}$$

Rearranging the above inequality, we obtain the theorem. □

Note that the above result holds for any data-dependent estimator. Specifically, for the empirical risk minimization estimator $\hat{p}$ that is defined through (4), we have $\hat{L}(\hat{p}) - \hat{L}(p_C) \leq 0$. Therefore the second term on the right-hand-side of the bound is non-positive, which can be ignored. That is, we have

$$L(\hat{p}) \leq L(p_C) + \lambda c_l\|\hat{p} - p_C\|_H^2 + \frac{c_u^2\epsilon_\lambda(t)^2}{c_l}[b_C + M_H\|\hat{p} - p_C\|_H]^2.$$

Since the above bound holds for all $\lambda > 0$, we may choose a more specific λ to simplify the bound. The following result gives such a simplified version of Theorem 4.1.

Corollary 4.1 *Under the assumptions of Theorem 4.1,* $\forall t \geq 0$ *and* $\forall \lambda$ *such that* $\lambda n \geq \max(D_\lambda, 1) M_H^2$*, the following bound holds with probability of at least* $1 - 4\exp(-t)$*,*

$$L(\hat{p}) \leq L(p_C) + \frac{2c_u}{c_l}[\hat{L}(\hat{p}) - \hat{L}(p_C)] + \lambda \frac{2c_u^2(2t+1)^2}{c_l M_H^2}[b_C + M_H \|\hat{p} - p_C\|_H]^2.$$

Proof We shall apply Theorem 4.1. Using the bound $M_\lambda \leq M_H/\sqrt{\lambda}$, we obtain

$$\epsilon_\lambda(t)^2 \leq (4tD_\lambda n + 8t^2 M_H^2/\lambda) n^{-2} \leq \lambda(4t + 8t^2)/M_H^2.$$

We also have

$$\lambda c_l \|\hat{p} - p_C\|_H^2 \leq \frac{\lambda c_u^2}{c_l M_H^2}[b_C + M_H \|\hat{p} - p_C\|_H]^2.$$

Substituting into Theorem 4.1, we obtain the desired bound. □

Since when $\lambda \to \infty$, $D_\lambda \to 0$, the assumption of Corollary 4.1 can always be satisfied with a parameter λ that is sufficiently large. The quality $\|\hat{p} - p_C\|_H$ is never larger than the diameter of C: $d_H(C) = \sup\{\|p_1 - p_2\|_H : p_1, p_2 \in C\}$. It can thus be upper bounded by a constant when $d_H(C)$ is finite.

The parameters c_l, c_u and b_C depend on the loss function ϕ. For the least squares loss, $\phi(p, y) = (p - y)^2$, we have $c_l = c_u = 1$. Now assume that $\sup\{\|p\|_H : p \in C\} \leq A$, and $y \in [-b, b]$. Then we have $d_H(C) \leq 2A$ and $b_C \leq 2(AM_H + b)$.

For the empirical risk minimization estimator defined in (4), we obtain the following bound from Corollary 4.1 with probability of at least $1 - 4\exp(-t)$:

$$E_{x,y}(\hat{p}(x) - y)^2 \leq \inf_{p \in C} E_{x,y}(p(x) - y)^2 + \lambda \frac{8(2t+1)^2}{M_H^2}[2M_H A + b]^2,$$

when $\lambda n \geq D_\lambda M_H^2$.

The optimal bound can be obtained when we pick λ as the solution of the fixed point equation $\lambda n = D_\lambda M_H^2$. The solution exists since when $\lambda = 0$, the left hand side is smaller than the right hand side, and when $\lambda \to \infty$, the right hand side is smaller than the left hand side. As we shall see in Section 5, the rate obtained using the optimally chosen λ agrees with the minimax rate for various problems. It is also similar to the rate from the localized Rademacher analysis with the complexity estimate given in Mendelson (2003).

5 Examples

As we can see from Section 3, the concept of effective dimension is motivated through direct computation of the bias-variance decomposition in the signal reconstruction problem. One can thus expect that the generalization bound we obtained in Theorem 4.1 is quite tight. Indeed, in this section, we show that rates we obtain achieve the best possible minimax rates for some kernel regression problems for which the optimal rates are known.

We will only consider empirical estimator $\hat{p}$ that minimizes $\hat{L}(p)$ in C. In this case, $[\hat{L}(\hat{p}) - \hat{L}(p_C)] \leq 0$ in Corollary 4.1. Therefore we obtain the following bound: with probability of at least $1 - 4\exp(-t)$:

$$L(\hat{p}) \leq L(p_C) + a_C(t) \inf\{\lambda : \lambda n \geq D_\lambda M_H^2\},$$

where

$$a_C(t) = \frac{2c_u^2(2t+1)^2}{c_l M_H^2}[b_C + M_H d_H(C)]^2$$

can be regarded as a constant.

5.1 Worst case effective dimensionality and generalization

In the worst case, we have $D_\lambda \leq M_H^2/\lambda$. Therefore we can let $\lambda = M_H^2/\sqrt{n}$ in Corollary 4.1, which leads to the following bound with probability at least $1 - 4\exp(-t)$:

$$L(\hat{p}) \leq L(p_C) + \frac{2c_u^2(2t+1)^2}{c_l\sqrt{n}}[b_C + M_H\|\hat{p} - p_C\|_H]^2.$$

This implies that the convergence rate is $O(1/\sqrt{n})$ in the worst case. This is the best possible kernel-independent rate (compare with Section 5.3).

5.2 Finite dimensional problems

We can use the bound $D_\lambda \leq \dim(H)$. Now let $\lambda = \dim(H)M_H^2/n$ in Corollary 4.1, we obtain with probability at least $1 - 4\exp(-t)$:

$$L(\hat{p}) \leq L(p_C) + \frac{2c_u^2(2t+1)^2\dim(H)}{c_l n}[b_C + M_H\|\hat{p} - p_C\|_H]^2.$$

It is well known that the convergence rate of the order $O(\dim(H)/n)$ is optimal in this case.

5.3 Smoothing splines

We only consider 1-dimensional problems. For smoothing splines, the corresponding Hilbert space consists of functions p satisfying the smoothness condition that $\int[p^{(s)}(x)]^2dx$ is bounded ($p^{(s)}$ is the s-th derivative of p and $s > 1/2$). We may consider periodic functions (or their restrictions in an interval) and the condition corresponds to a decaying Fourier coefficients condition. Specifically, the space can be regarded as the reproducing kernel Hilbert space with kernel

$$K(x_1, x_2) = \sum_{k\geq 0}(k+1)^{-2s}(\sin(kx_1)\sin(kx_2) + \cos(kx_1)\cos(kx_2)).$$

Now, using Proposition A.4, we have $D_\lambda \leq \inf_{k\geq 1}\left[2k + \frac{2/\lambda}{(2s-1)k^{2s-1}}\right]$. Therefore $D_{k^{-2s}} \leq \frac{4sk}{2s-1}$. Note that we may take $M_H^2 = 2s/(2s-1)$. Therefore we can let $\lambda = k^{-2s}$ in Corollary 4.1 where k is the largest integer such that $k^{2s+1} \leq \frac{(2s-1)^2 n}{8s^2}$. This gives the following bound (with probability at least $1 - 4\exp(-t)$).

$$L(\hat{p}) \leq L(p_C) + \left(\frac{8s^2}{(2s-1)^2 n}\right)^{2s/(2s+1)} \frac{c_u^2(2s-1)(2t+1)^2}{c_l s}[b_C + M_H d_H(C)]^2.$$

This rate of $O(n^{-2s/(2s+1)})$ matches the best possible convergence rate for any data-dependent estimator. Note that the lower bound is well-known in the non-parametric statistical literature (for example, see Stone (1982)).

5.4 Exponential kernel

Exponential kernel has recently been popularized by Vapnik. Again for simplicity we consider 1-dimensional problems where $x \in [-1, 1]$. The kernel function is given by

$$K(x_1, x_2) = \exp(x_1 x_2) = \sum_{i=0}^{n} \frac{1}{i!} x_1^i x_2^i.$$

Now, using Proposition A.4, we have $D_\lambda \le \inf_{k \ge 0}[k + 2 + \frac{1}{\lambda k!}]$. Therefore $D_{k^{-k}} \le 2k + 3$. Note that $M_H^2 \le e$, therefore for sufficiently large n, we can let $\lambda = k_*^{-k_*}$ in Corollary 4.1 with $10 k_*^{k_*+1} = n$, where the optimal solution $k_* \le \frac{2 \ln n}{\ln \ln n}$. Therefore we can take $\lambda = \frac{10 k_*}{n} \le \frac{20 \ln n}{n \ln \ln n}$. This means that at the optimal scale, the effective dimension is at most $O(\ln n / \ln \ln n)$. Now Corollary 4.1 implies a generalization bound of the form $L(\hat{p}) \le L(p_C) + O(\frac{\ln n}{n \ln \ln n})$.

6 Conclusion

In this paper, we introduced a concept of scale-sensitive effective data dimension, and used it to derive generalization bounds for some kernel regression problems. The resulting convergence rates are optimal for various learning formulations.

The effective dimension at the appropriately chosen optimal scale can be sample-size dependent and behaves like $\sqrt{n}$ in the worst case. This shows that despite the claim that a kernel method learns a predictor from an infinite dimensional Hilbert space, the effective dimension is sub-polynomial in the sample size. Therefore kernel methods are not more powerful than learning in an appropriately chosen finite dimensional space.

The formulation of effective dimension in Appendix A suggests that we may use the largest few eigen-functions in the kernel PCA to approximate the effective dimensions. This has interesting computation implications: in certain applications, one may be able to obtain the eigen-functions relatively easily.[2] For such problems, it is can be more efficient computationally to solve a regression problem in the subspace spanned by the eigen-functions corresponding to the largest eigenvalues, instead of using the standard kernel method. This is because with n-samples, kernel methods require n parameters (one for each data point) in the computation. However as we have shown, the effective number of parameters (effective dimension) is not more than $O(\sqrt{n})$. Therefore it could be possible to significantly reduce the computational cost of kernel methods by explicitly parameterizing the effective dimensions using the principle eigen-functions (if they can be obtained relatively easily).

A Properties of the effective dimension

We give properties of the scale-sensitive data dimension D_λ.

The following characterization of D_λ is very useful for estimating the quantity. It relates D_λ to the eigen-decomposition of the data in the feature space (kernel PCA).

Proposition A.1 *Consider a complete set of ortho-normal eigen-pairs $\{(\lambda_i, u_i) : i \ge 1\}$ of the operator $E_x \psi_x \psi_x^T$, where $u_i \cdot u_j = 0$ if $i \ne j$ and $u_i \cdot u_i = 1$. We have the identity: $D_\lambda = \sum_i \frac{\lambda_i}{\lambda_i + \lambda}$.*

[2] Another possibility is to direct model the eigen-functions without going through the kernels.

Proof Clearly $\{u_i\}$ forms a complete set of eigen-vectors for the operator $(E_x\psi_x\psi_x^T+\lambda I)^{-1}E_x\psi_x\psi_x^T$. The corresponding eigenvalues are $\lambda_i/(\lambda_i+\lambda)$. Since the trace of an operator is the sum of its eigenvalues, we obtain the equality. □

The following result implies that the quantity D_λ behaves like dimension if the underlying space H is finite dimensional.

Proposition A.2 *If H is a finite dimensional space, then $D_\lambda \leq \dim(H)$. In addition, $\lim_{\lambda\to 0} D_\lambda = \mathrm{rank}(E_x\psi_x\psi_x^T)$, where we use* rank *to denote the rank of a matrix.*

Proof The number of nonzero eigenvalues r of $E_x\psi_x\psi_x^T)$ is its rank, where $r \leq \dim(H)$. Let the corresponding eigenvalues be $\lambda_i, \ldots \lambda_r$. Therefore $D_\lambda = \sum_{i=1}^r \lambda_i/(\lambda_i+\lambda) \leq r$ and $D_\lambda \to r$ as $\lambda \to 0$. □

The following results give an upper bound of effective dimension D_λ at an approximation scale λ that is independent of the feature-space dimensionality.

Proposition A.3 *For all Hilbert spaces H, we have the following bound $D_\lambda \leq M_H^2/\lambda$, where M_H is defined in (12).*

Proof $D_\lambda \leq \mathrm{tr}(E_x\psi_x\psi_x^T)/\lambda = E_x\|\psi_x\|_H^2/\lambda \leq M_H^2/\lambda$. □

In many cases, we can find a so-called feature representation of the kernel function $K(x_1,x_2) = \psi_{x_1}\cdot\psi_{x_2}$ as $K(x_1,x_2)=\sum_{j=1}^\infty \psi_j(x_1)\cdot\psi_j(x_2)$, where each $\psi_j(x)$ is a real-valued function which gives a feature component of data x. Under this representation, we can obtain the following bound on the effective dimension.

Proposition A.4 *Consider the following feature space decomposition of kernel: $\psi_{x_1}\cdot\psi_{x_2} = \sum_i \psi_j(x_1)\psi_j(x_2)$, where each ψ_j is a real valued function. If $\lambda_1 \geq \lambda_2\cdots$, then we have the following bound: $\sum_{j\geq k}\lambda_j \leq E_x\sum_{j\geq k}\psi_j(x)^2$. This implies*

$$D_\lambda \leq \inf_{k\geq 0}\left[k + E_x\sum_{j>k}\psi_j(x)^2/\lambda\right].$$

Proof We can represent each ψ_x using the feature representation as: $\psi_x(\cdot)=\sum_j \psi_j(x)\psi_j(\cdot)$. This induces a norm preserving embedding $\sum_i \alpha_i\psi_{x_i} \in H \to \{\sum_i \alpha_i\psi_j(x_i)\} \in L_2$. Therefore technically, we may consider H belonging to $\tilde{H}$ which are spanned by $\{\psi_j(x)\}$, with the quotient norm induced from the above embedding. For the sake of simplicity, we still denote $\tilde{H}$ by H. Now for any k, let $P_k : H \to H$ be the projection operator onto the subspace spanned by $\{\psi_j(\cdot)\}_{j=1}^k$. Let $P_k\psi_x = \phi_x^k$ and $(I-P_k)\psi_x = \psi_x - \phi_x^k = \bar{\phi}_x^k$. It is clearly that $\|\bar{\phi}_x^k\|_H^2 \leq \sum_{j>k}\psi_j(x)^2$. Therefore

$$\begin{aligned}
D_\lambda =&\mathrm{tr}[(E_x\psi_x\psi_x^T+\lambda I)^{-1}E_x\phi_x^k\phi_x^{k\,T}] + \mathrm{tr}[(E_x\psi_x\psi_x^T+\lambda I)^{-1}E_x\bar{\phi}_x^k\bar{\phi}_x^{k\,T}]\\
\leq&\mathrm{tr}[(E_x\phi_x^k\phi_x^{k\,T}+\lambda I)^{-1}E_x\phi_x^k\phi_x^{k\,T}] + \mathrm{tr}[(\lambda I)^{-1}E_x\bar{\phi}_x^k\bar{\phi}_x^{k\,T}]\\
\leq&k + E_x\|\bar{\phi}_x^k\|_H^2/\lambda\\
\leq&k + E_x\sum_{j>k}\psi_j(x)^2/\lambda.
\end{aligned}$$

Note that in the above derivation, the second inequality follows from Proposition A.2 and the fact that ϕ_x^k belongs to a subspace of H which has dimension at most k. □

References

Bartlett, P., Bousquet, O., and Mendelson, S. (2002). Localiazed Radeacher complexity. In *Proceedings of the Annual Conference on Computational Learning Theory*, volume 2375 of *LNAI*, Sydney. Springer.

Chucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49.

Guo, Y., Bartlett, P. L., Shawe-Taylor, J., and Williamson, R. C. (2002). Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1):239–250.

Lee, W., Bartlett, P., and Williamson, R. (1998). The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory*, 44(5):1974–1980.

Mendelson, S. (2002). Improving the sample complexity using global data. *IEEE Tran. Inf. Theory*, 48(7):1977–1991.

Mendelson, S. (2003). On the performance of kernel classes. *JMLR*, 4:759–771.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053.

van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS-NSF Regional Conference series in applied mathematics. SIAM, Philadelphia, PA.

Williamson, R. C., Smola, A., and Schlkpof, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532.

Yurinsky, V. (1995). *Sums and Gaussian vectors.* Springer-Verlag, Berlin.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550.

Zhang, T. (2003a). Effective dimension and generalization of kernel learning. In *NIPS 2002*.

Zhang, T. (2003b). Leave-one-out bounds for kernel methods. *Neural Computation*, 15:1397–1437.