Covering Numbers for Support Vector Machines

Ying Guo, *Student Member, IEEE*, Peter L. Bartlett, *Member, IEEE*, John Shawe-Taylor, *Member, IEEE*, and Robert C. Williamson, *Member, IEEE*

***Abstract*—Support vector (SV) machines are linear classifiers that use the maximum margin hyperplane in a feature space defined by a kernel function. Until recently, the only bounds on the generalization performance of SV machines (within Valiant's probably approximately correct framework) took no account of the kernel used except in its effect on the margin and radius. More recently, it has been shown that one can bound the relevant covering numbers using tools from functional analysis. In this paper, we show that the resulting bound can be greatly simplified. The new bound involves the eigenvalues of the integral operator induced by the kernel. It shows that the effective dimension depends on the rate of decay of these eigenvalues. We present an explicit calculation of covering numbers for an SV machine using a Gaussian kernel, which is significantly better than that implied by previous results.**

***Index Terms*—Covering numbers, entropy numbers, kernel machines, statistical learning theory, support vector (SV) machines.**

I. Introduction

SUPPORT vector (SV) machines [5], [6] are learning algorithms based on maximum margin hyperplanes [4] which make use of an implicit mapping into feature space by using a general kernel function in place of the standard inner product. Consequently, one can apply an analysis for the maximum margin algorithm directly to SV machines. However, such a process ignores the effect of the kernel. Intuitively one would expect that a "smoother" kernel would somehow reduce the capacity of the learning machine thus leading to better bounds on generalization error if the machine could attain a small training error.

In [15], [16] it has been shown that this intuition is justified. The main result there (quoted later) gives a bound on the covering numbers for the class of functions computed with SV machines. This bound along with statistical results in [3] and [11] results in bounds that explicitly depend on the kernel used. The intuitive idea that eigenvalues of kernels must have something to say about generalization performance has also been previously explored by others in a different analysis framework: see the (simultaneous and independent) development in terms of regularization theory in [14] and [7]. One can also recover a dependence of covering numbers on eigenvalues in a different setting: in [13] it was shown how the eigenvalues of the empirical gram matrix can bound the empirical covering numbers and in turn how generalization results can be obtained that way. The covering number bounds of the present paper do not depend on the particular data observed.

In the traditional viewpoint of statistical learning theory, one is given a class of functions $\mathcal{F}$, and the generalization performance attainable using $\mathcal{F}$ is determined via the covering numbers $\mathcal{N}(\epsilon, \mathcal{F})$ (precise definitions are given in what follows). Many generalization error bounds can be expressed in terms of $\mathcal{N}(\epsilon, \mathcal{F})$. The main method of bounding $\mathcal{N}(\epsilon, \mathcal{F})$ has been to use the Vapnik–Chervonenkis dimension or one of its generalizations (see [1], [2] for an overview).

In [15], [16], the class $\mathcal{F}$ is viewed as being generated by an integral operator induced by the kernel, and properties of this operator are used to bound the required covering numbers. The result is in a form that is not particularly easy to use (see (15) and (16)).

The main technical result of this paper is a covering number bound based on this result that is amenable to direct calculation. We illustrate the new result by bounding the covering numbers of SV machines which use Gaussian radial basis function (RBF) kernels with variance σ^2. The result shows the influence of the variance on the covering numbers: the covering number bound decreases when σ^2 increases. More generally, the main result makes model order selection possible using any parameterized family of kernel functions, since it describes how the capacity of the class is affected by changes to the kernel.

For $0 < p \leq \infty$ and $d \in \mathbb{N}$, define the spaces

$$\ell_p^d := \{\boldsymbol{x} \in \mathbb{R}^d \colon \|\boldsymbol{x}\|_{\ell_p^d} < \infty\}$$

where the p-norms are

$$\|\boldsymbol{x}\|_{\ell_p^d} := \left(\sum_{j=1}^{d} |x_j|^p\right)^{\frac{1}{p}}, \qquad \text{for } 0 < p < \infty$$

$$\|\boldsymbol{x}\|_{\ell_\infty^d} := \sup_{j=1,\ldots,d} |x_j|, \qquad \text{for } p = \infty.$$

For $0 < p \leq \infty$, we write $\ell_p = \ell_p^\infty$ and the norms are defined similarly.

For $\epsilon > 0$ and a subset $\mathcal{F}$ of a metric space, an *ϵ-cover for $\mathcal{F}$* with respect to the metric ρ is a subset $\hat{\mathcal{F}}$ of the metric space for which every $f \in \mathcal{F}$ has a $\hat{f} \in \hat{\mathcal{F}}$ satisfying $\rho(f, \hat{f}) \leq \epsilon$. The *$\epsilon$-covering number of $\mathcal{F}$ with respect to the metric ρ* denoted $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ is the size of the smallest ϵ-cover for $\mathcal{F}$. Given

Manuscript received February 7, 2000; revised November 6, 2000. This work was supported by the Australian Research Council. This research was carried out while P. L. Bartlett and J. Shawe-Taylor were at the Australian National University. The material in this paper was presented at the Twelfth Annual Conference on Computational Learning Theory, Santa Cruz, CA, July 1999.

Y. Guo and R. C. Williamson are with the Department of Telecommunications Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra 0200, Australia (e-mail: Ying.Guo@anu.edu.au; Bob.Williamson@anu.edu.au).

P. L. Bartlett is with BIOwulf Technologies, Berkeley, CA 94704 USA (e-mail: Peter.Bartlett@anu.edu.au).

J. Shawe-Taylor is with the Department of Computer Science, Royal Holloway College, University of London, Egham, TW20 0EX, U.K. (e-mail: jst@dcs.rhbnc.ac.uk).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(02)00050-0.

m points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \ell_p^d$, we use the shorthand $\boldsymbol{X}^m = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$. Suppose $\mathcal{F}$ is a class of functions defined on $\mathbb{R}^d$. The ℓ_∞^d norm *with respect to* $\boldsymbol{X}^m$ of $f \in \mathcal{F}$ is defined as

$$\|f\|_{\ell_\infty^{\boldsymbol{X}^m}} := \max_{i=1,\ldots,m} |f(\boldsymbol{x}_i)|.$$

To simplify notation, we use $\ell_\infty^{\boldsymbol{X}^m}$ to denote both the space and the metric induced by the norm in that space. The input space is taken to be $\mathcal{X}$, a compact subset of $\mathbb{R}^d$.

Let k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel satisfying the hypotheses of Mercer's theorem (see Theorem 2). Given m points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}$ as the input data, we will map the input data into a feature space $\mathcal{S}$ (which is in fact a Hilbert space) via a mapping Φ. We let $\tilde{\boldsymbol{x}} := \Phi(\boldsymbol{x})$, and denote by $\mathcal{F}_{R_{\boldsymbol{w}}}(\boldsymbol{X}^m)$ the hypothesis class implemented by SV machines on an m-sample with weight vector (in feature space $\mathcal{S}$) bounded by $R_{\boldsymbol{w}}$

$$\mathcal{F}_{R_{\boldsymbol{w}}}(\boldsymbol{X}^m) := \{\boldsymbol{X}^m \mapsto (\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}_1\rangle, \ldots, \langle \boldsymbol{w}, \tilde{\boldsymbol{x}}_m\rangle)\colon \boldsymbol{x}_i \in \boldsymbol{X}^m, \boldsymbol{X}^m \in \mathcal{X}^m, \boldsymbol{w} \in \mathcal{S}, \|\boldsymbol{w}\| \le R_{\boldsymbol{w}}\} \quad (1)$$

and the hypothesis class $\mathcal{F}_{R_{\boldsymbol{w}}}$ on $\mathcal{X}^m$ is defined as

$$\mathcal{F}_{R_{\boldsymbol{w}}} = \bigcup_{m=1}^{\infty} \bigcup_{\boldsymbol{X}^m \in \mathcal{X}^m} \mathcal{F}_{R_{\boldsymbol{w}}}(\boldsymbol{X}^m). \quad (2)$$

Here, $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{S}$. Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ be the eigenvalues of the integral operator

$$T_k\colon L_2(\mathcal{X}) \to L_2(\mathcal{X})$$
$$T_k\colon f \mapsto \int_{\mathcal{X}} k(\cdot, \boldsymbol{y}) f(\boldsymbol{y})\, d\boldsymbol{y}$$

and denote by $\psi_n(\cdot)$, $n \in \mathbb{N}$ the corresponding eigenfunctions. (The eigenvalues are real and nonnegative because of the assumptions on k—see the next section.) For translation invariant kernels (such as $k(x, y) = \exp((x-y)^2/\sigma^2)$), the eigenvalues are given by

$$\lambda_i = \sqrt{2\pi}\, K(j\omega_0) \quad (3)$$

for $j \in \mathbb{Z}$, where $K(\omega) = F[k(x)](\omega)$ is the Fourier transform of $k(\cdot)$ (see [15], [16] for further details; and see Section IV for an explanation of ω_0). For smoother kernels, the Fourier transform $F(j\omega_0)$ decreases faster. (There are fewer"high-frequency components.") Thus, for smooth kernels, λ_i decreases to zero rapidly for increasing i.

Theorem 1 (Main Result): Suppose k is a kernel satisfying the hypothesis of Mercer's theorem. Let the hypothesis class $\mathcal{F}_{R_{\boldsymbol{w}}}$, eigenfunctions $\psi_n(\cdot)$ and eigenvalues $(\lambda_i)_i$ be defined as above. Suppose

$$C_k := \sup_n \|\psi_n\|_{L_\infty} < \infty. \quad (4)$$

Then, for $n \in \mathbb{N}$, the minimum

$$j_n^* = \min\left\{ j\colon \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} \right\}$$

always exists. Define

$$\epsilon_n^* = 6R_{\boldsymbol{w}}C_k \sqrt{j_n^* \left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2}\right)^{\frac{1}{j_n^*}} + \sum_{i=j_n^*+1}^{\infty} \lambda_i}. \quad (5)$$

Then

$$\sup \mathcal{N}\left(\epsilon_n^*, \mathcal{F}_{R_{\boldsymbol{w}}}, \ell_\infty^{\boldsymbol{X}^m}\right) \le n \quad (6)$$

where the supremum is taken over all m-tuples of data points, i.e., $\boldsymbol{X}^m \in \mathcal{X}^m$.

Although the left-hand side of (6) depends on m, the inequality remains true for all m. The quantity ϵ_n^* is an upper bound on the *entropy number* of $\mathcal{F}_{R_{\boldsymbol{w}}}$, which is the functional inverse of the covering number. In this theorem, the number j_n^* has a natural interpretation. If $j_n^* = d$ is independent of n, then from (5) we can obtain

$$\epsilon_n(\mathcal{F}_{R_{\boldsymbol{w}}}) = O\left(\left(\frac{1}{n}\right)^{\frac{1}{d}}\right)$$
$$\Rightarrow \sup_{\boldsymbol{X}^m \in \mathcal{X}^m} \mathcal{N}\left(\epsilon, \mathcal{F}_{R_{\boldsymbol{w}}}, \ell_\infty^{\boldsymbol{X}^m}\right) = O\left(\left(\frac{1}{\epsilon}\right)^d\right).$$

Hence, for a given value of n, j_n^* can be viewed as the *effective dimension* of the function class. Clearly, this effective dimension depends on the rate of decay of the eigenvalues. As expected, for smooth kernels (which have rapidly decreasing eigenvalues), the effective dimension is small. In the following, we write j^* for j_n^*.

Before proceeding with the formal part of the paper, we very briefly outline (for those unfamiliar with the setting) the learning model used in order to motivate the results. More details can be found in [15] and references therein or in several recent textbooks such as [2]. The setting is learning from examples. The learning machine is given a training sequence $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ where the $\boldsymbol{x}_i$ are drawn independently from some unknown probability distribution P. The y_i are given by some teacher (think of it as a function so $y_i = f(\boldsymbol{x}_i)$). Of course, f is not known, the aim being to learn it. Algorithmically, one can minimize the *empirical risk*

$$R_{\text{emp}}(\hat{f}) := \sum_{i=1}^{m} (y_i - \hat{f}(\boldsymbol{x}_i))^2$$

of some estimate $\hat{f}$ but what one would really like to minimize the *expected risk*

$$R(\hat{f}) = E_P((y_i - \hat{f}(\boldsymbol{x}_i))^2).$$

(Here we are using the "squared loss"; other choices are possible.) An important theoretical question then is: "If for some $\hat{f}$, $R_{\text{emp}}(\hat{f})$ is small, does this mean $R(\hat{f})$ is as well?" It turns out that one can bound the difference between $R_{\text{emp}}(\hat{f})$ and $R(\hat{f})$ in a probabilistic sense, and such bounds are in terms of the covering numbers of the class of hypotheses from which $\hat{f}$ had the possibility of being drawn from—a better bound being obtained using smaller covering number estimates. The particular covering numbers needed are those used in the statement of the main theorem above.

The remainder of the paper is organized as follows. We start by introducing notation and definitions (Section II). Section III contains the main result (the proof is in Appendix A). Section IV contains an example of the application of the main result. Section V concludes the paper.

II. Definitions and Previous Results

Let $\mathfrak{L}(E, F)$ be the set of all bounded linear operators T between the normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ (defined over the field of complex numbers), i.e., operators such that the image of the (closed) unit ball

$$U_E := \{x \in E\colon \|x\|_E \leq 1\} \tag{7}$$

is bounded. The smallest such bound is called the *operator norm*

$$\|T\| := \sup_{x \in U_E} \|Tx\|_F. \tag{8}$$

The nth *entropy number of a set* $M \subset E$, for $n \in \mathbb{N}$, is

$$\epsilon_n(M) := \inf\{\epsilon > 0\text{: there exists an } \epsilon\text{-cover for } M \text{ in the metric } \|\cdot\|_E \text{ containing } n \text{ or fewer points}\}. \tag{9}$$

In case of ambiguity, we will sometime write $\epsilon_n(M, \|\cdot\|_E)$ to explicitly indicate the metric that the covering number is taken with respect to. (The function $n \mapsto \epsilon_n(M)$ can be thought of as the functional inverse of the function $\epsilon \mapsto \mathcal{N}(\epsilon, M, d)$ where d is the metric induced by $\|\cdot\|_E$.) The *entropy numbers of an operator* $T \in \mathfrak{L}(E, F)$ are defined as

$$\epsilon_n(T) := \epsilon_n(T(U_E)). \tag{10}$$

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

In the following, k will always denote a kernel, and d and m will be the input dimensionality and the number of training examples, respectively, so that the training data is a sequence

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}. \tag{11}$$

Let $\log$ denote the logarithm to base 2.

Given a class of functions $\mathcal{F}$, the generalization performance attainable using $\mathcal{F}$ can be bounded in terms of the covering numbers of $\mathcal{F}$. More precisely, for some set $\mathcal{X}$, and $\boldsymbol{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$, define the *uniform ϵ-covering number* of the function class $\mathcal{F}$ on $\mathcal{X}$ as

$$\mathcal{N}^m(\epsilon, \mathcal{F}) := \sup_{X^m \in \mathcal{X}^m} \mathcal{N}\left(\epsilon, \mathcal{F}, \ell_\infty^{\boldsymbol{X}^m}\right) \tag{12}$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\boldsymbol{X}^m})$ is the ϵ-covering number of $\mathcal{F}$ with respect to $\ell_\infty^{\boldsymbol{X}^m}$. Many generalization error bounds can be expressed in terms of $\mathcal{N}^m(\epsilon, \mathcal{F})$ (see, for example, [2], [3], [12]).

Assume $\mathcal{X}$ is a measurable space, given some $1 \leq p < \infty$ and a function $f\colon \mathcal{X} \to \mathbb{R}$ we define

$$\|f\|_{L_p(\mathcal{X})} := \left(\int |f(x)|^p \, d(x)\right)^{1/p}$$

if the integral exists and

$$\|f\|_{L_\infty(\mathcal{X})} := \operatorname{ess\,sup}_{x \in \mathcal{X}} |f(x)|.$$

(See, e.g., [8] for the definition of the essential supremum.) For $1 \leq p \leq \infty$, we let

$$L_p(\mathcal{X}) := \{f\colon \mathcal{X} \to \mathbb{R}\colon \|f\|_{L_p(\mathcal{X})} < \infty\}.$$

We sometimes write $L_p = L_p(\mathcal{X})$.

Suppose $T\colon E \to E$ is a linear operator mapping a normed space E into itself. We say that $x \in E$ is an *eigenvector* of T if for some scalar λ, $Tx = \lambda x$. Such a λ is called the *eigenvalue* associated with x. When E is a function space (e.g., $E = L_2(\mathcal{X})$), the eigenvectors are, of course, functions, and are usually called *eigenfunctions*. Thus, ψ_n is an eigenfunction of $T\colon L_2(\mathcal{X}) \to L_2(\mathcal{X})$ if $T\psi_n = \lambda\psi_n$. In general, λ is complex, but in this paper all eigenvalues are real (because of the symmetry of the kernels used to induce the operators). The inner product in $L_2(\mathcal{X})$ is defined as $\langle f, g\rangle = \int_{\mathcal{X}} f(\tau)g(\tau)\, d\tau$.

We will make use of Mercer's theorem. The version stated below is a special case of the theorem proven in [9, p. 145].

Theorem 2 (Mercer): Suppose $k \in L_\infty(\mathcal{X} \times \mathcal{X})$ is a symmetric kernel (i.e., $k(x, x') = k(x', x)$) such that the integral operator $T_k\colon L_2(\mathcal{X}) \to L_2(\mathcal{X})$

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, \boldsymbol{y}) f(\boldsymbol{y})\, d\boldsymbol{y} \tag{13}$$

is positive (i.e., for all $f \in L_2(\mathcal{X})$, $\langle T_k(f), f\rangle \geq 0$; for symmetric k this is equivalent to $\lambda_i \geq 0$ for all i). Let $\psi_j \in L_2(\mathcal{X})$ be the eigenfunction of T_k associated with the eigenvalue $\lambda_j \neq 0$ and normalized such that $\|\psi_j\|_{L_2} = 1$ and let $\overline{\psi_j}$ denote its complex conjugate. Suppose ψ_j is continuous for all $j \in \mathbb{N}$. Then

1) $(\lambda_j(T))_j \in \ell_1$;
2) $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j \in \mathbb{N}} \lambda_j \psi_j(\boldsymbol{x}) \psi_j(\boldsymbol{y})$ holds for all $(\boldsymbol{x}, \boldsymbol{y})$, where the series converges absolutely and uniformly for almost all $(\boldsymbol{x}, \boldsymbol{y})$.

We will call a kernel satisfying the conditions of this theorem a *Mercer kernel*.

Note that in an early version of [15] we made use of an incorrect additional conclusion of Mercer's theorem to the effect that any Mercer kernel satisfies (4) (we propagated the error from [9, p. 145]). Steve Smale (private communication) has shown one can construct a counterexample to such a statement. In [10], a re-derivation of the main result of [15] is made without the need for (4) to hold. The idea is to replace λ_j by $l_j := \sup_{x \in \mathcal{X}} \lambda_j |\psi_j(x)|^2$ and proceed as in the original argument. The bottom line is that all of the results of the present paper then hold for kernels for which C_k may be infinite as long as l_j is finite for all j. For simplicity of presentation, we have assumed $C_k < \infty$ here and thus $l_j = C_k^2 \lambda_j$ and the results are explicitly stated in terms of C_k and λ_j. We are unaware of any kernel used in practice for which l_j is infinite.

In [15], an upper bound on the entropy numbers was given in terms of the eigenvalues of the kernel used. The result is in terms of the entropy numbers of a scaling operator A. The notation $(a_s)_s \in \ell_p$ denotes the sequence $(a_1, a_2, \ldots)$.

Theorem 3 (Entropy Numbers for $\Phi(\mathcal{X})$): Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a Mercer kernel satisfying (4). Choose $a_j > 0$ for $j \in \mathbb{N}$ such that $(\sqrt{\lambda_s}/a_s)_s \in \ell_2$, and define $A: \ell_2 \rightarrow \ell_2$ by

$$A(x_j)_j = (R_A a_j x_j)_j \tag{14}$$

with $R_A := C_k \|(\sqrt{\lambda_j}/a_j)_j\|_{\ell_2}$, where $(x_j)_j$ is a sequence in ℓ_2 and x_j a real number, respectively. Then

$$\epsilon_n(A) \leq \sup_{j \in \mathbb{N}} 6C_k \left\| \left(\sqrt{\lambda_s} \Big/ a_s \right)_s \right\|_{\ell_2} \left(\frac{a_1 \cdots a_j}{n} \right)^{\frac{1}{j}}. \tag{15}$$

This result leads to the following bounds for SV classes.

Theorem 4 (Bounds for SV Classes): Let k be a Mercer kernel satisfying (4). Then for all $n \in \mathbb{N}$

$$\epsilon_n(\mathcal{F}_{R_{\boldsymbol{w}}}, \ell_\infty^m) \leq R_{\boldsymbol{w}} \inf_{(a_s)_s:\, (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \epsilon_n(A) \tag{16}$$

where A is defined as in Theorem 3. Notice that while $\epsilon_n(\mathcal{F}_{R_{\boldsymbol{w}}})$ depends on m, its upper bound does not.

Combining (15) and (16) gives effective bounds on $\mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\boldsymbol{w}}})$ since

$$\epsilon_n\left(\mathcal{F}_{R_{\boldsymbol{w}}}, \ell_\infty^m\right) \leq \epsilon_0 \Rightarrow \mathcal{N}^m\left(\epsilon_0, \mathcal{F}_{R_{\boldsymbol{w}}}\right) \leq n.$$

These results thus give a method to obtain bounds on the entropy numbers for kernel machines. In (15) and (16), we can choose $(a_s)_s$ and j to optimize the bound. The key technical contribution of this paper is the explicit determination of the best choice of $(a_s)_s$ and j.

We assume henceforth that $(\lambda_s)_s$ is fixed and sorted in nonincreasing order, and $a_s > 0$ for all s. For $j \in \mathbb{N}$, we define the set

$$A_j = \left\{ (a_s)_s:\, \sup_{i \in \mathbb{N}} \left(\frac{a_1 \cdots a_i}{n} \right)^{\frac{1}{i}} = \left(\frac{a_1 \cdots a_j}{n} \right)^{\frac{1}{j}} \right\}. \tag{17}$$

In other words, A_j is the set of $(a_s)_s$ such that the

$$\sup_{i \in \mathbb{N}} \left(\frac{a_1 a_2 \cdots a_i}{n} \right)^{\frac{1}{i}}$$

is attained at $i = j$.

Let

$$B((a_s)_s, n, j) = \left\| \left(\sqrt{\lambda_s} \Big/ a_s \right)_s \right\|_{\ell_2} \left(\frac{a_1 \cdots a_j}{n} \right)^{\frac{1}{j}}. \tag{18}$$

III. The Optimal Choice of $(a_s)_s$ and j

Our aim in this section is to show that the infimum in (16) and the supremum in (15) can be achieved and to give explicit expressions for the sequence $(a_s)_s$ and number j^* that achieve them. The main technical theorem is as follows.

Theorem 5: Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a Mercer kernel satisfying (4). Suppose $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of T_k. For any $n \in \mathbb{N}$, the minimum

$$j^* = \min \left\{ j:\, \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2} \right)^{\frac{1}{j}} \right\} \tag{19}$$

always exists, and

$$\inf_{(a_s)_s:\, (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \leq B((a_s^*)_s, n, j^*)$$

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i}, & \text{when } i \leq j^* \\ \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n} \right)^{\frac{1}{j^*}}, & \text{when } i > j^*. \end{cases} \tag{20}$$

This choice of $(a_s)_s$ results in a simple form for the bound of (16) in terms of n and $(\lambda_i)_i$.

Corollary 6: Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a Mercer kernel satisfying (4) and let $A: \ell_2 \rightarrow \ell_2$ be given by (14). Then for any $n \in \mathbb{N}$, the entropy numbers satisfy

$$\inf_{(a_s)_s:\, (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \epsilon_n(A) \leq 6C_k \sqrt{ j^* \left(\frac{\lambda_1 \cdots \lambda_{j^*}}{n^2} \right)^{\frac{1}{j^*}} + \sum_{i=j^*+1}^{\infty} \lambda_i } \tag{21}$$

with

$$j^* = \min \left\{ j:\, \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2} \right)^{\frac{1}{j}} \right\}.$$

This corollary, together with (16), implies Theorem 1.

Proof Outline

The proof of Theorem 5 is quite long and is in Appendix A. It involves the following four steps.

1) We first prove that for all $n \in \mathbb{N}$

$$\hat{j} = \min \left\{ j:\, \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2} \right)^{\frac{1}{j}} \right\} \tag{22}$$

exists, whenever $(\lambda_i)_i$ are the eigenvalues of a Mercer kernel.

2) We then prove that for any $n \in \mathbb{N}$

$$\inf_{(a_s)_s:\, (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \leq \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j). \tag{23}$$

3) The next step is to prove that the choice of $(a_s)_s$ and j described by (19) and (20) are optimal. It is separated into two parts:

   a) for any $j_0 \leq j^*$, and any $(a_s)_s \in A_{j_0}$,

   $$B((a_s)_s, n, j_0) \geq B((a_s^*)_s, n, j^*)$$

   holds;

   b) for any $j_0 > j^*$, and any $(a_s)_s \in A_{j_0}$,

   $$B((a_s)_s, n, j_0) \geq B((a_s^*)_s, n, j^*)$$

   also holds.

4) Finally, we show that $(a_s^*)_s \in A_j$ and $(\sqrt{\lambda_s}/a_s^*)_s \in \ell_2$ when $(a_s^*)_s$ is chosen according to (20).

IV. Example

We illustrate the results of this paper with an example. Consider the kernel $k(x, y) = k(x - y)$ where $k(x) = e^{-x^2/\sigma^2}$. (Here, $d = 1$.) For such kernels (RBF kernels), $\|\Phi(\boldsymbol{x})\|_{\ell_2} = 1$ for all $\boldsymbol{x} \in \mathcal{X}$. Thus, by Mercer's theorem, the class (1) can be written as

$$\mathcal{F}_{R_{\boldsymbol{w}}} = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \tilde{\boldsymbol{x}}\rangle \colon \tilde{\boldsymbol{x}} \in \mathcal{S}, \, \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le 1, \, \|\boldsymbol{w}\|_{\ell_2} \le R_{\boldsymbol{w}}\}.$$

(See, for example, [6] for a more detailed explanation of this point—it is the fundamental basis of viewing SV machines in feature space.) One can use the fat-shattering dimension to bound the covering number of the class of functions $\mathcal{F}_{R_{\boldsymbol{w}}}$ (see, for example, [2]).

Theorem 7: With $\mathcal{F}_{R_{\boldsymbol{w}}}$ as above, if $m \ge 16R_{\boldsymbol{w}}^2/\epsilon^2 \ge 1$

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\boldsymbol{w}}}) \le 48 \left(\frac{R_{\boldsymbol{w}}}{\epsilon}\right)^2 \log^2 \left(\frac{4eR_{\boldsymbol{w}}m}{\epsilon}\right). \tag{24}$$

In order to determine the eigenvalues of T_k, we need to periodize the kernel. This periodization is necessary in order to get a discrete set of eigenvalues since $k(x)$ has infinite support (see [15] for further details). For our purposes, we can assume a fixed period $2\pi/\omega_0$ for some $\omega_0 > 0$. Since the kernel is translation-invariant, the eigenfunctions are $\psi_n(x) = \sqrt{2}\cos(n\omega_0 x)$ and so $C_k = \sqrt{2}$ [15]. The $\sqrt{2}$ factor comes from the requirement in Theorem 2 that $\|\psi_j\|_{\ell_2} = 1$. The eigenvalues can be computed and are

$$\lambda_j = \sqrt{2}\pi\sigma e^{-\frac{\omega_0^2}{4}\sigma^2 j^2}.$$

Setting $c_1 = \sqrt{2}\pi\sigma$, $c_2 = \frac{\omega_0^2}{4}\sigma^2$, the eigenvalues can be written as

$$\lambda_j = c_1 e^{-c_2 j^2}. \tag{25}$$

From (19), we know that

$$\lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}}$$

implies $j^* \le j$. But (25) shows that this condition on the eigenvalues is equivalent to

$$c_1 e^{-c_2(j+1)^2} < n^{-\frac{2}{j}} \left(c_1^j \exp\left(-c_2 \sum_{i=1}^{j} i^2\right)\right)^{\frac{1}{j}} \tag{26}$$

which is equivalent to

$$c_2(j+1)^2 > \frac{2}{j}\ln n + \frac{c_2}{6}(j+1)(2j+1)$$

$$\Leftrightarrow \tfrac{2}{3} c_2 (j+1) j \left(j + \tfrac{5}{4}\right) > 2\ln n$$

which follows from

$$j > \left(\frac{12\ln n}{\omega_0^2 \sigma^2}\right)^{1/3}.$$

Hence,

$$j^* \le \left\lfloor \left(\frac{12\ln n}{\omega_0^2\sigma^2}\right)^{1/3} \right\rfloor + 1. \tag{27}$$

We can now use (5) to give an upper bound on ϵ_n. Since the λ_i decay so rapidly the tail $\sum_{i=j^*+1}^{\infty} \lambda_i$ in (5) is dominated by the first term. We obtain the following bound:

$$\epsilon_n^2 = O\left(j^* n^{-\frac{2}{j^*}} c_1 \exp\left(-\frac{c_2}{6}(j^*+1)(2j^*+1)\right)\right).$$

Substituting (27) shows that

$$\log \epsilon_n = O\left(\log\log n + \log\sigma - (\sigma \log n)^{\frac{2}{3}}\right). \tag{28}$$

We can get several results from (28).

The relationship between ϵ_n and n. For fixed σ, (28) shows that

$$\log 1/\epsilon_n = \Omega\left(\log^{\frac{2}{3}} n\right)$$

which implies

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\boldsymbol{w}}}) = O\left(\log^{\frac{3}{2}}\left(\frac{1}{\epsilon}\right)\right) \tag{29}$$

which is considerably better than Theorem 7. Note that (29) does not depend on m. This is a consequence of using (16) which also has no dependence on m. One can obtain a dependence in m if instead of (16) one uses [15, eq. (49)]. As explained in [15], for moderate decay rates of $(\lambda_i)_i$ the bounds obtained are no better by doing so.

The relationship between ϵ_n and σ^2. Here, σ^2 is the variance of the Gaussian functions. When σ^2 increases, the kernel function will be wider, so the class $\mathcal{F}_{R_{\boldsymbol{w}}}$ should be simpler. In (28), we notice that if σ decreases, ϵ_n decreases for fixed n. Similarly, if σ increases, n decreases for fixed ϵ_n. Since the entropy numbers (and the covering numbers) indicate the capacity of the learning machine, the more complicated the machine, the bigger are the covering numbers for fixed ϵ_n. Specifically, we see from (28) that

$$\log 1/\epsilon_n = \Omega\left(\sigma^{\frac{2}{3}}\right)$$

and that

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\boldsymbol{w}}}) = O(1/\sigma). \tag{30}$$

Figs. 1 and 2 illustrate the bounds on the effective dimension j^* (for $\sigma^2 = 1$) as a function of n and ϵ, respectively.

V. Conclusion

We have presented a new formula for bounding the covering numbers of SV machines in terms of the eigenvalues of an integral operator induced by the kernel. We showed, by way of an example using a Gaussian kernel, that the new bound is easily computed and considerably better than previous results that did not take account of the kernel. We showed explicitly the effect of the choice of width of the kernel in this case.

The "effective dimension," j^*, can illustrate the character of kernel expansions clearly. For a smooth kernel, the "effective dimension" j^* is small. The value of j^* depends on n which in turn depends on ϵ. Thus, j^* can be considered analogous to existing "scale-sensitive" dimensions, such as the fat-shattering dimension. A key difference is that we now have bounds for j^* that explicitly depend on the kernel.

The bounds obtained apply to any dimension d. However, repeated eigenvalues become generic for isotropic translation invariant kernels. It is possible to obtain bounds that can be tighter

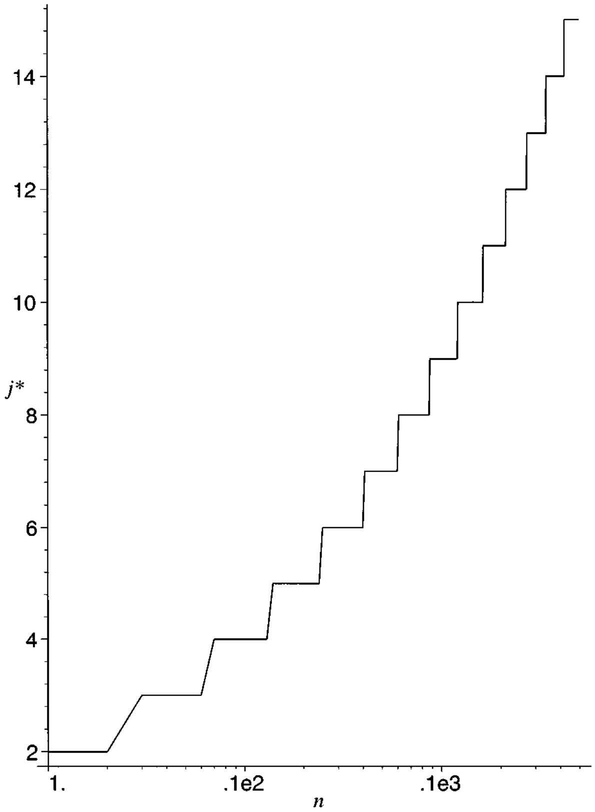

Fig. 1. j^* versus n for a Gaussian kernel.

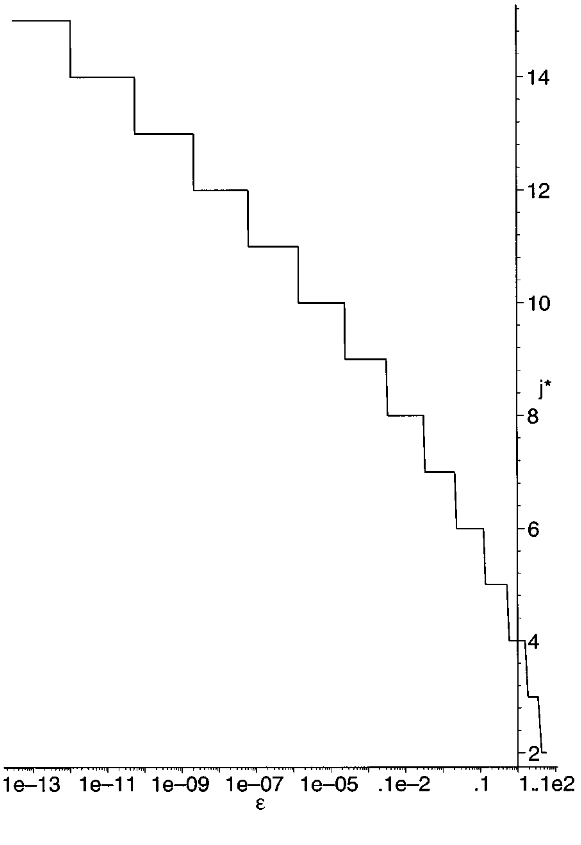

Fig. 2. j^* versus ϵ for a Gaussian kernel. Since j^* can be interpreted as an "effective dimension," this clearly illustrates why the bound on the covering numbers for Gaussian kernels grows so slowly as $\epsilon \downarrow 0$. Even when $\epsilon = 10^{-9}$, j^* is only 13.

in some cases, by using a slightly more refined argument; see [15].

Appendix A
Proof of Theorem 1

Step One

As indicated in Section III, we will first prove the existence of $\hat{j}$, which is defined in (22).

Lemma 8: Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ is a nonincreasing sequence of nonnegative numbers and $\lim_{j\to\infty} \lambda_j = 0$. Then, for all $n \in \mathbb{N}$, there exists $\hat{j} \in \mathbb{N}$ such that

$$\lambda_{\hat{j}+1} < \left(\frac{\lambda_1\cdots\lambda_{\hat{j}}}{n^2}\right)^{\frac{1}{\hat{j}}}. \tag{31}$$

Proof: Let $P_{\hat{j}} = \frac{\lambda_{\hat{j}+1}^{\hat{j}}}{\lambda_1\cdots\lambda_{\hat{j}}}$. Observe that (31) can be written as $P_{\hat{j}} < \frac{1}{n^2}$, and hence for all n there is a $\hat{j}$ such that (31) is true iff $\lim_{j\to\infty} P_j = 0$. But

$$P_{\hat{j}} = \frac{\lambda_{\hat{j}+1}^{\hat{j}}}{\lambda_1\cdots\lambda_{\hat{j}}} = \frac{\lambda_{\hat{j}+1}}{\lambda_1}\prod_{i=2}^{\hat{j}}\frac{\lambda_{\hat{j}+1}}{\lambda_i} \leq \frac{\lambda_{\hat{j}+1}}{\lambda_1}$$

since $(\lambda_i)_i$ is nonincreasing. Since $\lim_{j\to\infty} \lambda_j = 0$, we get $\lim_{j\to\infty} P_j = 0$. Thus, for any $n \in \mathbb{N}$ there is a $\hat{j}$ such that (31) is true. □

Corollary 9: Suppose k is a Mercer kernel and T_k the associated integral operator. If $\lambda_i = \lambda_i(T_k)$, then the minimum $\hat{j}$ from (22) always exists.

Proof: By Mercer's theorem, $(\lambda_i)_i \in \ell_1$ and so $\lim_{i\to\infty} \lambda_i = 0$. Lemma 8 can thus be applied. □

Step Two

Lemma 10: Suppose A_j and $B((a_s)_s, n, j)$ are defined as in (17) and (18), $(\sqrt{\lambda_s}/a_s^*)_s \in \ell_2$, j^* and $(a_s^*)_s \in A_{j^*}$ satisfy

$$B((a_s^*)_s, n, j^*) = \inf_{j\in\mathbb{N}} \inf_{(a_s)_s\in A_j} B((a_s)_s, n, j). \tag{32}$$

Then

$$\inf_{(a_s)_s:\left(\sqrt{\lambda_s}/a_s\right)_s\in\ell_2} \sup_{j\in\mathbb{N}} B((a_s)_s, n, j) \leq \inf_{j\in\mathbb{N}} \inf_{(a_s)_s\in A_j} B((a_s)_s, n, j). \tag{33}$$

Proof: Since $(\sqrt{\lambda_s}/a_s^*)_s \in \ell_2$

$$\inf_{(a_s)_s:\left(\sqrt{\lambda_s}/a_s\right)_s\in\ell_2} \sup_{j\in\mathbb{N}} B((a_s)_s, n, j) \leq \sup_{j\in\mathbb{N}} B((a_s^*)_s, n, j). \tag{34}$$

But $(a_s^*)_s \in A_{j^*}$, following the definition of A_j and equality (32) we get

$$\sup_{j\in\mathbb{N}} B((a_s^*)_s, n, j) = B((a_s^*)_s, n, j^*) = \inf_{j\in\mathbb{N}} \inf_{(a_s)_s\in A_j} B((a_s)_s, n, j). \quad □$$

In fact, we can show that inequality (33) is an equality. The proof is in Appendix B.

It is now easier to calculate the optimal bound of the entropy number using Lemma 10.

Step Three

In this step, we will prove that the choice of $(a_s^*)_s$ and j^* given in Theorem 5 are optimal. We will first prove a useful technical result.

Lemma 11: Suppose A_j and $(\lambda_i)_i$ are defined as above, $(a_s)_s \in A_{j_0}$. Then we have

$$\left(\sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\right)\left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty}\lambda_i \ge 0. \tag{35}$$

Proof: Since $(a_s)_s \in A_{j_0}$, the following inequality must be true for $k \in \mathbb{N}$:

$$\left(\frac{a_1\cdots a_{j_0}\cdots a_{j_0+k}}{n}\right)^{\frac{1}{j_0+k}} \le \left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{1}{j_0}} \tag{36}$$

which implies

$$\begin{aligned} &\left(\frac{a_1\cdots a_{j_0}\cdots a_{j_0+k}}{n}\right) \le \left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{j_0+k}{j_0}}\\ \Rightarrow&\\ &a_{j_0+1}\cdots a_{j_0+k} \le \left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{k}{j_0}}, \qquad \forall k \in \mathbb{N}. \end{aligned} \tag{37}$$

Set

$$\psi = \left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{1}{j_0}}.$$

Then (37) can be rewritten as

$$a_{j_0+1}\cdots a_{j_0+k} \le \psi^k, \qquad \forall k \in \mathbb{N}. \tag{38}$$

Hence, the left-hand side of (35) can be rewritten as

$$\sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\psi^2 - \sum_{i=j_0+1}^{\infty}\lambda_i = \psi^2 \sum_{i=j_0+1}^{\infty}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right). \tag{39}$$

From (38), we get $a_{j_0+1} \le \psi$, so

$$\frac{1}{a_{j_0+1}^2} - \frac{1}{\psi^2} \ge 0.$$

Suppose $\frac{1}{a_i^2} - \frac{1}{\psi^2} < 0$ for some $i \in \mathbb{N}$. We will separate the sum into several parts. Set

$$\left.\begin{aligned} k_0 &= j_0\\ k_m &= \max\Big\{n > l_m\colon \tfrac{1}{a_i^2} < \tfrac{1}{\psi^2},\\ &\qquad\qquad \forall i \in \{l_m+1,\ldots,n\}\Big\}\\ l_m &= \max\Big\{n > k_{m-1}\colon \tfrac{1}{a_i^2} \ge \tfrac{1}{\psi^2},\\ &\qquad\qquad \forall i \in \{k_{m-1}+1,\ldots,n\}\Big\} \end{aligned}\right\} \tag{40}$$

where we set k_m and l_m to ∞ if the $\max$ does not exist. Since $(\lambda_i)_i$ is a nonincreasing sequence, from (40) we know

$$\begin{aligned} &\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) \ge \lambda_{i+c}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\qquad\qquad \forall i \in \{k_{m-1}+1,\ldots,l_m\},\ c \in \mathbb{N}\\ &\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) > \lambda_{i-c}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\qquad\qquad \forall i \in \{l_m+1,\ldots,k_m\}, \quad \forall c \in \{1,\ldots,i-1\} \end{aligned}$$

for $m \in \mathbb{N}$. Hence, if l_m is finite,

$$\begin{aligned} &\sum_{i=k_{m-1}+1}^{k_m}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\quad \ge \lambda_{l_m}\sum_{i=k_{m-1}+1}^{l_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) + \lambda_{l_m}\sum_{i=l_m+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\quad = \lambda_{l_m}\sum_{i=k_{m-1}+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right). \end{aligned} \tag{41}$$

And if l_m is infinite, this inequality is clearly true. We will exploit the inequality of the arithmetic and geometric means

$$x_1 + x_2 + \cdots + x_m \ge m(x_1\cdots x_m)^{\frac{1}{m}}, \qquad \text{for } x_i > 0. \tag{42}$$

Now (42) implies that for any $k_0 + 1 \le j \le k_m$, we have

$$\sum_{i=k_0+1}^{j}\frac{1}{a_i^2} \ge (j-k_0)\left(\prod_{i=k_0+1}^{j}\frac{1}{a_i^2}\right)^{\frac{1}{j-k_0}} \tag{43}$$

which together with (38) gives

$$\sum_{i=k_0+1}^{j}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) = \sum_{i=k_0+1}^{j}\frac{1}{a_i^2} - \frac{j-k_0}{\psi^2} \ge 0. \tag{44}$$

Hence, for any k_m, finite or infinite,

$$\sum_{i=k_0+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) \ge 0. \tag{45}$$

Now, for all k_m, using (41) and (45) repeatedly, we get

$$\begin{aligned} &\sum_{i=k_0+1}^{k_m}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &= \sum_{i=k_0+1}^{k_1}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) + \cdots + \sum_{i=k_{m-1}+1}^{k_m}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\ge \lambda_{l_1}\sum_{i=k_0+1}^{k_1}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) + \cdots + \lambda_{l_m}\sum_{i=k_{m-1}+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\ge \lambda_{l_2}\sum_{i=k_0+1}^{k_2}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) + \cdots + \lambda_{l_m}\sum_{i=k_{m-1}+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\ge \cdots \ge \lambda_{l_m}\sum_{i=k_0+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) \ge 0 \end{aligned}$$

for all $m \in \mathbb{N}$. Hence,

$$\psi^2\sum_{i=j_0+1}^{\infty}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right) \ge 0. \tag{46}$$

Noticing (39), inequality (35) is true. □

Now, let us prove the main result.

Lemma 12: Let A_j and $B((a_s)_s, n, j)$ be defined as above. Then we have

$$B((a_s^*)_s, n, j^*) = \inf_{j_0 \in \mathbb{N}}\ \inf_{(a_s)_s \in A_{j_0}} B((a_s)_s, n, j_0), \tag{47}$$

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i}, & \text{when } i \le j^* \\ \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}}, & \text{when } i > j^* \end{cases} \tag{48}$$

$$j^* = \min\left\{ j\colon \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} \right\}. \tag{49}$$

Proof: The main idea is to compare $B^2((a_s)_s, n, j_0)$ with $B^2((a_s^*)_s, n, j^*)$ and show

$$B^2((a_s)_s, n, j_0) \ge B^2((a_s^*)_s, n, j^*)$$

for all $j_0 \in \mathbb{N}$ and any $(a_s)_s \in A_{j_0}$. From the definition of $B((a_s)_s, n, j)$, we know

$$B^2((a_s)_s, n, j_0) = \left(\sum_{i=1}^{\infty} \frac{\lambda_i}{a_i^2}\right)\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}}$$

and

$$B^2((a_s^*)_s, n, j^*) = j^*\left(\frac{\lambda_1 \cdots \lambda_{j^*}}{n^2}\right)^{\frac{1}{j^*}} + \sum_{i=j^*+1}^{\infty} \lambda_i.$$

For convenience, we set

$$\Lambda = \left(\frac{\lambda_1 \cdots \lambda_{j^*}}{n^2}\right)^{\frac{1}{j^*}}.$$

Hence,

$$\begin{aligned} &B^2((a_s)_s, n, j_0) - B^2((a_s^*)_s, n, j^*) \\ &\quad = \sum_{i=1}^{\infty} \frac{\lambda_i}{a_i^2}\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j^*\Lambda - \sum_{i=j^*+1}^{\infty} \lambda_i. \end{aligned} \tag{50}$$

Part a): For the Condition $j_0 \le j^*$: Rewrite (50)

$$\begin{aligned} &B^2((a_s)_s, n, j_0) - B^2((a_s^*)_s, n, j^*) \\ &= \sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2}\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} + \sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} \\ &\quad - \left(j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} + \sum_{i=j_0+1}^{\infty} \lambda_i \right) \\ &\quad - \left(j^*\Lambda + \sum_{i=j^*+1}^{\infty} \lambda_i - j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i \right) \\ &= \left\{ \sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2}\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} \right\} \\ &\quad + \left\{ \sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i \right\} \\ &\quad + \left\{ j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} + \sum_{i=j_0+1}^{\infty} \lambda_i \right. \\ &\qquad \left. - \left(j^*\Lambda + \sum_{i=j^*+1}^{\infty} \lambda_i \right) \right\} \\ &\doteq E_1 + E_2 + E_3. \end{aligned} \tag{51}$$

We will show $E_1 \ge 0$, $E_2 \ge 0$, and $E_3 \ge 0$.

To Prove $E_1 \ge 0$: Since $\lambda_i \ge 0$ and $a_i \ge 0$, we exploit the inequality of the arithmetic and geometric means (42) again. Hence

$$\begin{aligned} E_1 &\ge j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{a_1^2 \cdots a_{j_0}^2}\left(\frac{a_1^2 \cdots a_{j_0}^2}{n^2}\right)^{\frac{j_0}{j_0}}\right)^{\frac{1}{j_0}} - j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} \\ &= j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} - j_0\left(\frac{\lambda_1 \cdots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} = 0. \end{aligned} \tag{52}$$

To Prove $E_2 \ge 0$: Applying Lemma 11 shows $E_2 \ge 0$.

To Prove $E_3 \ge 0$: In order to prove $E_3 \ge 0$, let us define the function

$$g(j) = j\left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} + \sum_{i=j+1}^{\infty} \lambda_i. \tag{53}$$

We will show that $g(j)$ is a nonincreasing function of j, for $j \le j^*$. Set

$$\beta_j = \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}}, \qquad \beta_{j-1} = \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j-1}}$$

we have

$$\begin{aligned} g(j-1) - g(j) &= (j-1)\beta_{j-1} + \lambda_j - j\beta_j \\ &= (\lambda_j - \beta_{j-1}) - j\,(\beta_j - \beta_{j-1}). \end{aligned} \tag{54}$$

Noticing $\beta_{j-1}^{j-1}\lambda_j = \beta_j^j$, (54) can be modified to

$$\begin{aligned} &g(j-1) - g(j) \\ &\quad = \beta_{j-1}^{-(j-1)}\left((\beta_j^j - \beta_{j-1}^j) - j\beta_{j-1}^{j-1}(\beta_j - \beta_{j-1})\right). \end{aligned} \tag{55}$$

Since $j \le j^*$, following (49), we get

$$\lambda_j \ge \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j-1}}, \qquad \forall\, j \le j^*. \tag{56}$$

So

$$\begin{aligned} \beta_j &= \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j}} \lambda_j^{\frac{1}{j}} \\ &\ge \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j} + \frac{1}{j(j-1)}} \\ &= \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j-1}} = \beta_{j-1}. \end{aligned}$$

Making use of the formula

$$x^n - y^n = (x-y)\sum_{i=1}^{n} x^{n-i}y^{i-1} \tag{57}$$

we obtain

$$\begin{aligned} \beta_j^j - \beta_{j-1}^j &= (\beta_j - \beta_{j-1})\sum_{i=1}^{j} \beta_j^{j-i}\beta_{j-1}^{i-1} \\ &\ge j\,\beta_{j-1}^{j-1}(\beta_j - \beta_{j-1}). \end{aligned}$$

Together with $\beta_{j-1} > 0$ and (55), we obtain

$$(\lambda_j - \beta_{j-1}) - j(\beta_j - \beta_{j-1}) \ge 0.$$

Hence,

$$g(j-1) \geq g(j).$$

Since $j_0 \leq j^*$, we get

$$E_3 = g(j_0) - g(j^*) \geq 0. \tag{58}$$

Combining the above results, we get

$$B^2((a_s)_s, n, j_0) - B^2((a_s^*)_s, n, j^*) \geq 0, \qquad \forall\, j_0 \leq j^*. \tag{59}$$

Part b): For the Condition $j_0 > j^*$: Rewrite (50)

$$\begin{aligned} & B^2((a_s)_s, n, j_0) - B^2((a_s^*)_s, n, j^*) \\ &= \left(\sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2} + \sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\right) \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} \\ &\quad - j^*\Lambda - \sum_{i=j^*+1}^{j_0} \lambda_i - \sum_{i=j_0+1}^{\infty} \lambda_i \\ &= \left\{\sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j^*\Lambda - \sum_{i=j^*+1}^{j_0} \lambda_i\right\} \\ &\quad + \left\{\sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i\right\} \\ &= F_1 + F_2. \end{aligned} \tag{60}$$

We will show $F_1 \geq 0$ and $F_2 \geq 0$.

To Prove $F_1 \geq 0$: For convenience, we set

$$D_i = \left(\frac{a_1 \cdots a_i}{n}\right)^{\frac{2}{i}}.$$

F_1 can be rewritten as

$$\begin{aligned} & \left(\sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} + \sum_{i=j^*+1}^{j_0} \frac{\lambda_i}{a_i^2}\right) D_{j_0} - j^*\Lambda - \sum_{i=j^*+1}^{j_0} \lambda_i \\ &= \left\{D_{j_0} \sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} - j^*\Lambda\right\} \\ &\quad + \left\{\sum_{i=j^*+1}^{j_0} \lambda_i \left(\frac{D_{j_0}}{a_i^2} - 1\right)\right\} = P_1 + P_2. \end{aligned} \tag{61}$$

Let us consider P_1 at first

$$\begin{aligned} P_1 &= (D_{j_0} + D_{j^*} - D_{j^*}) \sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} - j^*\Lambda \\ &= D_{j^*} \sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} - j^*\Lambda + (D_{j_0} - D_{j^*}) \sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2}. \end{aligned}$$

Since $(\lambda_i/a_i^2) > 0$, using the inequality of the arithmetic and geometric mean (42) again, we get

$$\sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} \geq j^* \left(\frac{\lambda_1 \cdots \lambda_{j^*}}{a_1^2 \cdots a_{j^*}^2}\right)^{\frac{1}{j^*}} \frac{n^2}{n^2} = \frac{j^*}{D_{j^*}} \Lambda.$$

Since $(a_s)_s \in A_{j_0}$, we get $D_{j_0} \geq D_i$ for any $i \neq j_0$ and

$$\lambda_{j^*+1} < \left(\frac{\lambda_1 \cdots \lambda_{j^*}}{n^2}\right)^{\frac{1}{j^*}}$$

holds based on (49). Hence,

$$\begin{aligned} P_1 &\geq 0 + (D_{j_0} - D_{j^*}) \sum_{i=1}^{j^*} \frac{\lambda_i}{a_i^2} \\ &\geq (D_{j_0} - D_{j^*}) \frac{1}{D_{j^*}} j^*\Lambda \\ &> (D_{j_0} - D_{j^*}) \frac{1}{D_{j^*}} j^*\lambda_{j^*+1} \geq 0. \end{aligned} \tag{62}$$

Let us consider P_2 now. If $P_2 \geq 0$, then $F_1 \geq 0$.

So let us prove that $F_1 \geq 0$ is also true when $P_2 \leq 0$. Observing $a_i^2 = D_i^i / D_{i-1}^{i-1}$ and $D_{j_0} \geq D_i$ for any $i \neq j_0$, the last element of P_2

$$\lambda_{j_0} \left(\frac{D_{j_0}}{a_{j_0}^2} - 1\right) = \lambda_{j_0} \left(\left(\frac{D_{j_0-1}}{D_{j_0}}\right)^{j_0-1} - 1\right) \leq 0.$$

Using a similar method as before, suppose $\frac{D_{j_0}}{a_i^2} - 1 > 0$ for some $i \in (j^*, j_0)$. We separate P_2 into several parts. Set

$$\begin{aligned} k_0 &= j_0 + 1 \\ l_m &= \min\left\{n < k_m\colon \frac{D_{j_0}}{a_i^2} - 1 \leq 0,\right. \\ &\qquad\qquad \left.\forall\, i \in \{n, \ldots, k_m - 1\}\right\} \\ k_m &= \min\left\{n < l_{m-1}\colon \frac{D_{j_0}}{a_i^2} - 1 > 0,\right. \\ &\qquad\qquad \left.\forall\, i \in \{n, \ldots, l_{m-1} - 1\}\right\}. \end{aligned} \tag{63}$$

Since $(\lambda_i)_i$ is a nonincreasing sequence, from (63) we know

$$\begin{aligned} & \lambda_i \left(\frac{D_{j_0}}{a_i^2} - 1\right) \geq \lambda_{i+c} \left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &\qquad \forall\, i \in \{k_{m+1}, \ldots, l_m - 1\},\ c \in \mathbb{N} \\ & \lambda_i \left(\frac{D_{j_0}}{a_i^2} - 1\right) > \lambda_{i-c} \left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &\qquad \forall\, i \in \{l_m, \ldots, k_m - 1\},\ \forall\, c \in \{1, \ldots, i-1\}. \end{aligned} \tag{64}$$

Using (64), we have

$$\begin{aligned} & \sum_{i=k_{m+1}}^{k_m-1} \lambda_i \left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &\geq \lambda_{l_m} \sum_{i=k_{m+1}}^{l_m-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right) + \lambda_{l_m} \sum_{i=l_m}^{k_m-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &= \lambda_{l_m} \sum_{i=k_{m+1}}^{k_m-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right). \end{aligned} \tag{65}$$

Hence,

$$\begin{aligned} 0 \geq P_2 &= \sum_{i=j^*+1}^{j_0} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_i \\ &= \sum_{i=j^*+1}^{k_1-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_i + \sum_{i=k_1}^{k_0-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_i \\ &\geq \sum_{i=j^*+1}^{k_1-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_i + \lambda_{l_1}\sum_{i=k_1}^{k_0-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right). \end{aligned}$$

If

$$\lambda_{l_1}\sum_{i=k_1}^{k_0-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right) > 0$$

we get

$$P_2 > \sum_{i=j^*+1}^{k_1-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_i.$$

If

$$\lambda_{l_1}\sum_{i=k_1}^{k_0-1} \left(\frac{D_{j_0}}{a_i^2} - 1\right) \leq 0$$

we can use (64) and (65) repeatedly. Finally, using (42) and $a_i^2 = D_i^i / D_{i-1}^{i-1}$ again, we can get

$$\begin{aligned} 0 \geq P_2 &\geq \sum_{i=j^*+1}^{j^*+l} \left(\frac{D_{j_0}}{a_i^2} - 1\right)\lambda_{j^*+l} \\ &= \lambda_{j^*+l}\sum_{i=j^*+1}^{j^*+l} \left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &\geq \lambda_{j^*+l}\, l\left(\left(\frac{1}{a_{j^*+1}^2 \cdots a_{j^*+l}^2}\right)^{\frac{1}{l}} D_{j_0} - 1\right) \\ &= \lambda_{j^*+l}\, l\left(D_{j_0}\left(\frac{D_{j^*}^{j^*}}{D_{j^*+l}^{j^*+l}}\right)^{\frac{1}{l}} - 1\right) \\ &\geq \lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j^*+l}}\right)^{\frac{j^*}{l}} - 1\right) \\ &\geq \lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}} - 1\right), \qquad \text{with } l \in \{1, \ldots, k_0 - 1\}. \end{aligned} \tag{66}$$

Combining (62) and (66), we have

$$F_1 = P_1 + P_2 > \frac{j^*\lambda_{j^*+1}(D_{j_0} - D_{j^*})}{D_{j^*}} + \lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}} - 1\right). \tag{67}$$

In order to show $F_1 \geq 0$, we just need to show

$$\frac{j^*\lambda_{j^*+1}(D_{j_0} - D_{j^*})}{D_{j^*}} + \lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}} - 1\right) \geq 0. \tag{68}$$

When $D_{j^*} = D_{j_0}$

$$\frac{j^*\lambda_{j^*+1}(D_{j_0} - D_{j^*})}{D_{j^*}} + \lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}} - 1\right) = 0.$$

Inequality (68) holds.

When $D_{j_0} > D_{j^*}$, setting $\Phi_0 = D_{j_0}^{\frac{1}{l}}$ and $\Phi_* = D_{j^*}^{\frac{1}{l}}$, the inequality (68) can be rewritten as

$$\frac{1}{\Phi_*^l}\lambda_{j^*+1}j^*(\Phi_0^l - \Phi_*^l) \geq \frac{1}{\Phi_0^{j^*}}\lambda_{j^*+l}\, l(\Phi_0^{j^*} - \Phi_*^{j^*}).$$

Noticing

$$\lambda_{j^*+l}\, l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}} - 1\right) = \frac{\lambda_{j^*+l}\, l(\Phi_*^{j^*} - \Phi_0^{j^*})}{\Phi_0^{j^*}} \leq 0 \tag{69}$$

we only need to show

$$\frac{\lambda_{j^*+1}j^*\Phi_0^{j^*}(\Phi_0^l - \Phi_*^l)}{\lambda_{j^*+l}\, l\Phi_*^l(\Phi_0^{j^*} - \Phi_*^{j^*})} > 1. \tag{70}$$

Since $\lambda_{j^*+1} \geq \lambda_{j^*+l}$, the left-hand side of (70) becomes

$$\frac{\lambda_{j^*+1}j^*\Phi_0^{j^*}(\Phi_0^l - \Phi_*^l)}{\lambda_{j^*+l}\, l\Phi_*^l(\Phi_0^{j^*} - \Phi_*^{j^*})} > \frac{j^*\Phi_0^{j^*}(\Phi_0^l - \Phi_*^l)}{l\Phi_*^l(\Phi_0^{j^*} - \Phi_*^{j^*})}.$$

Making use of (57) again, we obtain

$$\begin{aligned} \frac{j^*\Phi_0^{j^*}(\Phi_0^l - \Phi_*^l)}{l\Phi_*^l(\Phi_0^{j^*} - \Phi_*^{j^*})} &= \frac{j^*\Phi_0^{j^*}(\Phi_0 - \Phi_*)\sum_{i=1}^{l}\Phi_0^{l-i}\Phi_*^{i-1}}{l\Phi_*^l(\Phi_0 - \Phi_*)\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{i-1}} \\ &= \frac{j^*\Phi_0^{j^*}\sum_{i=1}^{l}\Phi_0^{l-i}\Phi_*^{i-1}}{l\Phi_*^l\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{i-1}} \\ &= \frac{j^*\sum_{i=1}^{l}\Phi_0^{j^*+l-i}\Phi_*^{i-1}}{l\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{l+i-1}} \\ &= \frac{\sum_{k=1}^{j^*}\sum_{i=1}^{l}\Phi_0^{j^*+l-i}\Phi_*^{i-1}}{\sum_{k=1}^{l}\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{l+i-1}}. \end{aligned} \tag{71}$$

Observe the numerator and the denominator both have $j^* \times l$ elements represented as $\Phi_0^m\Phi_*^n$. But we know $\Phi_0 > \Phi_*$ since $D_{j_0} > D_{j^*}$, hence from (71), we obtain

$$\begin{aligned} \frac{\sum_{k=1}^{j^*}\sum_{i=1}^{l}\Phi_0^{j^*+l-i}\Phi_*^{i-1}}{\sum_{k=1}^{l}\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{l+i-1}} &> \frac{\sum_{k=1}^{j^*}\sum_{i=1}^{l}\Phi_0^{j^*}\Phi_*^{l-1}}{\sum_{k=1}^{l}\sum_{i=1}^{j^*}\Phi_0^{j^*-1}\Phi_*^{l}} \\ &= \frac{j^*l\Phi_0^{j^*}\Phi_*^{l-1}}{j^*l\Phi_0^{j^*-1}\Phi_*^{l}} = \frac{\Phi_0}{\Phi_*} > 1. \end{aligned}$$

So

$$\frac{j^*\Phi_0^{j^*}(\Phi_0^l - \Phi_*^l)}{l\Phi_*^l(\Phi_0^{j^*} - \Phi_*^{j^*})} \geq 1.$$

Hence,

$$F_1 = P_1 + P_2 > 0 \tag{72}$$

is proved for $j_0 = j^* + k$ with all $k \in \mathbb{N}$.

To Prove $F_2 \geq 0$: Using Lemma 11 again, we get

$$F_2 \geq 0. \tag{73}$$

Combining (72) and (73), we get

$$B^2((a_s)_s, n, j_0) - B^2((a_s^*)_s, n, j^*) \geq 0 \qquad \forall j_0 > j^*. \tag{74}$$

Combining (59) and (74), (47) is proved true. □

Step Four

We supposed that $(a_s^*)_s \in A_{j^*}$ in the above proof. Now let us show it. First, for $j > j^*$, from (20)

$$\begin{aligned}
&\left(\frac{a_1^* \cdots a_{j^*}^* \cdots a_j^*}{n}\right)^{\frac{1}{j}} \\
&= \left(\frac{a_1^* \cdots a_{j^*}^*}{n}\right)^{\frac{1}{j}} \left(a_{j^*+1}^* \cdots a_j^*\right)^{\frac{1}{j}} \\
&= \left(\frac{\sqrt{\lambda_1} \cdots \sqrt{\lambda_{j^*}}}{n}\right)^{\frac{1}{j}} \left(\left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^* j}}\right)^{j-j^*} \\
&= \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j} + \frac{j-j^*}{j^* j}} \\
&= \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}} = \left(\frac{a_1^* \cdots a_{j^*}^*}{n}\right)^{\frac{1}{j^*}}.
\end{aligned}$$

Second, for $j \leq j^*$. From (56), we get

$$\begin{aligned}
\left(\frac{a_1^* \cdots a_j^*}{n}\right)^{\frac{1}{j}} &= \left(\frac{\sqrt{\lambda_1 \cdots \lambda_j}}{n^2}\right)^{\frac{1}{j}} \geq \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j-1}}}{n^2}\right)^{\frac{1}{j-1}} \\
&= \left(\frac{a_1^* \cdots a_{j-1}^*}{n}\right)^{\frac{1}{j-1}}.
\end{aligned}$$

Thus, $(a_s^*)_s \in A_{j^*}$.

We can also show $(\sqrt{\lambda_s}/a_s^*)_s \in \ell_2$

$$\left(\sqrt{\lambda_s}/a_s^*\right)_s = \sqrt{\sum_{i=1}^{\infty} \frac{\lambda_i}{a_i^2}} = \sqrt{j^* + \frac{1}{\Lambda^2} \sum_{i=j^*+1}^{\infty} \lambda_i}. \tag{75}$$

When $k(x, y)$ and n are given, $(\lambda_i)_i$ and j^* are determined. So $\Lambda = n^{-\frac{2}{j^*}}(\lambda_1 \cdots \lambda_{j^*})^{\frac{1}{j^*}}$ is a constant. By Mercer's theorem, $(\lambda_i)_i \in \ell_1$ and thus $\sum_{i=j^*+1}^{\infty} \lambda_i$ is finite. So (75) is finite. Hence $(\sqrt{\lambda_s}/a_s^*)_s \in \ell_2$ is proved.

Conclusion

Following the proof above, we get the following corollary.

Corollary 13: Suppose A_j and $B((a_s)_s, n, j)$ are defined as in (17) and (18). Then we have

$$B((a_s^*(j^*)), n, j^*) = \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j) \tag{76}$$

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i}, & \text{when } i \leq j^* \\ \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}}, & \text{when } i > j^* \end{cases} \tag{77}$$

$$j^* = \min\left\{j \colon \lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}}\right\}. \tag{78}$$

Theorem 1 is then established.

Appendix B
Proof That (33) Cannot be Improved

Lemma 14: Suppose A_j and $B((a_s)_s, n, j)$ are defined as above. Let $j \in \mathbb{N}$ and $(a_s)_s \in A_j$. Suppose j^* and $(a_s^*)_s$ exist. Then

$$\begin{aligned}
&\inf_{(a_s)_s \colon (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \\
&\qquad = \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j).
\end{aligned} \tag{79}$$

Proof: Let us prove

$$\begin{aligned}
&\inf_{(a_s)_s \colon (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \\
&\qquad \geq \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j).
\end{aligned} \tag{80}$$

Choose an $(a_s^*)_s$ to realize the infimum on the left-hand side; then $(a_s^*)_s \in A_{j^*}$, where j^* is the j that realizes the inner supremum. Then

$$\begin{aligned}
&\inf_{(a_s)_s \colon (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \\
&= \sup_{j \in \mathbb{N}} B((a_s^*)_s, n, j) \\
&= B((a_s^*)_s, n, j^*) \geq \inf_{(a_s)_s \in A_{j^*}} B((a_s)_s, n, j^*) \\
&\geq \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j).
\end{aligned}$$

We have already proved

$$\begin{aligned}
&\inf_{(a_s)_s \colon (\sqrt{\lambda_s}/a_s)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s)_s, n, j) \\
&\qquad \leq \inf_{j \in \mathbb{N}} \inf_{(a_s)_s \in A_j} B((a_s)_s, n, j).
\end{aligned}$$

So, (79) is proved to be true. □

Acknowledgment

The authors wish to thank B. Schölkopf and A. J. Smola for useful discussions, R. Herbrich for some corrections, and the anonymous reviewers for useful suggestions.

References

[1] M. Anthony, “Probabilistic analysis of learning in artificial neural networks: The pac model and its variants,” *Neural Computing Surveys*, vol. 1, pp. 1–47, 1997. [Online]. Available: http://www.icsi.berkeley.edu/~jagota/NCS.

[2] M. Anthony and P. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[3] P. L. Bartlett, “The sample complexity of pattern classsification with neural networks: The size of the weights is more important than the size of the network,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 525–536, Mar. 1998.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *5th Annu. ACM Workshop on COLT*, D. Haussler, Ed. Pittsburgh, PA: ACM, 1992, pp. 144–152.

[5] C. Cortes and V. Vapnik, “Support vector networks,” *Machine Learning*, vol. 20, pp. 273–297, 1995.

[6] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[7] F. Girosi, “An equivalence between sparse approximation and support vector machines,” *Neural Comput.*, vol. 10, no. 6, pp. 1455–1480, 1998.

[8] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1970.

[9] H. König, *Eigenvalue Distribution of Compact Operators*. Basel, Switzerland: Birkhäuser, 1986.

[10] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[11] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, “Structural risk minimization over data-dependent hierarchies,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 1926–1940, Sept. 1998.

[12] J. Shawe-Taylor and N. Cristianini, “Margin distribution bounds on generalization,” Royal Holloway College, Univ. London, London, U.K., NeuroCOLT Tech. Rep. NC-TR-98-030, 1998.

[13] J. Shawe-Taylor and R. C. Williamson, “Generalization performance of classifiers in terms of observed covering numbers,” in *Proc. EUROCOLT’99*, 1999, pp. 274–284.

[14] A. J. Smola, B. Schölkopf, and K.-R. Müller, “The connection between regularization operators and support vector kernels,” *Neural Networks*, vol. 11, pp. 637–649, 1998.

[15] R. C. Williamson, A. J. Smola, and B. Schölkopf, “Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators,” Royal Holloway College, London, U.K., NeuroCOLT Tech. Rep. NC-TR-98-019, 1998. Submitted for publication to *IEEE Trans. Inform. Theory*.

[16] ——, “Entropy numbers, operators and support vector kernels,” in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 127–144.