Empirical Effective Dimension and Optimal Rates for Regularized Least Squares Algorithm

Andrea Caponnetto [a] Lorenzo Rosasco [b] Ernesto De Vito [c] Alessandro Verri [d]

[a] *C.B.C.L.,McGovern Institute, Massachusetts Institute of Technology, Bldg.E25-201, 45 Carleton St., Cambridge, MA 02142*

[b] *Dipartimento di Informatica e Scienza dell'Informazione, Università degli Studi di Genova, Via Dodecaneso 35,16146 Genova, Italy*

[c] *Dipartimento di Matematica, Università di Modena e Reggio Emilia, Via Campi 213/B, 41100 Modena, Italy and I.N.F.N., Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy*

[d] *Dipartimento di Informatica e Scienza dell'Informazione, Università degli Studi di Genova, Via Dodecaneso 35,16146 Genova, Italy*

Abstract

This paper presents an approach to model selection for regularized least-squares on reproducing kernel Hilbert spaces in the semi-supervised setting. The role of *effective dimension* was recently shown to be crucial in the definition of a rule for the choice of the regularization parameter, attaining asymptotic optimal performances in a minimax sense. The main goal of the present paper is showing how the effective dimension can be replaced by an empirical counterpart while conserving optimality. The *empirical effective dimension* can be computed from independent unlabelled samples. This makes the approach particularly appealing in the semi-supervised setting.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL), as well as in the Dipartimento di Informatica e Scienze dell'Informazione (DISI) at University of Genoa, Italy.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1.

Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

This research has been partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

1 Introduction

The semi-supervised setting in statistical learning theory has been investigated in various recent papers [3], [2],[18]. The interest for this problem is especially motivated by the large variety of applications where a large amount of unlabelled data are available, but for which the process of labelling may be expensive or impractical. The practitioner is then faced with the problem of somehow exploiting all the available information on the phenomenon coded in the unlabelled data. Traditionally statistical learning theory has mostly studied the learning process in the so-called supervised setting [16],[11],[7],[6], [5],[13] where a set of input-output couples is given. It is clear that unlabelled data give some extra information regarding the marginal probability distribution on the input space. A natural starting point to a theoretically founded approach to semi-supervised learning is the analysis of the optimal rates achievable when the marginal distribution is known. This was the main goal of [4] and [8] where a criterion for the choice of the regularization parameter for regularized least-squares (RLS) on reproducing kernel Hilbert spaces (RKHS) was shown, leading to optimal rates in a marginal dependent minimax sense. In that case the optimal regularization parameter was expressed in terms of the *effective dimension*, the trace of a certain operator defined by the kernel and the marginal distribution itself.

This paper considers the following natural step in the analysis of semi-supervised learning: exploiting unlabelled data in order to replace effective dimension with an empirical version of it while conserving asymptotically optimal performances.

The plan of the paper is as follows. In section 2 we briefly recall the main concepts of statistical learning and define the RLS algorithm on RKHS. We also overview the main result of [4] giving a rule for the optimal choice of the regularization parameter in terms of the effective dimension. In section 3 we define the empirical counterpart of effective dimension. This can be expressed quite naturally by the empirical kernel matrix associated to a set of independent unlabelled data. The main result of this section is a concentration result relating empirical effective dimension and effective dimension. Finally in section 4 we generalize the main theorem of [4] to the empirical case, prove asymptotic optimality, and present a sketch of an explicit procedure that can achieve optimal rates when enough independent unlabelled samples are available. Let us stress that the procedures presented here have not been designed to be computationally effective but rather to be simple and instructive. In fact the aim of this analysis is focusing on the theoretical issues that should be considered while developing model selection techniques in the semi-supervised setting.

2 Learning Theory

We consider a compact input space $X \subset \mathbb{R}^d$ and an output space $Y = [-M, M] \subset \mathbb{R}$. The space $Z = X \times Y$ is endowed with a probability measure $\rho(x, y) = \rho_X(x)\rho(y|x)$, where $\rho_X(x)$ denotes the marginal probability measure on X and $\rho(y|x)$ the conditional probability measure of y given x. The probability measure ρ is fixed but unknown. The data we are given is a training set of ℓ pairs of examples $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = \{(x_i, y_i)\}_{i=1}^{\ell}$ drawn i.i.d. with respect to ρ, that is $\mathbf{z} \in Z^{\ell}$. Roughly speaking the goal of learning is to design a procedure that, given the training set $\mathbf{z}$, provide us with a function $f_{\mathbf{z}}$ that will correctly estimate the label y given new points x, i.e. we want $f_{\mathbf{z}}$ to *generalize*. In this paper we analyze regularized least-squares (RLS) algorithm when the hypothesis space is chosen to be a reproducing kernel Hilbert space (RKHS). For any given $\lambda > 0$ and training set $\mathbf{z}$, RLS algorithm defines an estimator $f_{\mathbf{z}}^{\lambda}$ as the solution of the following minimization problem

$$\min_{f \in \mathcal{H}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \}, \tag{1}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm in the RKHS $\mathcal{H}$ [1]. Roughly speaking the first term measures how much the estimator f fits the data whereas the second term is a penalization which constraints the complexity of the solution. The parameter λ balances out the two terms. The intuition behind the algorithm is that the regularization parameter λ allows to pass from overfitting to oversmoothing so that a good choice of the regularization parameter on the basis of the given data, $\lambda = \lambda(\mathbf{z}, \ell)$, allows to prevent both. In this sense we can think of the regularization parameter choice as a model selection procedure*. The question is then how to choose λ in order to obtain good generalization properties. To formalize the problem we can consider the squared loss function $(y - f(x))^2$ and introduce the expected loss

$$I[f] = \int_{X \times Y} (y - f(x))^2 d\rho(x, y).$$

We assume that the above functional admits a minimizer on $\mathcal{H}$ that we denote with $f_{\mathcal{H}}$. If $\mathcal{H}$ is dense ** in the space of square integrable functions with respect to ρ, then $f_{\mathcal{H}}$ is the regression function

$$f_{\rho} = \int_{Y} y d\rho(y|x).$$

In this case the problem is to find an estimator whose error is close to that of $f_{\mathcal{H}}$. In this paper we study a consistency property of RLS, in fact we want to define a choice for $\lambda_{\ell} = \lambda(\ell)$ such that for every $\varepsilon > 0$

$$\lim_{\ell \to \infty} \mathrm{P} \left[I[f_{\mathbf{z}}^{\lambda_{\ell}}] - I[f_{\rho}] \geq \epsilon \right] = 0.$$

* Though it might happen that no explicit structure of models (spaces) is considered.

**This is the case for universal kernels (for example for gaussian kernel) and we refer to [14] for details on the subject.

While studying consistency the crucial issue is indeed the convergence rate. In fact, this gives information on the finite sample behavior of the considered algorithm. Unfortunately there are classic results [10] showing that convergence rates are obtainable only under suitable assumptions on the probability distribution underlying the learning problem. Hence as we look for convergence rates we previously have to clarify the class of probability distributions $\mathcal{M}$ to which we restrict our analysis. The natural question arising, before starting the consistency analysis of a given algorithm, is that of the best attainable rates under the prior assumption $\rho \in \mathcal{M}$. The answer to this question is then given in term of minmax optimality results, that is studying lower bounds of the quantity

$$\inf_{f_{\mathbf{z}}} \sup_{\rho \in \mathcal{M}} [\mathbb{E}(I[f_{\mathbf{z}}] - I[f_{\mathcal{H}}])]$$

where the infimum is taken w.r.t. all the possible learning algorithms $\mathbf{z} \to f_{\mathbf{z}}$. Usually the class $\mathcal{M}$ is characterized through some assumption on the minimizer $f_{\mathcal{H}}$, for example smoothness or approximability properties. In [4] upper and lower bounds for RLS are proposed in the case where $f_{\mathcal{H}}$ has approximability properties in a RKHS.

After recalling some basic concepts about RKHS we briefly review the main results in [4] that we develop in the following sections.

2.1 RKHS and Covariance Operators

We briefly recall some ideas on RKHS we use in the following (see [1] for a broader introduction to the subject). A RKHS is a Hilbert space of functions uniquely defined by a symmetric positive definite function $K : X \times X \to \mathbb{R}$, namely the kernel. We say that K is positive definite if for all $m > 0$, $x_1, \dots x_m \in X$ and $c_1, \dots c_m \in \mathbb{R}$ the following inequality holds

$$\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geq 0.$$

We will assume throughout that the kernel is bounded, that is

$$\sup_{x \in X} K(x, x) \leq \kappa.$$

It will be useful to recall that the following operators are naturally defined

- Covariance operator $T : \mathcal{H} \to \mathcal{H}$

$$T := \int_X \langle \cdot, K_x \rangle_{\mathcal{H}} K_x d\rho_X(x).$$

- Empirical covariance operator $T_{\mathbf{x}} : \mathcal{H} \to \mathcal{H}$.

$$T_{\mathbf{x}} := \frac{1}{\ell} \sum_{i=1}^{\ell} \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} K_{x_i},$$

where we set $K_x = K(\cdot, x)$ and $\mathbf{x} = (x_i)_{i=1}^{\ell}$. The operators T and $T_{\mathbf{x}}$ can be proved to be positive, self-adjoint, Hilbert-Schmidt and trace-class (see, for example, the appendix in [9]).

2.2 Optimal a Priori Choice for Regularized Least Squares

We now recall the results in [4] about RLS algorithm that we develop in the following sections.

Since we look for convergence rates we first clarify the assumptions on the probability measure ρ we consider. To this aim we define the family of priors

$$\mathcal{F}(a, R) := \{f \in \mathcal{H} \mid T^{-a} f \in \mathcal{H} \ \text{ with } \ \left\|T^{-a} f\right\|_{\mathcal{H}} \leq R\},$$

where $a \in (0, 1/2]$ and $R > 0$. Moreover we consider the population version of the RLS algorithm and let f^{λ} be the solution of the problem

$$\min_{f \in \mathcal{H}} \{\int_{X \times Y} (y - f(x))^2 d\rho(x, y) + \lambda \|f\|_{\mathcal{H}}^2\}.$$

If we let $\|\cdot\|_{\rho}$ be the norm in the space of square integrable functions with respect to ρ, according to [4] we can define the following quantities:

$$\mathcal{A}(\lambda) := \left\|f^{\lambda} - f_{\mathcal{H}}\right\|_{\rho}^2 \qquad \mathcal{B}(\lambda) := \left\|f^{\lambda} - f_{\mathcal{H}}\right\|_{\mathcal{H}}^2$$

measuring the approximation error in the norm $\|\cdot\|_{\rho}$ and in the norm $\|\cdot\|_{\mathcal{H}}$ respectively. Moreover we define the effective dimension

$$\mathcal{N}(\lambda) := \mathrm{Tr}[(T + \lambda)^{-1} T]$$

which plays the role of a capacity measure for the RLS algorithm. When $f_{\mathcal{H}} \in \mathcal{F}(a, R)$ the following inequalities hold

$$\mathcal{A}(\lambda) \leq \lambda^c \left\|T^{-a} f\right\|_{\mathcal{H}}^2, \qquad \mathcal{B}(\lambda) \leq \lambda^{c-1} \left\|T^{-a} f\right\|_{\mathcal{H}}^2, \tag{2}$$

where $c = 2a + 1$ (see Lemma 6 in [4]). Moreover if the eigenvalues $(t_n)_{n=1}^{\infty}$ of the operator T fulfill $t_n = O(n^{-b})$ for some $b > 0$ then

$$\mathcal{N}(\lambda) = O(\lambda^{-1/b}), \tag{3}$$

see again Lemma 6 in [4]. Finally we recall that the stochastic order symbol is defined by the following equivalence [15]

$$X_{\ell} = O_P(h_{\ell}) \Leftrightarrow \lim_{D \to 0} \lim_{\ell \to \infty} \mathrm{supP}\left[\, |X_{\ell}| > D h_{\ell}\right] = 0.$$

The following theorem summarizes the main result in [4].

Theorem 1 *Let* $\mathbf{z}$ *be a training set drawn i.i.d according to* ρ *and* $f_{\mathbf{z}}^{\lambda}$ *the RLS estimator.*

(1) Let $0 < \eta < 1$. *If*

$$\ell \geq \frac{C_\eta \kappa}{2\lambda} \max\{\mathcal{N}(\lambda), \sqrt{2/C_\eta}\}$$

then with probability greater than $1 - \eta$,

$$I[f_{\mathbf{z}}^{\lambda}] - I[f_{\mathcal{H}}] \leq C_\eta \left(\mathcal{A}(\lambda) + \frac{\kappa^2 \mathcal{B}(\lambda)}{\ell^2 \lambda} + \frac{\kappa \mathcal{A}(\lambda)}{\ell \lambda} + \frac{\kappa M}{\ell^2 \lambda} + \frac{M \mathcal{N}(\lambda)}{\ell} \right)$$

where $C_\eta = 128 \log^2(8/\eta)$.

(2) Assume $f_\rho \in \mathcal{F}(a, R)$, $a \in (0, 1/2]$ *and that the eigenvalues* $(t_n)_{n=1}^{\infty}$ *of the operator* T *fulfill* $t_n = O(n^{-b})$ *for some* $b > 0$. *If we choose the unique value* $\lambda_0 = \lambda_0(\ell)$ *such that*

$$\mathcal{N}(\lambda_0) = \ell \lambda_0^c$$

then

$$I[f_{\mathbf{z}}^{\lambda_0}] - I[f_{\mathcal{H}}] = O_P(\ell^{-\frac{cb}{cb+1}}) \tag{4}$$

where $c = 2a + 1$.

Remark 1 *The rate in (4) can be shown to be optimal in a minmax sense with respect to the considered class of probability distributions [4].*

Remark 2 *If the hypothesis space* $\mathcal{H}$ *is finite dimensional then the convergence rate is* ℓ^{-1}.

3 Empirical Effective Dimension

In this section we show that effective dimension can be empirically estimated from a set of unlabelled data.

Definition 1 *Let* $\mathbf{x} = (x)_{i=1}^m$ *a set of* m *inputs. We define the empirical effective dimension as*

$$\mathcal{N}_{\mathbf{x}}(\lambda) := \mathrm{Tr}[(T_{\mathbf{x}} + \lambda)^{-1} T_{\mathbf{x}}].$$

The main result of this section is the following concentration result relating the effective dimension to the empirical effective dimension.

Theorem 2 *Let* $\lambda > 0$, $m \in \mathbb{N}$ *and* $\mathbf{x} = (x)_{i=1}^m$ *a set of* m *input values drawn i.i.d. according to* ρ_X. *Let* $0 < \eta < 1, \Delta > 0$ *if*

$$m \geq \Gamma(\Delta, \eta, \lambda) := \left(4 \frac{\kappa}{\lambda \Delta} \left(1 + \frac{\kappa}{\lambda} \right) \ln \frac{2}{\eta} \right)^2.$$

the following inequality holds with probability $1 - 2\eta$

$$|\mathcal{N}(\lambda) - \mathcal{N}_{\mathbf{x}}(\lambda)| \leq \Delta.$$

3.1 Proof

To prove the above theorem we need the following probabilistic inequality for random variables in Hilbert spaces due to [12]. We use in particular the following simple restatement of of Th. 3.3.4 of [17], whose proof can be found in [4].

Lemma 3 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and ξ be a random variable on Ω taking values in a real separable Hilbert space K. Assume that there are two positive constants H and σ such that*

$$\|\xi(\omega)\|_K \leq \frac{H}{2} \quad \text{a.}s \tag{5}$$

$$\mathbb{E}[\|\xi\|_K^2] \leq \sigma^2. \tag{6}$$

Let $\ell \in \mathbb{N}$ and $0 < \eta < 1$, then

$$\mathbf{P}^\ell \left[(\omega_1, \ldots, \omega_\ell) \in \Omega^\ell \mid \left\| \frac{1}{\ell} \sum_{i=1}^{\ell} \xi(\omega_i) - \mathbb{E}[\xi] \right\|_K \leq \left(\frac{H}{\ell} + \frac{\sigma}{\sqrt{\ell}} \right) 2 \log \frac{2}{\eta} \right] \geq 1 - \eta. \tag{7}$$

We can now prove Theorem (2).

PROOF. We first claim that

$$|\mathcal{N}(\lambda) - \mathcal{N}_{\mathbf{x}}(\lambda)| = |\operatorname{Tr}((T+\lambda)^{-1}T - (T_{\mathbf{x}}+\lambda)^{-1}T_{\mathbf{x}})| \leq \Delta N_1(\mathbf{x}) + \Delta N_2(\mathbf{x}), \tag{8}$$

where

$$\Delta N_1(\mathbf{x}) = \operatorname{Tr}((T+\lambda)^{-1}(T - T_{\mathbf{x}})) \quad \text{and} \quad \Delta N_2(\mathbf{x}) = \frac{\kappa}{\lambda^2} \|T - T_{\mathbf{x}}\|$$

where $\|\cdot\|$ is the norm in the Banach space of linear bounded operators form $\mathcal{H}$ to $\mathcal{H}$. We start by considering the following simple algebraic equalities

$$\begin{aligned} &(T+\lambda)^{-1}T - (T_{\mathbf{x}}+\lambda)^{-1}T_{\mathbf{x}} = \\ &(T+\lambda)^{-1}(T - T_{\mathbf{x}}) + [(T+\lambda)^{-1} - (T_{\mathbf{x}}+\lambda)^{-1}]T_{\mathbf{x}} = \\ &(T+\lambda)^{-1}(T - T_{\mathbf{x}}) + (T+\lambda)^{-1}(T_{\mathbf{x}} - T)(T_{\mathbf{x}}+\lambda)^{-1}T_{\mathbf{x}}. \end{aligned} \tag{9}$$

Recalling that

$$\left\|(T+\lambda)^{-1}\right\| \leq \frac{1}{\lambda} \qquad \left\|(T_{\mathbf{x}}+\lambda)^{-1}\right\| \leq \frac{1}{\lambda}$$

we have

$$\begin{aligned} &\operatorname{Tr}((T+\lambda)^{-1}(T_{\mathbf{x}} - T)(T_{\mathbf{x}}+\lambda)^{-1}T_{\mathbf{x}}) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left\langle (T+\lambda)^{-1}(T_{\mathbf{x}} - T)(T_{\mathbf{x}}+\lambda)^{-1}K_{x_i}, K_{x_i} \right\rangle_{\mathcal{H}} \leq \frac{\kappa}{\lambda^2} \|T - T_{\mathbf{x}}\|, \end{aligned} \tag{10}$$

and (8) follows by taking the trace of (9) and using Inequality (10).

To finish the proof we need to give probabilistic bounds on $\Delta N_1(\mathbf{x})$ and $\Delta N_2(\mathbf{x})$, to this purpose we are going to use Lemma 3. We first give the bound on $\Delta N_1(\mathbf{x})$. Let us consider the random variable $\xi_1 : X \to \mathbb{R}$

$$\xi_1(x) := \left\langle K_x, (T+\lambda)^{-1} K_x \right\rangle_{\mathcal{H}}.$$

It is straightforward to check that

$$|\xi_1(x)| \leq \frac{\kappa}{\lambda}$$

and moreover

$$\frac{1}{m}\sum_{i=1}^{m} \xi_1(x_i) = \mathrm{Tr}((T+\lambda)^{-1} T_{\mathbf{x}}) \qquad \mathbb{E}[\xi_1] = \mathrm{Tr}((T+\lambda)^{-1} T).$$

We can then apply Lemma 3 with $H = \sigma = \kappa/\lambda$ to get with probability greater than $1-\eta$

$$\Delta N_1(\mathbf{x}) \leq \frac{2\kappa}{\lambda} \ln \frac{2}{\eta} \left(\frac{1}{m} + \frac{1}{\sqrt{m}} \right).$$

Finally we study the bound on $\Delta N_2(\mathbf{x})$. Recall that both T and $T_{\mathbf{x}}$ are Hilbert-Schmidt operators so that we can apply Lemma (3). If we let $\mathcal{L}_2(\mathcal{H})$ be the space of Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}$ and we denote with $\|\cdot\|_{\mathcal{L}_2(\mathcal{H})}$ the Hilbert-Schmidt norm the uniform norm is dominated by the norm $\|\cdot\|_{\mathcal{L}_2(\mathcal{H})}$. Then we can introduce the random variable $\xi_2 : X \to \mathcal{L}_2(\mathcal{H})$

$$\xi_2(x) := \langle \cdot, K_x \rangle_{\mathcal{H}} K_x.$$

Again it is easy to check that

$$\|\xi_2\|_{\mathcal{L}_2(\mathcal{H})} \leq \kappa$$

and moreover

$$\frac{1}{m}\sum_{i=1}^{m} \xi_2(x_i) = T_{\mathbf{x}} \qquad \mathbb{E}[\xi_2] = T.$$

Applying Lemma 3 with $H = \sigma = \kappa$ we get with probability greater $1-\eta$

$$\Delta N_2(\mathbf{x}) \leq \frac{2\kappa^2}{\lambda^2} \ln \frac{2}{\eta} \left(\frac{1}{m} + \frac{1}{\sqrt{m}} \right),$$

and the theorem is proved.

4 Optimal parameter choice in Semi-supervised Setting

In Theorem 1 to define an optimal a priori choice for the regularization parameter for a given prior we need to know the effective dimension $\mathcal{N}(\lambda)$. In a semi-supervised setting we can use the concentration result of the previous section to replace $\mathcal{N}(\lambda)$ with an empirical estimate based on unlabelled data. The goal of this section is to show that using such an estimate we can define a data-dependent parameter choice achieving the optimal convergence rate.

Recall that if we know $\mathcal{N}(\lambda)$ we can choose the value λ_0 according to Theorem 1 to achieve the optimal rate. The idea behind our parameter choice is to replace $\mathcal{N}(\lambda)$ with an approximation based on unlabelled data. Roughly speaking, to ensure that the parameter choice based on unlabelled data is still optimal we have to suitably control the quality of the empirical estimate for $\mathcal{N}(\lambda)$. To clarify this we let $0 < \alpha_- < 1 < \alpha_+$ be two fixed constants and define the values λ_ℓ^+ and λ_ℓ^- such that

$$\alpha_+ \mathcal{N}(\lambda_\ell^+) = \ell(\lambda_\ell^+)^c \quad \text{and} \quad \alpha_- \mathcal{N}(\lambda_\ell^-) = \ell(\lambda_\ell^-)^c. \tag{11}$$

It is possible to show that if we choose either λ_ℓ^+ or λ_ℓ^- we get the same convergence rate as choosing λ_0. Intuitively we want our estimates for $\mathcal{N}(\lambda)$ to lie, with high probability, between $\alpha_+\mathcal{N}(\lambda)$ and $\alpha_-\mathcal{N}(\lambda)$ for each value of λ. In this case we expect to be able to select λ so that the good asymptotic properties are maintained.

We now formalize the above idea. The first step toward the definition of our parameter choice rule is to consider a suitable discretization criterion for λ. This is most reasonable from a practical point of view and will not prevent us to obtain optimal convergence results. The following assumption describe the discretization procedure that we are going to consider.

Assumption 1 *We discretize the possible values for the regularization parameter considering $0 < \lambda_k \leq \lambda_{k-1}$ with $k = 1, 2, \ldots$ such that*

$$\lambda_k \geq q\lambda_{k-1}. \tag{12}$$

The following assumption describes the regularization parameter choice we consider.

Assumption 2 *We assume the index $k(\ell) \in \mathbb{N}$ be such that if we let $\hat{\lambda}_\ell^+ := \lambda_{k(\ell)}$ and $\hat{\lambda}_\ell^- = \lambda_{k(\ell)-1}$ then, the following conditions hold true*

$$\alpha_- \mathcal{N}(\hat{\lambda}_\ell^+) \leq \ell(\hat{\lambda}_\ell^+)^c \quad \textit{and} \quad \alpha_+ \mathcal{N}(\hat{\lambda}_\ell^-) \geq \ell(\hat{\lambda}_\ell^-)^c. \tag{13}$$

In Section 4.3 we show how to actually find $\hat{\lambda}_\ell^+$ in an iterative way. Next theorem shows that choosing $\lambda = \hat{\lambda}_\ell^+$ we can actually achieve the same rate of the optimal value λ_0.

Theorem 4 *Under the same hypotheses of Theorem 1 Item 2, if Assumption 1 holds and the random variables $(k(\ell))_{\ell\in\mathbb{N}}$, with values on $\mathbb{N}$, fulfill Assumption 2 with probability greater than $1 - \bar{\eta}(\ell)$ for some $\bar{\eta}(\ell) \to 0$. Then defining $\lambda_\ell := \hat{\lambda}_\ell^+ := \lambda_{k(\ell)}$ one has*

$$I[f_{\mathbf{z}}^{\lambda_\ell}] - I[f_{\mathcal{H}}] = O_P(\ell^{-\frac{cb}{cb+1}}) \tag{14}$$

where $c = 2a + 1$.

4.2 Proof

Before giving the proof of Theorem (4) we collect a few simple results on our parameter choice in the following Lemma.

Lemma 5 *Let $\hat{\lambda}_\ell^+, \hat{\lambda}_\ell^-$ as in Assumption 2 and $\lambda_\ell^+, \lambda_\ell^-$ as in (11). Then*

(1) the following inequalities hold

$$\hat{\lambda}_\ell^+ \geq \lambda_\ell^- \quad \text{and} \quad \hat{\lambda}_\ell^- \leq \lambda_\ell^+. \tag{15}$$

(2) $\hat{\lambda}_\ell^+ \to 0$ as $\ell \to \infty$.

(3) The following inequality holds true

$$(\hat{\lambda}_\ell^+)^c + \mathcal{N}(\hat{\lambda}_\ell^+)/\ell \leq (q^c + (\alpha_-)^{-1})(\lambda_\ell^+)^c \tag{16}$$

.

PROOF. We first prove Item 1 by contradiction. If we let $\hat{\lambda}_\ell^+ < \lambda_\ell^-$ then

$$\mathcal{N}(\hat{\lambda}_\ell^+) \geq \mathcal{N}(\lambda_\ell^-)$$

so that by Assumption 2

$$\alpha_- \mathcal{N}(\lambda_\ell^-) \leq \ell(\hat{\lambda}_\ell^+)^c < \ell(\lambda_\ell^-)^c$$

which is impossible because of (11). Similarly one can prove that $\hat{\lambda}_\ell^- \leq \lambda_\ell^+$. The proof of Item 2 follows from Item 1 and Assumption 1, In fact we have

$$\hat{\lambda}_\ell^+ \leq q\hat{\lambda}_\ell^- \leq q\lambda_\ell^+$$

and the proof follows since $\lambda_\ell^+ \to 0$ as $\ell \to \infty$. Finally from Assumption 1 and Item 1 we have

$$(\hat{\lambda}_\ell^+)^c \leq q^c(\hat{\lambda}_\ell^-)^c \leq q^c(\lambda_\ell^+)^c$$

and from Item 1 and (11)

$$\mathcal{N}(\hat{\lambda}_\ell^+)/\ell \leq \mathcal{N}(\lambda_\ell^-)/\ell = (\alpha_-)^{-1}(\lambda_\ell^-)^c \leq (\alpha_-)^{-1}(\lambda_\ell^+)^c.$$

Putting the above inequalities together we get (16)

We are now ready to prove Theorem (4).

PROOF. [Theorem 4] The proof is similar to that of Theorem (1) Item 2 (see [4]). We assume that $k(\ell)$ is a random variable fulfilling Assumption 2 with confidence

level $1-\overline{\eta}(\ell)$ where $\overline{\eta}(\ell) \to 0$ as $\ell \to \infty$. Recall that $c > 1$, if we let $0 < \eta < 1$ then with confidence level at least $1-\overline{\eta}(\ell)$

$$\ell\hat{\lambda}_\ell^+ = \ell(\hat{\lambda}_\ell^+)^{1-c}(\hat{\lambda}_\ell^+)^c \geq (\hat{\lambda}_\ell^+)^{1-c}\alpha_-\mathcal{N}(\hat{\lambda}_\ell^+).$$

Since $1-c < 0$, from Lemma (5) Item 2 we know that it exists $\ell(\eta) \in \mathbb{N}$ such that

$$\mathrm{P}\left[\ell \leq \frac{C_\eta \kappa}{2\hat{\lambda}_\ell^+}\max\{\mathcal{N}(\hat{\lambda}_\ell^+), \sqrt{2/C_\eta}\}\right] \leq \overline{\eta}(\ell)$$

for $\ell > \ell(\eta)$. If we now consider to choose the value $\hat{\lambda}_\ell^+$ then we have

$$\mathrm{P}\left[X_\ell > C_\eta\left(\mathcal{A}(\hat{\lambda}_\ell^+) + \frac{\kappa^2\mathcal{B}(\hat{\lambda}_\ell^+)}{\ell^2\hat{\lambda}_\ell^+} + \frac{\kappa\mathcal{A}(\hat{\lambda}_\ell^+)}{\ell\hat{\lambda}_\ell^+} + \frac{\kappa M}{\ell^2\hat{\lambda}_\ell^+} + \frac{M\mathcal{N}(\hat{\lambda}_\ell^+)}{\ell}\right)\Big|\hat{\lambda}_\ell^+\right] \leq \eta$$

where $C_\eta = 128\log^2(8/\eta)$ and $X_\ell = I[f_{\mathbf{z}}^{\hat{\lambda}_\ell^+}] - I[f_{\mathcal{H}}]$. Using Lemma 6 in [4] we can simplify the form of the above bound. In fact it is easy to show (see the proof of Theorem (1) in [4]) that asymptotically the first and the last term in the bound prevail so that a positive constant C' and a natural number $\ell'(\eta)$ exist for which

$$\mathrm{P}\left[X_\ell > C'C_\eta\left(\mathcal{A}(\hat{\lambda}_\ell^+) + \frac{\mathcal{N}(\hat{\lambda}_\ell^+)}{\ell}\right)\Big|\hat{\lambda}_\ell^+\right] \leq \eta, \quad \forall \ell > \ell'(\eta).$$

If we now apply (2) and Lemma (5) Item 3, we can rewrite the above bound using stochastic order symbol [15]. In fact a positive constant C'' exist for which

$$\mathrm{P}\left[\, X_\ell > D(\lambda_\ell^+)^c\right] \leq 8e^{-\sqrt{D/128C''}} + \overline{\eta}(\ell),$$

if $D > 128C''(q^c + (\alpha_-)^{-1})\log^2 8$ and $\ell > \ell''(D)$, then we have

$$I[f_{\mathbf{z}}^{\hat{\lambda}_\ell^+}] - I[f_{\mathcal{H}}] = O_P\left((\lambda_\ell^+)^c\right).$$

Recalling (see again Lemma 6 in [4]) that if the eigenvalues of T satisfy $t_n = O(n^{-b})$ then $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{b}})$, from the definition of λ_ℓ^+ we have

$$\ell(\lambda_\ell^+)^c = O((\lambda_\ell^+)^{-\frac{1}{b}})$$

which implies $(\lambda_\ell^+)^c = O(\ell^{-\frac{bc}{bc+1}})$ and the theorem is proved.

4.3 Model Selection from Unlabelled Data

In this section we present an explicit procedure to find from unlabelled data the index $k(\ell)$ satisfying Assumption 2 with a given confidence level $1-\bar{\eta}(\ell)$. The corresponding regularization parameter choice $\lambda_\ell := \hat{\lambda}_\ell^+ := \lambda_{k(\ell)}$ plays the central

role in Theorem 4. The fundamental condition to accomplish our scheme is to have unlabelled data available, from now on we indicate with

$$\texttt{unlabelled}(m) \rightarrow \mathbf{x} \quad \text{with} \quad |\mathbf{x}| = m \tag{17}$$

the procedure providing us with $m \in \mathbb{N}$ unlabelled examples.

First we describe the procedure that for each value of λ provides us with the approximation of $\mathcal{N}(\lambda)$ we need for a fixed confidence level $1-\eta$ and relative error $0 < \delta < 1$. We let

$$\Gamma(\Delta, \eta, \lambda)$$

as in Theorem (2) and recall that

$$\mathcal{N}_{\mathbf{x}}(\lambda) = \mathrm{Tr}((T_{\mathbf{x}} + \lambda I)^{-1} T_{\mathbf{x}}) = \mathrm{Tr}((\mathbf{K} + \lambda I)^{-1}\mathbf{K}).$$

where $\mathbf{K}_{ij} = k(x_i, x_j)$. We now first give the procedure `eff_dim`(λ, η) and then briefly explain it.

`eff_dim`(λ, η)

- $j = 1$
- `do unlabelled`$(\Gamma(2^{-j}, 2^{-(j+1)}\eta, \lambda)) \rightarrow \mathbf{x}$;j+=1
- `until` $\mathcal{N}_{\mathbf{x}}(\lambda) \geq 2^{-j+1}$
- `unlabelled`$(\Gamma(2^{-j}\delta, 2^{-1}\eta, \lambda)) \rightarrow \mathbf{x}$
- `return` $\mathcal{N}_{\mathbf{x}}(\lambda)$

It is easy to show, applying theorem 2, that with probability greater than $1-\eta$, $\mathcal{N}$, the output of `eff_dim`(λ, η), is bounded from above and from below in terms of $\mathcal{N}(\lambda)$, formally we have

$$\mathrm{P}\left[\,(1-\delta)\mathcal{N}(\lambda) \leq \mathcal{N} \leq (1+\delta)\mathcal{N}(\lambda)\right] \geq 1-\eta,$$

where δ is the constant appearing in the text of `eff_dim`.

`eff_dim` is called by the procedure `mod_sel` given below. `mod_sel`(ℓ, η) returns the integer $k(\ell)$ fulfilling with confidence level $1-\eta$ Assumption 2 used in the previous section. The idea behind the procedure is simply exploring the grid $(\lambda_k)_k$ until a crossing between the approximation term $\ell\lambda^c$ and our estimate of $\mathcal{N}(\lambda)$ is encountered. Clearly this strategy is performed while properly controlling the accuracy of the estimate of $\mathcal{N}(\lambda)$ and its confidence level.

`mod_sel`(ℓ, η)

- $k, j = 1$
- $\sigma = \text{sign}(\texttt{eff\_dim}(\lambda_k, 2^{-1}\eta(\ell)) - \ell(\lambda_k)^c)$
- `do` $k = k + \sigma$; $j+ = 1$
- `until` $\sigma(\texttt{eff\_dim}(\lambda_k, 2^{-j}\eta(\ell)) - \ell(\lambda_k)^c) \geq 0$
- `return` $k + \frac{\sigma - 1}{2}$

Acknowledgments

We would like to thank T. Poggio for useful discussions and suggestions.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. *COLT*, 2004.

[3] M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56, , Special Issue on Clustering:209–239, 2004.

[4] A. Caponnetto and E. De Vito. Fast rates for regularized least-squares algorithm. Technical report, Massachusetts Institute of Technology, Cambridge, MA, April 2005. CBCL Memo#248/CSAIL Memo#2005-013.

[5] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computationals Mathematics*, 2:413–428, 2002.

[6] F. Cucker and S. Smale. On the mathematical foundation of learning. *Bull. A.M.S.*, 39:1–49, 2002.

[7] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[8] E. De Vito and A. Caponnetto. Risk bounds for regularized least-squares algorithm with operator-valued kernels. Technical report, Massachusetts Institute of Technology, Cambridge, MA, May 2005. CBCL Memo#249/CSAIL Memo#2005-015.

[9] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, May 2005.

[10] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[11] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

[12] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 30(1):143–148, 1985.

[13] Tomaso Poggio and Steve Smale. The mathematics of learning: dealing with data. *Notices Amer. Math. Soc.*, 50(5):537–544, 2003.

[14] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[15] Sara Van De Geer. *Empirical Processes in M-Estimation*. CBMS-NSF Regional Conference Series in Applied Mathematics. Cambridge University Press, 2000.

[16] V. N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.

[17] V. Yurinsky. *Sums and Gaussian vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

[18] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In T. Fawcett and N. Mishra, editors, *Procedings of the 20th Iternational Conference of Machine Learning*. AAAI Press, 2003.