Covering Numbers for Support Vector Machines

Ying Guo
Department of Engineering
Australian National University
Canberra 0200, Australia
guo@hilbert.anu.edu.au

Peter L. Bartlett
RSISE
Australian National University
Canberra 0200, Australia
Peter.Bartlett@anu.edu.au

John Shawe-Taylor
Department of Computer Science
Royal Holloway College,
University of London
Egham, TW20 0EX, UK
jst@dcs.rhbnc.ac.uk

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

Abstract

Support vector machines are a type of learning machine related to the maximum margin hyperplane. Until recently, the only bounds on the generalization performance of SV machines (within the PAC framework) were via bounds on the fat-shattering dimension of maximum margin hyperplanes. This result took no account of the kernel used. More recently, it has been shown [8] that one can bound the relevant covering numbers using some tools from functional analysis. The resulting bound is quite complex and seemingly difficult to compute. In this paper we show that the bound can be greatly simplified and as a consequence we are able to determine some interesting quantities (such as the effective number of dimensions used). The new bound is quite a simple formula involving the eigenvalues of the integral operator induced by the kernel. We present an explicit calculation of covering numbers for an SV machine using a Gaussian kernel which is significantly better than that implied by the maximum margin fat-shattering result.

1 INTRODUCTION

Support Vector (SV) Machines [5] are learning algorithms based on maximum margin hyperplanes [4] which make use of an implicit mapping into feature space by using a more 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 completely 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 can attain a small training error.

In [9, 8] it has been shown that this intuition is justified. The main result there (quoted below) gives a bound on the covering numbers for the class of functions computed with support vector machines. This bound along with statistical results of the form given in [7] result in bounds that do explicitly depend on the kernel used.

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 below). 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] for an overview).

In [9, 8] an alternative viewpoint is taken where the class $\mathcal{F}$ is viewed as being generated by an integral operator induced by the kernel. 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 (13) below).

The main technical result of this paper is an explicit reformulation of this bound which is amenable to direct calculation. We illustrate the new result by bounding the covering numbers of SV machines which use Gaussian RBF kernels. The result shows the influence of σ^2 on the covering numbers: the covering numbers will decrease when σ^2 increases. Here σ^2 is the variance of the Gaussian function used for the kernel. More generally, the main result makes model order selection possible using any parametrized family of kernel functions: we can describe precisely how the capacity of the class is affected by changes to the kernel.

For $d \in \mathbb{N}$, $\mathbb{R}^d$ denotes the d-dimensional space of vectors $\mathbf{x} = (x_1, \ldots, x_d)$. For $0 < p \leq \infty$, define the spaces

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

where the p-norms are

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

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

For $d = \infty$, we write $l_p = l_p^\infty$ and the norms are defined similarly (by formally subsituting ∞ for d in the above definitions).

The *ϵ-covering number of $\mathcal{F}$ with respect to the metric d* denoted $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the size of the smallest ϵ-cover for $\mathcal{F}$ using the metric d. Given m points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \ell_p^d$, we use the shorthand $\mathbf{X}^m = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Suppose $\mathcal{F}$ is a class of functions defined on $\mathbb{R}^d$. The ℓ_∞^d norm *with respect to* $\mathbf{X}^m$ of $f \in \mathcal{F}$ is defined as $\|f\|_{\ell_\infty^{\mathbf{X}^m}} := \max_{i=1,\ldots,m} |f(\mathbf{x}_i)|$. The input space is taken to be $\mathcal{X}$, a compact subset of $\mathbb{R}^d$.

Our main result is a bound for the covering number of SV machines. We only discuss the case when $d = 1$. (In fact the result does hold for general d; see the discussion in the conclusion).

Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel satisfying the hypotheses of Mercer's theorem (Theorem 2). Given m points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}$. Denote by $\mathcal{F}_{R_\mathbf{w}}$ the hypothesis class implemented by SV machines on an m-sample with weight vector (in feature space) bounded by $R_\mathbf{w}$:

$$\mathcal{F}_{R_\mathbf{w}} = \left\{ \mathbf{x} \mapsto \sum \alpha_i k(\mathbf{x}, \mathbf{x}_i) \colon \sum_i \sum_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \le R_\mathbf{w}^2 \right\}. \tag{1}$$

Let $\lambda_1 \ge \lambda_2 \ge \cdots$ be the eigenvalues of the integral operator

$$\begin{aligned} T_k\colon L_2(\mathcal{X}) &\to L_2(\mathcal{X}) \\ T_k\colon f &\mapsto \int_\mathcal{X} k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}. \end{aligned}$$

and denote by $\psi_n(\cdot)$, $n \in \mathbb{N}$ the corresponding eigenfunctions. (See the next section for a reminder of what this means.) For translation invariant kernels (such as $k(x, y) = \exp((x - y)^2/\sigma^2)$), the eigenvalues are given by

$$\lambda_i = \sqrt{2\pi} K(j\omega_0) \tag{2}$$

for $j \in \mathbb{Z}$, where $K(\omega) = F[k(x)](\omega)$ is the Fourier transform of $k(\cdot)$ (see [9, 8] for further details). For a smooth kernel, the Fourier transform $F(j\omega_0)$ decreases faster. (There are less "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. Hypothesis class $\mathcal{F}_{R_\mathbf{w}}$, eigenfunctions $\psi_n(\cdot)$ and eigenvalues (λ_i) are defined as above. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}$ be m data points. Let*

$$C_k = \sup_n \|\psi_n\|_{L_\infty}.$$

For $n \in \mathbb{N}$ set

$$\epsilon_n^* = 6 R_\mathbf{w} C_k \sqrt{j^* \left(\frac{\lambda_1 \ldots \lambda_{j^*}}{n^2}\right)^{\frac{1}{j^*}} + \sum_{i=j^*+1}^{\infty} \lambda_i}, \tag{3}$$

with

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

Then $C_k < \infty$ and

$$\sup_{\mathbf{X}^m \in \mathcal{X}^m} \mathcal{N}(\epsilon_n^*, \mathcal{F}_{R_\mathbf{w}}, \ell_\infty^{\mathbf{X}^m}) \le n.$$

The quantity ϵ_n^* is an upper bound on the *entropy number* of $\mathcal{F}_{R_\mathbf{w}}$, which is the functional inverse of the covering number. In this theorem, the number j^* has a natural interpretation: For a given value of n, it 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. It turns out that all kernels satisfying Mercer's conditions are sufficiently smooth for j^* to be finite.

The remainder of the paper is organized as follows. We start by introducing notation and definitions (Section 2). Section 3 contains the main result (the proof is in Appendix A). Section 4 contains an example application of the main result. Section 5 concludes.

2 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 $(F, \|\cdot\|_F)$, i.e. operators such that the image of the (closed) unit ball

$$U_E := \{x \in E\colon \|x\|_E \le 1\} \tag{4}$$

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

$$\|T\| := \sup_{x \in U_E} \|Tx\|_F. \tag{5}$$

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

$$\begin{aligned} \epsilon_n(M) \quad := \inf\{\epsilon > 0\colon &\text{ there exists an } \epsilon\text{-cover} \\ &\text{for } M \text{ in } E \text{ containing } n \text{ or fewer points}\}. \end{aligned} \tag{6}$$

(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

$$\epsilon_n(T) := \epsilon_n(T(U_E)). \tag{7}$$

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

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

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}. \tag{8}$$

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

We will map the input data into a feature space $\mathcal{S}$ via a mapping Φ. We let $\tilde{\mathbf{x}} := \Phi(\mathbf{x})$, and

$$\mathcal{F}_{R_\mathbf{w}} := \{\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle\colon \tilde{\mathbf{x}} \in \mathcal{S}, \|\mathbf{w}\| \le R_\mathbf{w}\} \subseteq \mathbb{R}^\mathcal{S}.$$

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 $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$, define the *ϵ-growth function* of the function class $\mathcal{F}$ on $\mathcal{X}$ as

$$\mathcal{N}^m(\epsilon, \mathcal{F}) := \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m}), \tag{9}$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m})$ is the ϵ-covering number of $\mathcal{F}$ with respect to $\ell_\infty^{\mathbf{X}^m}$. Many generalization error bounds can be expressed in terms of $\mathcal{N}^m(\epsilon, \mathcal{F})$.

Given some set $\mathcal{X}$, 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 $\|f\|_{L_\infty(\mathcal{X})} := \operatorname{ess\,sup}_{x \in \mathcal{X}} |f(x)|$. 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* 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 $T : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ 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).

We will make use of Mercer's theorem. The version stated below is a special case of the theorem proven in [6, p. 145].

Theorem 2 (Mercer) *Suppose $k \in L_\infty(\mathcal{X}^2)$ is a symmetric kernel such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,*

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \tag{10}$$

is positive. 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. $\psi_j \in L_\infty(\mathcal{X})$ *and* $\sup_j \|\psi_j\|_{L_\infty} < \infty$.
3. $k(\mathbf{x}, \mathbf{y}) = \sum_{j \in \mathbb{N}} \lambda_j \psi_j(\mathbf{x}) \overline{\psi_j(\mathbf{y})}$ *holds for all* $(\mathbf{x}, \mathbf{y})$, *where the series converges absolutely and uniformly for all* $(\mathbf{x}, \mathbf{y})$.

We will call a kernel satisfying the conditions of this theorem a *Mercer kernel*. From statement 2 of Mercer's theorem there exists some constant $C_k \in \mathbb{R}^+$ depending on $k(\cdot, \cdot)$ such that

$$|\psi_j(\mathbf{x})| \leq C_k \text{ for all } j \in \mathbb{N} \text{ and } \mathbf{x} \in \mathcal{X}. \tag{11}$$

This conclusion is the only reason we have added the condition that ψ_n is continuous; it is not necessary for the theorem as stated, but it is convenient to bundle all of our assumptions into the one place. In any case it is not a very restrictive assumption: if $\mathcal{X}$ is compact and k is continuous, then ψ_j is automatically continuous (see e.g. [3]). Alternatively, if k is translation invariant, then ψ_j are scaled cosine functions and thus continuous.

In [8] 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 $(a_s)_s \in l_p$ donates the sequence $(a_1, a_2, \cdots)$ such that $\sum_{s=0}^{\infty} |a_s| < \infty$.

Theorem 3 (Entropy numbers for $\Phi(\mathcal{X})$) *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. Choose $a_j > 0$ for $j \in \mathbb{N}$ such that $\left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2$, and define*

$$A\colon (x_j)_j \mapsto (R_A a_j x_j)_j \tag{12}$$

with $R_A := C_k \|(\sqrt{\lambda_j}/a_j)_j\|_{\ell_2}$. Then

$$\epsilon_n(A\colon \ell_2 \to \ell_2) \leq \sup_{j \in \mathbb{N}} 6 C_k \left\| \left(\sqrt{\lambda_s}/a_s\right)_s \right\|_{\ell_2} \left(\frac{a_1 \cdots a_j}{n}\right)^{\frac{1}{j}}. \tag{13}$$

This result leads to the following bounds for SV classes.

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

$$\epsilon_n(\mathcal{F}_{R_{\mathbf{w}}}) \leq R_{\mathbf{w}} \inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \epsilon_n(A), \tag{14}$$

where A is defined as in Theorem 3.

Combining Equations (13) and (14) gives effective bounds on $\mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\mathbf{w}}})$ since

$$\epsilon_n(T\colon \ell_2 \to \ell_\infty^m) \leq \epsilon_0 \;\Rightarrow\; \mathcal{N}^m(\epsilon_0, \mathcal{F}_{R_{\mathbf{w}}}) \leq n.$$

These results thus give a method to obtain bounds on the entropy numbers for kernel machines. In Inequality (14), we can choose $(a_s)_s$ to optimize the bound. The key technical contribution of this paper is the explicit determination of the best choice of $(a_s)_s$.

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

$$A_j = \left\{ (a_s)_s\colon \quad \sup_{i \in \mathbb{N}} \left(\frac{a_1 \cdots a_i}{n}\right)^{\frac{1}{i}} = \left(\frac{a_1 \cdots a_j}{n}\right)^{\frac{1}{j}} \right\}. \tag{15}$$

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

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

where for notational simplicity, we write (a_s) for $(a_s)_s$.

3 THE OPTIMAL CHOICE OF $(a_s)_s$ AND j

Our main aim in this section is to show that the infimum in (14) and the supremum in (13) can be achieved and to give an explicit recipe for the sequence (a_s) and number j^* that achieve them. The main technical theorem is as follows.

Theorem 5 *Let $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a Mercer kernel. Suppose $\lambda_1,\lambda_2,\cdots$ are the eigenvalues of T_k. For any $n\in\mathbb{N}$, the minimum*

$$j^* = \min\left\{ j\colon \quad \lambda_{j+1} < \left(\frac{\lambda_1\ldots\lambda_j}{n^2}\right)^{\frac{1}{j}} \right\} \tag{16}$$

always exists, and

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

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i} & \text{when } i\le j^* \\ \left(\frac{\sqrt{\lambda_1\cdots\lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}} & \text{when } i>j^*. \end{cases} \tag{17}$$

This choice of (a_s) results in a simple form for the bound of (14) in terms of n and (λ_i):

Corollary 6 *Let $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a Mercer kernel and let A be given by (12). Then for any $n\in\mathbb{N}$, the entropy numbers satisfy*

$$\begin{aligned} &\inf_{(a_s)_s:\left(\sqrt{\lambda_s}/a_s\right)_s\in\ell_2} \epsilon_n(A\colon \ell_2\to\ell_2) \\ &\quad\le\quad 6C_k\sqrt{j^*\left(\frac{\lambda_1\ldots\lambda_{j^*}}{n^2}\right)^{\frac{1}{j^*}} + \sum_{i=j^*+1}^{\infty}\lambda_i}, \end{aligned} \tag{18}$$

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

This corollary, together with (14), 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 $n\in\mathbb{N}$,
$$\hat{j} = \min\left\{ j\colon \quad \lambda_{j+1} < \left(\frac{\lambda_1\ldots\lambda_j}{n^2}\right)^{\frac{1}{j}} \right\} \tag{19}$$
exists, whenever (λ_i) are the eigenvalues of a Mercer kernel.

2. We then prove that for any $n\in\mathbb{N}$
$$\begin{aligned} &\inf_{(a_s)_s:\left(\sqrt{\lambda_s}/a_s\right)_s\in\ell_2} \sup_{j\in\mathbb{N}} B((a_s),n,j) \\ &\quad\le\quad \inf_{j\in\mathbb{N}} \inf_{(a_s)\in A_j} B((a_s),n,j). \end{aligned} \tag{20}$$

3. The next step is to prove that the choice of (a_s) and j described by (16) and (17) are optimal. It is separated into two parts:
   (a) For any $j_0\le j^*$, and any $(a_s)\in A_{j_0}$,
   $$B((a_s),n,j_0) \ge B((a_s^*),n,j^*)$$
   holds.
   (b) For any $j_0 > j^*$, and any $(a_s)\in A_{j_0}$,
   $$B((a_s),n,j_0) \ge B((a_s^*),n,j^*)$$
   also holds.

4. Finally we show that $(a_s^*)\in A_j$ and $\left(\sqrt{\lambda_s}/a_s^*\right)_s\in\ell_2$ when (a_s^*) is chosen according to (17).

4 EXAMPLE

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

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

One can use the fat-shattering dimension to bound the covering number of the class of functions $\mathcal{F}_{R_{\mathbf{w}}}$ (see [2]).

Lemma 7 *With $\mathcal{F}_{R_{\mathbf{w}}}$ as above,*

$$\text{fat}_{\mathcal{F}_{R_{\mathbf{w}}}}(\epsilon) \le \left(\frac{R_{\mathbf{w}}}{\epsilon}\right)^2. \tag{21}$$

Theorem 8 *If $\mathcal{F}$ is class of functions mapping from a set $\mathcal{X}$ into the interval $[0,B]$, then for any ϵ, if $m\ge \text{fat}_{\mathcal{F}}(\epsilon/4)\ge 1$,*

$$\log\mathcal{N}^m(\epsilon,\mathcal{F}) \le 3\text{fat}_{\mathcal{F}}(\epsilon/4)\log^2\left(\frac{4eBm}{\epsilon}\right). \tag{22}$$

Combining these results we have the following bound with which we shall compare our new bound.

$$\log\mathcal{N}^m(\epsilon,\mathcal{F}_{R_{\mathbf{w}}}) \le 48\left(\frac{R_{\mathbf{w}}}{\epsilon}\right)^2\log^2\left(\frac{4eBm}{\epsilon}\right). \tag{23}$$

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 [9] for further details). For the purpose of the present paper, 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 $C_k=\sqrt{2}$. The $\sqrt{2}$ comes from the requirement in Theorem 2 that $\|\psi_j\|_{\ell_2}=1$. The eigenvalues 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

$$\lambda_j = c_1 e^{-c_2 j^2}. \tag{24}$$

From (16), we know that $\lambda_{j+1} < \left(\frac{\lambda_1\cdots\lambda_j}{n^2}\right)^{\frac{1}{j}}$ implies $j^*\le j$. But (24) shows that this condition on the eigenvalues is equivalent to

$$c_1 e^{-c_2(j+1)^2} < n^{-\frac{2}{j}}\left(c_1^j e^{-c_2\sum_{i=1}^j i^2}\right)^{\frac{1}{j}}, \tag{25}$$

which is equivalent to

$$\begin{aligned} c_2(j+1)^2 &> \frac{2}{j}\ln n + \frac{c_2}{6}(j+1)(2j+1) \\ \Leftrightarrow\quad \frac{2}{3}c_2(j+1)j\left(j+\frac{5}{4}\right) &> 2\ln n, \end{aligned}$$

which follows from

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

Hence,

$$j^* \leq \left\lfloor \left(\frac{12 \ln n}{\omega_0^2 \sigma^2} \right)^{1/3} \right\rfloor + 1. \qquad (26)$$

We can now use (18) to give an upper bound on ϵ_n. The tail $\sum_{i=j^*+1}^{\infty} \lambda_i$ in (18) is dominated by the first term, hence we obtain the following bound.

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

Substituting (26) shows that

$$\log \epsilon_n = O\left(\log\log n + \log\sigma - (\sigma \log n)^{\frac{2}{3}} \right) \qquad (27)$$

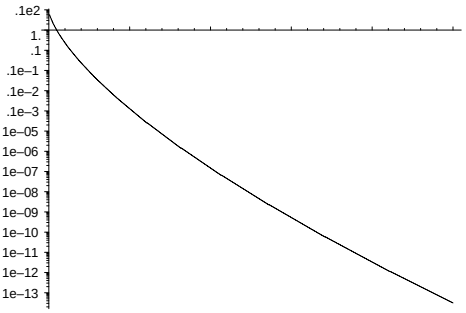

Figure 1: ϵ_n versus n for a Gaussian kernel as given by Corollary 6.

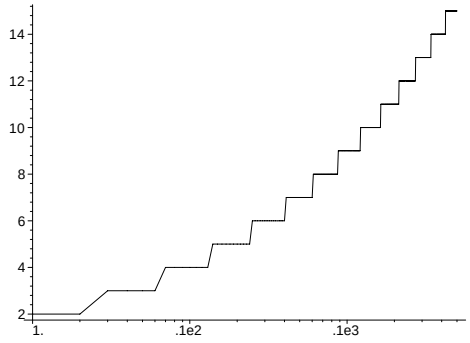

Figure 2: j^* versus n for a Gaussian kernel.

We can get several results from Equation (27).

The relationship between ϵ_n and n. For fixed σ, (27) shows that

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

which implies

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\mathbf{w}}}) = O\left(\log^{\frac{3}{2}} \left(\frac{1}{\epsilon} \right) \right), \qquad (28)$$

which is considerably better than (23). This can also be seen in Figure 1.

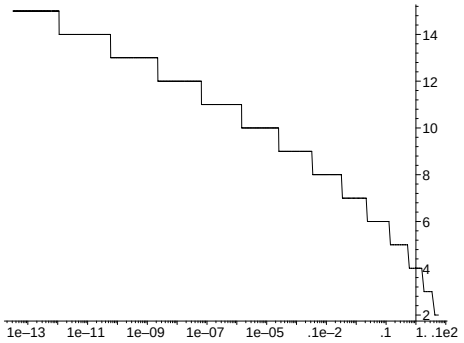

Figure 3: j^* versus ϵ for a Gaussian kernel. Since j^* can be interpreted as an "effective number of dimensions", this clearly illustrates why the bound on the covering numbers for Gaussian kernels grows so slowly as $\epsilon \downarrow 0$. Even when $\epsilon = 10^{-9}$, j^* is only 13.

The relationship between σ^2 and ϵ_n. Here, σ^2 is the variance of the Gaussian functions. When σ^2 increases, the kernel function will be wider, so the class $\mathcal{F}_{R_{\mathbf{w}}}$ should be simpler. In Equation (27), 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 number (and the covering number) indicates the capacity of the learning machine, the more complicated the machine is, the bigger the covering number for fixed ϵ_n. Specifically we see from Equation (27) that

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

and that

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

Figures (1) to (3) illustrate our bounds (for $\sigma^2 = 1$).

5 CONCLUSIONS

We have presented a new formula for bounding the covering numbers of support vector 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 number of dimensions", j^*, can illustrate the characters of the kernel functions clearly. For a smooth kernel, the "effective number of the dimensions" 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-shatterring dimension. A key difference is that we now have bounds for j^* that explicitly depend on the kernel.

We have discussed the result for the situation where the input dimension is 1. The main complication arising when $d > 1$ is that repeated eigenvalues become generic for isotropic translation invariant kernels. This does not break the bounds as stated (as long as one properly counts the multiplicity of eigenvalues). However, it is possible to obtain bounds that can be tighter in some cases, by using a slightly more refined argument [9].

References

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

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

[3] Robert Ash. *Information Theory*. Interscience Publishers, New York, 1965.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

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

[6] H. König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

[7] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[8] R. Williamson, A. Smola, and B. Schölkopf. Entropy numbers, operators and support vector kernels. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 127–144, Cambridge, MA, 1999. MIT Press.

[9] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. NeuroCOLT NC-TR-98-019, Royal Holloway College, 1998.

A PROOF OF THEOREM 1

STEP ONE

As indicated above, we will first prove the existence of $\hat{j}$, which is defined in (19).

Lemma 9 *Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ is a non-increasing sequence of non-negative 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*

$$\lambda_{\hat{j}+1} < \left(\frac{\lambda_1 \cdots \lambda_{\hat{j}}}{n^2}\right)^{\frac{1}{\hat{j}}}. \tag{29}$$

Proof Let $P_{\hat{j}} = \frac{\lambda_{\hat{j}+1}^{\hat{j}}}{\lambda_1 \cdots \lambda_{\hat{j}}}$. Observe that (29) can be written as $P_{\hat{j}} < \frac{1}{n^2}$, and hence for all n there is a $\hat{j}$ such that (29) 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 (λ_i) is non-increasing. 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 (29) is true. ∎

Corollary 10 *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 (19) always exists.*

Proof By Mercer's Theorem, $(\lambda_i) \in \ell_1$ and so $\lim_{i\to\infty} \lambda_i = 0$. Lemma 9 can thus be applied. ∎

STEP TWO

Lemma 11 *Suppose A_j and $B((a_s), n, j)$ are defined as above, $\left(\sqrt{\lambda_s}/a_s^*\right)_s \in \ell_2$, j^* and $(a_s^*) \in A_{j^*}$ satisfy*

$$B((a_s^*), n, j^*) = \inf_{j\in\mathbb{N}} \inf_{(a_s)\in A_j} B((a_s), n, j). \tag{30}$$

Then

$$\inf_{(a_s):\left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j\in\mathbb{N}} B((a_s), n, j) \leq \inf_{j\in\mathbb{N}} \inf_{(a_s)\in A_j} B((a_s), n, j). \tag{31}$$

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

$$\inf_{(a_s)_s:\left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j\in\mathbb{N}} B((a_s), n, j) \leq \sup_{j\in\mathbb{N}} B((a_s^*), n, j). \tag{32}$$

But $(a_s^*) \in A_{j^*}$, following the definition of A_j and equality (30) we get

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

∎

In fact, we can show that the inequality in (31) is in fact an equality. The proof is in appendix B.

It is now easier to calculate the optimal bound of the entropy number using Lemma 11.

STEP THREE

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

Lemma 12 *Suppose A_j and (λ_i) are defined as above, $(a_s) \in A_{j_0}$. Then we have*

$$\left(\sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2}\right) \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i \geq 0. \tag{33}$$

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

$$\left(\frac{a_1 \ldots a_{j_0} \ldots a_{j_0+k}}{n}\right)^{\frac{1}{j_0+k}} \leq \left(\frac{a_1 \ldots a_{j_0}}{n}\right)^{\frac{1}{j_0}}. \tag{34}$$

which implies

$$\left(\frac{a_1 \ldots a_{j_0} \ldots a_{j_0+k}}{n}\right) \leq \left(\frac{a_1 \ldots a_{j_0}}{n}\right)^{\frac{j_0+k}{j_0}}$$

$\Rightarrow$

$$a_{j_0+1}\dots a_{j_0+k} \le \Big(\frac{a_1\dots a_{j_0}}{n}\Big)^{\frac{k}{j_0}},\ \ \forall k\in\mathbb{N}. \tag{35}$$

Set

$$\psi = \Big(\frac{a_1\dots a_{j_0}}{n}\Big)^{\frac{1}{j_0}}.$$

Then (35) can be rewritten as:

$$a_{j_0+1}\dots a_{j_0+k} \le \psi^k,\ \ \forall k\in\mathbb{N}. \tag{36}$$

Hence, the left hand side of (33) can be rewritten as

$$\begin{aligned}&\sum_{i=j_0+1}^{\infty}\frac{\lambda_i}{a_i^2}\psi^2 - \sum_{i=j_0+1}^{\infty}\lambda_i\\ &= \psi^2\sum_{i=j_0+1}^{\infty}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right).\end{aligned} \tag{37}$$

From (36), 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

$$\begin{aligned}k_0 &= j_0,\\ k_m &= \max\{n>l_m: \ \tfrac{1}{a_i^2}<\tfrac{1}{\psi^2},\\ &\qquad \forall i\in\{l_m+1,\cdots,n\}\},\\ l_m &= \max\{n>k_{m-1}: \ \tfrac{1}{a_i^2}\ge\tfrac{1}{\psi^2},\\ &\qquad \forall i\in\{k_{m-1}+1,\cdots,n\}\},\end{aligned} \tag{38}$$

where we set k_m and l_m to ∞ if the max does not exist. Since (λ_i) is a non-increasing sequence, from (38) we know

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

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

$$\begin{aligned}&\sum_{i=k_{m-1}+1}^{k_m}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\ge \lambda_{l_m}\sum_{i=k_{m-1}+1}^{l_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &\quad +\lambda_{l_m}\sum_{i=l_m+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\\ &= \lambda_{l_m}\sum_{i=k_{m-1}+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right).\end{aligned} \tag{39}$$

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

$$x_1+x_2+\ldots+x_m\ge m(x_1\cdots x_m)^{\frac{1}{m}} \text{ for } x_i>0. \tag{40}$$

Now (40) implies that for any $k_0+1\le j\le k_m$, we have

$$\sum_{i=k_0+1}^{j}\frac{1}{a_i^2}\ge(j-k_0)\left(\prod_{i=k_0+1}^{j}\frac{1}{a_i^2}\right)^{\frac{1}{j-k_0}}, \tag{41}$$

which together with (36) gives

$$\sum_{i=k_0+1}^{j}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)=\sum_{i=k_0+1}^{j}\frac{1}{a_i^2}-\frac{j-k_0}{\psi^2}\ge 0. \tag{42}$$

Hence, for any k_m, finite or infinite,

$$\sum_{i=k_0+1}^{k_m}\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\ge 0. \tag{43}$$

Now, for all k_m, using (39) and (43) repeatedly, we get

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

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

$$\psi^2\sum_{i=j_0+1}^{\infty}\lambda_i\left(\frac{1}{a_i^2}-\frac{1}{\psi^2}\right)\ge 0. \tag{44}$$

Noticing (37), inequality (33) is true. ∎

Now, let us prove the main result.

Lemma 13 *Let A_j and $B((a_s),n,j)$ be defined as above. Then we have*

$$B((a_s^*),n,j^*)=\inf_{j_0\in\mathbb{N}}\inf_{(a_s)\in A_{j_0}}B((a_s),n,j_0), \tag{45}$$

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i} & \text{when } i \le j^* \\ \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}} & \text{when } i > j^*, \end{cases} \quad (46)$$

$$j^* = \min\left\{ j: \quad \lambda_{j+1} < \left(\frac{\lambda_1 \ldots \lambda_j}{n^2}\right)^{\frac{1}{j}} \right\}. \quad (47)$$

Proof The main idea is to compare $B^2((a_s), n, j_0)$ with $B^2((a_s^*), n, j^*)$ and show $B^2((a_s), n, j_0) \ge B^2((a_s^*), n, j^*)$ for all $j_0 \in \mathbb{N}$ and any $(a_s) \in A_{j_0}$. From the definition of $B((a_s), n, j)$, we know

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

and

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

For convenience, we set

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

Hence,

$$\begin{aligned} & B^2((a_s), n, j_0) - B^2((a_s^*), n, j^*) \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j^*\Lambda - \sum_{i=j^*+1}^{\infty} \lambda_i. \end{aligned} \quad (48)$$

Part a: For the condition $j_0 \le j^*$.

Rewrite (48):

$$\begin{aligned} & B^2((a_s), n, j_0) - B^2((a_s^*), n, j^*) \\ &= \sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} + \sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} \\ &\quad - \left(j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} + \sum_{i=j_0+1}^{\infty} \lambda_i \right) \\ &\quad - \left(j^*\Lambda + \sum_{i=j^*+1}^{\infty} \lambda_i - j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i \right) \\ &= \left\{ \sum_{i=1}^{j_0} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} \right\} \\ &\quad + \left\{ \sum_{i=j_0+1}^{\infty} \frac{\lambda_i}{a_i^2} \left(\frac{a_1 \cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty} \lambda_i \right\} \\ &\quad + \left\{ j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} + \sum_{i=j_0+1}^{\infty} \lambda_i \right. \\ &\quad \left. - \left(j^*\Lambda + \sum_{i=j^*+1}^{\infty} \lambda_i \right) \right\} \\ &= E_1 + E_2 + E_3. \end{aligned} \quad (49)$$

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 (40) again. Hence

$$\begin{aligned} E_1 &\ge j_0 \left(\frac{\lambda_1 \cdots \lambda_{j_0}}{a_1^2 \cdots a_{j_0}^2} \left(\frac{a_1^2 \cdots a_{j_0}^2}{n^2}\right)^{\frac{j_0}{j_0}} \right)^{\frac{1}{j_0}} \\ &\quad - j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} \\ &= j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} - j_0 \left(\frac{\lambda_1 \ldots \lambda_{j_0}}{n^2}\right)^{\frac{1}{j_0}} \\ &= 0. \end{aligned} \quad (50)$$

To prove $E_2 \ge 0$.

Applying Lemma 12 shows $E_2 \ge 0$.

To prove $E_3 \ge 0$.

In order to prove $E_3 \ge 0$, let us define function

$$g(j) = j \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} + \sum_{i=j+1}^{\infty} \lambda_i. \quad (51)$$

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

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

we have

$$\begin{aligned} & g(j-1) - g(j) \\ &= (j-1)\beta_{j-1} + \lambda_j - j\beta_j \\ &= (\lambda_j - \beta_{j-1}) - j\left(\beta_j - \beta_{j-1}\right). \end{aligned} \quad (52)$$

Noticing $\beta_{j-1}^{j-1}\lambda_j = \beta_j^j$, (52) can be modified to

$$\begin{aligned} & g(j-1) - g(j) \\ &= \beta_{j-1}^{-(j-1)} \left((\beta_j^j - \beta_{j-1}^j) - j\beta_{j-1}^{j-1}(\beta_j - \beta_{j-1}) \right) \end{aligned} \quad (53)$$

Since $j \le j^*$, following (47), we get

$$\lambda_j \ge \left(\frac{\lambda_1 \cdots \lambda_{j-1}}{n^2}\right)^{\frac{1}{j-1}} \quad \forall j \le j^*. \quad (54)$$

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

$$x^n - y^n = (x-y) \sum_{i=1}^{n} x^{n-i} y^{i-1}, \quad (55)$$

we obtain

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

Together with $\beta_{j-1} > 0$ and (53), we obtain

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

Hence,

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

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

$$E_3 = g(j_0) - g(j^*) \geq 0. \tag{56}$$

Combining the above results, we get

$$B^2((a_s), n, j_0) - B^2((a_s^*), n, j^*) \geq 0 \ \ \forall j_0 \leq j^*. \tag{57}$$

Part b: For the condition $j_0 > j^*$.

Rewrite (48):

$$\begin{aligned}&B^2((a_s), n, j_0) - B^2((a_s^*), n, j^*) \\ &= \left(\sum_{i=1}^{j_0}\frac{\lambda_i}{a_i^2} + \sum_{i=j_0+1}^{\infty}\frac{\lambda_i}{a_i^2}\right)\left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} \\ &\quad - j^*\Lambda - \sum_{i=j^*+1}^{j_0}\lambda_i - \sum_{i=j_0+1}^{\infty}\lambda_i \\ &= \left\{\sum_{i=1}^{j_0}\frac{\lambda_i}{a_i^2}\left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - j^*\Lambda - \sum_{i=j^*+1}^{j_0}\lambda_i\right\} \\ &\quad + \left\{\sum_{i=j_0+1}^{\infty}\frac{\lambda_i}{a_i^2}\left(\frac{a_1\cdots a_{j_0}}{n}\right)^{\frac{2}{j_0}} - \sum_{i=j_0+1}^{\infty}\lambda_i\right\} \\ &= F_1 + F_2.\end{aligned} \tag{58}$$

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:

$$\begin{aligned}&\left(\sum_{i=1}^{j^*}\frac{\lambda_i}{a_i^2} + \sum_{i=j^*+1}^{j_0}\frac{\lambda_i}{a_i^2}\right)D_{j_0} - j^*\Lambda - \sum_{i=j^*+1}^{j_0}\lambda_i \\ &= \left\{D_{j_0}\sum_{i=1}^{j^*}\frac{\lambda_i}{a_i^2} - j^*\Lambda\right\} + \left\{\sum_{i=j^*+1}^{j_0}\lambda_i\left(\frac{D_{j_0}}{a_i^2} - 1\right)\right\} \\ &= P_1 + P_2.\end{aligned} \tag{59}$$

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 (40) 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 $(a_s) \in A_{j_0}$, 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 (47). Hence

$$\begin{aligned}P_1 &\geq 0 + (D_{j_0} - D_{j^*})\sum_{i=1}^{j^*}\frac{\lambda_i}{a_i^2} \\ &\geq (D_{j_0} - D_{j^*})\frac{1}{D_{j^*}}j^*\Lambda \\ &> (D_{j_0} - D_{j^*})\frac{1}{D_{j^*}}j^*\lambda_{j^*+1} \geq 0.\end{aligned} \tag{60}$$

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

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

Using the 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

$$\begin{aligned}k_0 &= j_0 + 1, \\ l_m &= \min\{n < k_m : \ \tfrac{D_{j_0}}{a_i^2} - 1 \leq 0, \\ &\qquad \forall i \in \{n, \cdots, k_m - 1\}\}, \\ k_m &= \min\{n < l_{m-1} : \ \tfrac{D_{j_0}}{a_i^2} - 1 > 0, \\ &\qquad \forall i \in \{n, \cdots, l_{m-1} - 1\}\}.\end{aligned} \tag{61}$$

Since (λ_i) is a non-increasing sequence, from (61) we know

$$\begin{aligned}\lambda_i\left(\tfrac{D_{j_0}}{a_i^2} - 1\right) &\geq \lambda_{i+c}\left(\tfrac{D_{j_0}}{a_i^2} - 1\right) \\ &\qquad \forall i \in \{k_{m+1}, \cdots, l_m - 1\}, c \in \mathbb{N} \\ \lambda_i\left(\tfrac{D_{j_0}}{a_i^2} - 1\right) &> \lambda_{i-c}\left(\tfrac{D_{j_0}}{a_i^2} - 1\right) \\ &\qquad \forall i \in \{l_m, \cdots, k_m - 1\}, \\ &\qquad \forall c \in \{1, \cdots, i-1\}.\end{aligned} \tag{62}$$

Using (62), we have

$$\begin{aligned}&\sum_{i=k_{m+1}}^{k_m-1}\lambda_i\left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &\geq \lambda_{l_m}\sum_{i=k_{m+1}}^{l_m-1}\left(\frac{D_{j_0}}{a_i^2} - 1\right) + \lambda_{l_m}\sum_{i=l_m}^{k_m-1}\left(\frac{D_{j_0}}{a_i^2} - 1\right) \\ &= \lambda_{l_m}\sum_{i=k_{m+1}}^{k_m-1}\left(\frac{D_{j_0}}{a_i^2} - 1\right).\end{aligned} \tag{63}$$

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 (62) and (63) repeatedly. Finally, using (40) and $a_i^2 = D_i^i/D_{i-1}^{i-1}$ again, we can get

$$\begin{aligned}0 \quad\geq\quad P_2 &\geq \sum_{i=j^*+1}^{j^*+l}\left(\frac{D_{j_0}}{a_i^2}-1\right)\lambda_{j^*+l} \\ &= \lambda_{j^*+l}\sum_{i=j^*+1}^{j^*+l}\left(\frac{D_{j_0}}{a_i^2}-1\right) \\ &\geq \lambda_{j^*+l}\,l\left(\left(\frac{1}{a_{j^*+1}^2\cdots a_{j^*+l}^2}\right)^{\frac{1}{l}}D_{j_0}-1\right) \\ &= \lambda_{j^*+l}\,l\left(D_{j_0}\left(\frac{D_{j^*}^{j^*}}{D_{j^*+l}^{j^*+l}}\right)^{\frac{1}{l}}-1\right) \\ &\geq \lambda_{j^*+l}\,l\left(\left(\frac{D_{j^*}}{D_{j^*+l}}\right)^{\frac{j^*}{l}}-1\right) \\ &\geq \lambda_{j^*+l}\,l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}}-1\right) \\ &\qquad \text{with } l\in\{1,\cdots,k_0-1\}.\end{aligned} \tag{64}$$

Combining (60) and (64), we have

$$\begin{aligned}F_1 &= P_1+P_2 \\ &> \frac{j^*\lambda_{j^*+1}(D_{j_0}-D_{j^*})}{D_{j^*}} \\ &\quad + \lambda_{j^*+l}\,l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}}-1\right).\end{aligned} \tag{65}$$

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

$$\frac{j^*\lambda_{j^*+1}(D_{j_0}-D_{j^*})}{D_{j^*}} + \lambda_{j^*+l}\,l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}}-1\right)\geq 0. \tag{66}$$

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 (66) 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 (66) 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

$$\lambda_{j^*+l}\,l\left(\left(\frac{D_{j^*}}{D_{j_0}}\right)^{\frac{j^*}{l}}-1\right)=\frac{\lambda_{j^*+l}\,l(\Phi_*^{j^*}-\Phi_0^{j^*})}{\Phi_0^{j^*}}\leq 0, \tag{67}$$

we only need to show

$$\frac{\lambda_{j^*+1}j^*\Phi_0^{j^*}(\Phi_0^l-\Phi_*^l)}{\lambda_{j^*+l}\,l\,\Phi_*^l(\Phi_0^{j^*}-\Phi_*^{j^*})}>1. \tag{68}$$

Since $\lambda_{j^*+1}\geq\lambda_{j^*+l}$, the left hand side of (68) 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 the formula (55) again, we obtain

$$\begin{aligned}&\frac{j^*\Phi_0^{j^*}(\Phi_0^l-\Phi_*^l)}{l\,\Phi_*^l(\Phi_0^{j^*}-\Phi_*^{j^*})} \\ &\quad= \frac{j^*\Phi_0^{j^*}(\Phi_0-\Phi_*)\sum_{i=1}^{l}\Phi_0^{l-i}\Phi_*^{i-1}}{l\,\Phi_*^l(\Phi_0-\Phi_*)\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{i-1}} \\ &\quad= \frac{j^*\Phi_0^{j^*}\sum_{i=1}^{l}\Phi_0^{l-i}\Phi_*^{i-1}}{l\,\Phi_*^l\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{i-1}} \\ &\quad= \frac{j^*\sum_{i=1}^{l}\Phi_0^{j^*+l-i}\Phi_*^{i-1}}{l\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{l+i-1}} \\ &\quad= \frac{\sum_{k=1}^{j^*}\sum_{i=1}^{l}\Phi_0^{j^*+l-i}\Phi_*^{i-1}}{\sum_{k=1}^{l}\sum_{i=1}^{j^*}\Phi_0^{j^*-i}\Phi_*^{l+i-1}}.\end{aligned} \tag{69}$$

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 (69), we obtain

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

So

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

Hence

$$F_1 = P_1+P_2>0 \tag{70}$$

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

To prove $F_2 \geq 0$**.**

Using Lemma 12 again, we get

$$F_2 \geq 0. \tag{71}$$

Combining (70) and (71), we get

$$B^2((a_s), n, j_0) - B^2((a_s^*), n, j^*) \geq 0 \ \ \forall j_0 > j^*. \tag{72}$$

Combining (57) and (72), (45) is proved true. ∎

STEP FOUR

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

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

Second, for $j \leq j^*$. From (54), we get

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

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

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

$$\begin{aligned}
\left(\sqrt{\lambda_s}/a_s^*\right)_s &= \sqrt{\sum_{i=1}^{\infty} \frac{\lambda_i}{a_i^2}} \\
&= \sqrt{j^* + \frac{1}{\Lambda^2} \sum_{i=j^*+1}^{\infty} \lambda_i}.
\end{aligned} \tag{73}$$

When $k(x,y)$ and n are given, (λ_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, $(\lambda_i) \in \ell_1$ and thus $\sum_{i=j^*+1}^{\infty} \lambda_i$ is finite. So (73) is finite. Hence $\left(\sqrt{\lambda_s}/a_s^*\right)_s \in \ell_2$ is proved.

CONCLUSION

Following the proof above, we get

Corollary 14 *Suppose A_j and $B((a_s), n, j)$ are defined as above. Then we have*

$$B((a_s^*(j^*)), n, j^*) = \inf_{j \in \mathbb{N}} \inf_{(a_s) \in A_j} B((a_s), n, j), \tag{74}$$

where

$$a_i^* = \begin{cases} \sqrt{\lambda_i} & \text{when } i \leq j^* \\ \left(\frac{\sqrt{\lambda_1 \cdots \lambda_{j^*}}}{n}\right)^{\frac{1}{j^*}} & \text{when } i > j^*, \end{cases} \tag{75}$$

$$j^* = \min\left\{ j: \quad \lambda_{j+1} < \left(\frac{\lambda_1 \dots \lambda_j}{n^2}\right)^{\frac{1}{j}} \right\}. \tag{76}$$

Theorem 1 is then established.

B THE PROOF THAT INEQUALITY (31) CANNOT BE IMPROVED

Lemma 15 *Suppose A_j and $B((a_s), n, j)$ are defined as above. Let $j \in \mathbb{N}$ and $(a_s) \in A_j$. Suppose j^* and (a_s^*) exist. Then*

$$\begin{aligned}
&\inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s), n, j) \\
&= \inf_{j \in \mathbb{N}} \inf_{(a_s) \in A_j} B((a_s), n, j).
\end{aligned} \tag{77}$$

Proof Let us prove

$$\begin{aligned}
&\inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} B((a_s), n, j) \\
&\geq \inf_{j \in \mathbb{N}} \inf_{(a_s) \in A_j} B((a_s), n, j).
\end{aligned} \tag{78}$$

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

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

We have already proved

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

So, equation (77) is proved to be true. ∎

Acknowledgements

This work was supported by the Australian Research Council. Thanks to Bernhard Schölkopf and Alex J. Smola for useful discussions.