Sādhanā Vol. 31, Part 2, April 2006, pp. 155–171. © Printed in India

Complexity control in statistical learning

SAMEER M JALNAPURKAR

Department of Mathematics, Indian Institute of Science, Bangalore 560 012, India
e-mail: smj@math.iisc.ernet.in

Abstract. We consider the problem of determining a model for a given system on the basis of experimental data. The amount of data available is limited and, further, may be corrupted by noise. In this situation, it is important to control the *complexity* of the class of models from which we are to choose our model. In this paper, we first give a simplified overview of the principal features of learning theory. Then we describe how the method of regularization is used to control complexity in learning. We discuss two examples of regularization, one in which the function space used is finite dimensional, and another in which it is a reproducing kernel Hilbert space. Our exposition follows the formulation of Cucker and Smale. We give a new method of bounding the sample error in the regularization scenario, which avoids some difficulties in the derivation given by Cucker and Smale.

Keywords. Complexity control; learning theory; regularisation covering number.

1. Introduction

In this paper, we consider the problem of how we can choose the best possible model for a physical system on the basis of experimental data. In this introductory section, we try to give the reader an intuitive idea of some of the issues that arise in this process.

1.1 *Simplicity vs. accuracy of models*

The first question that we need to ask is what we mean by a "good" model. What are the attributes of a good model? First, we would like the model to be *accurate*, i.e., fit the data well. Second, we would like to model to be as simple as possible. These are conflicting requirements – to fit the data well, we may be forced to use a complex model.

The trade-off between simplicity and accuracy is well illustrated by the historical process through which we arrived at our understanding of the solar system. Ptolemy's model is not simple, and does not account for the observations too well either. Copernicus' heliocentric model, with planets going around the sun in circular orbits, is much simpler, but its accuracy is still not satisfactory. Kepler's model allows for elliptical orbits and is thus more complex than Copernicus' model. However, it is far more accurate. Later, Newton discovered the Law of Gravitation. This can be stated more compactly than Kepler's laws of planetary motion, yet Kepler's laws can be deduced from it. Thus, Newton's contribution resulted in a simplification of Kepler's model. Finally, Einstein's theory of General Relativity constitutes an increase in

complexity, but it can explain the minute perturbations to the orbit of Mercury caused by the rotation of the Sun on its axis.

1.2 *The optimal trade-off between simplicity and accuracy*

Suppose now that we are trying to fit a model to some experimental data. By considering more and more complex models, we can get better and better fits for the data. In fact, by taking a sufficiently complex model, we will be able to fit each data point exactly. But it is intuitively clear that this is not necessarily the best thing to do. We do want to fit the data well, but we do not want to "overfit". So, there must be some optimal level of complexity that will give us the best possible model.

What are the factors that will determine what this optimal level of complexity is? For one thing, our measuring instruments may not be very good, and the data may be corrupted by noise. By considering more and more complex models, we will only be expending our energy in trying to fit the noise, and the models we get will not be able to explain any new data. Another factor is the amount of data available to us. If we have only a small amount of data, there will be many models that fit this data well. Then, how are we to choose amongst them?

Thus, it appears that we should be willing to consider more complex models if our data is more accurate. Similarly, we should be willing to consider more complex models if we have a large amount of experimental data. One of the important tasks for us in the rest of this report is to make these intuitive notions more precise.

1.3 *An outline of this paper*

In §2 we mathematically formulate the learning problem. The approach is to associate with each model an *error*, which is indicative of the performance of that model. Our goal is to choose, from a class of models, a model with the least possible error. We discuss how the size of the class of models crucially affects the performance of our learning schemes. In §3, we discuss the use of "regularization" to control the effective size of our class of models. We also give two example to illustrate the theory. Finally, in §4, we describe several future directions that remain to be explored.

As is explained in §2, we use the concept of *covering numbers* to quantify the complexity of a class of models. There also other ways of quantifying complexity. For example, two such measures of complexity are the *VC-dimension* and the *fat shattering dimension*. These are not discussed here. An exposition of these measures of complexity, and learning methods based thereon, may be found, for example in the books by Vapnik (1998) and Vidyasagar (1997). The approach we follow here, based on covering numbers, is simpler.

We also mention in passing that the learning problem can be formulated as an *Ill-posed Inverse Problem*. Methods based on this approach do not explicitly use a measure of the complexity of a model class. Some references are DeVito *et al* (2005) and Smale & Zhou (2005).

2. Formulating the learning problem

Suppose that we have a system with inputs in a compact set $\mathcal{X} \subset \mathbb{R}^n$, and outputs in $\mathcal{Y} = \mathbb{R}$. Our observations of the system are thus elements of the space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The behaviour of our system is represented by a random variable $Z = (X, Y)$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in $\mathcal{Z}$. The distribution ρ of Z is not known. The goal is to discover

a functional relationship between the inputs and outputs. For a function $f : \mathcal{X} \to \mathcal{Y}$, the *loss* suffered with a data point $z = (x, y)$ is

$$l_f(x, y) := (y - f(x))^2.$$

$l_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called a loss function. The expected value of the l_f is called the *error* $\mathcal{E}(f)$ of f:

$$\mathcal{E}(f) = \int_{\mathcal{Z}} l_f(z)\rho(\mathrm{d}z).$$

Since ρ is not known, $\mathcal{E}(f)$ is not computable.

It is of interest to note that

$$\mathcal{E}(f) = \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} (y - f(x))^2 \rho(\mathrm{d}y|x) \right] \rho(\mathrm{d}x), \tag{1}$$

where $\rho(\mathrm{d}y|x)$ is the regular conditional distribution of Y, given X. We denote by $f_\rho(x)$ the conditional mean of the output, for a given input x. Note

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

The inner integral in (1) can be rewritten as follows:

$$\begin{aligned}\int_{\mathcal{Y}} (y - f(x))^2 \rho(\mathrm{d}y|x) &= \int_{\mathcal{Y}} ((y - f_\rho(x)) + (f_\rho - f(x)))^2 \rho(\mathrm{d}y|x) \\ &= \int_{\mathcal{Y}} (y - f_\rho(x))^2 \rho(\mathrm{d}y|x) \\ &\quad + \int_{\mathcal{Y}} (y - f_\rho(x))(f_\rho - f(x))\rho(\mathrm{d}y|x) + (f_\rho - f(x))^2.\end{aligned}$$

Since $\int_{\mathcal{Y}}(y)\rho(\mathrm{d}y|x) = f_\rho(x)$, the above equals

$$(f(x) - f_\rho(x))^2 + \int_{\mathcal{Y}} (y - f_\rho(x))^2 \rho(\mathrm{d}y|x).$$

The second term on the right hand side of the above equation is the conditional variance of the output. We denote this by $\sigma^2(x)$. Let the expected conditional variance be

$$\sigma_\rho^2 := \int_{\mathcal{X}} \sigma^2(x)\rho(\mathrm{d}x).$$

The larger the value of σ_ρ^2, the noisier is the data. Thus (1) becomes

$$\mathcal{E}(f) = \|f - f_\rho\|^2_{L^2_\rho(\mathcal{X})} + \sigma_\rho^2.$$

Here $L^2_\rho(\mathcal{X})$ denotes the space of functions on $\mathcal{X}$ that are square integrable with respect to the distribution of X. Note that $\mathcal{E}$ achieves its minimum at the conditional mean function f_ρ,

and the minimum value is σ_ρ^2. In the rest of this paper, we denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner products and norms in $L^2_\rho(\mathcal{X})$.

We restrict ourselves to functions that are elements of a *hypothesis space* $\mathcal{H}$, which is a subset of the space $\mathcal{C}(\mathcal{X})$ of continuous real valued functions on $\mathcal{X}$. $\mathcal{C}(\mathcal{X})$ is a normed linear space with the supremum norm. For now, we assume that $\mathcal{H}$ is compact in $\mathcal{C}(\mathcal{X})$.

We suppose that we are given m observations of the system, represented by the *iid* random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$, with $Z_i = (X_i, Y_i)$ having distribution ρ for $i = 1 \ldots, m$. We must choose a hypothesis $f \in \mathcal{H}$ based on this empirical data alone. Though we cannot compute $\mathcal{E}$, we can compute the *empirical error*, which is the empirical mean of the loss function:

$$\mathcal{E}_{\mathbf{Z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (Y_i - f(X_i))^2.$$

The natural thing to do now would be to choose as our hypothesis the function $f_{\mathbf{Z}}$ that minimizes the empirical error on $\mathcal{H}$, i.e.,

$$f_{\mathbf{Z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{Z}}(f)$$

Given that we have nothing but the empirical data to work with, this is perhaps the only thing we could do, for a given $\mathcal{H}$. But what makes the situation interesting is that we get to choose $\mathcal{H}$.

One consideration in choosing $\mathcal{H}$ is that it be rich enough to contain functions f for which $\mathcal{E}$ is small. A second consideration for choosing $\mathcal{H}$ comes from the observation that we are justified in using the minimizer of $\mathcal{E}_{\mathbf{Z}}$, rather than that of $\mathcal{E}$, *provided* $\mathcal{E}_{\mathbf{Z}}$ *is uniformly close to* $\mathcal{E}$ *on* $\mathcal{H}$. Thus, we need to choose $\mathcal{H}$ so that the difference $L_{\mathbf{Z}} = \mathcal{E}_{\mathbf{Z}} - \mathcal{E}$ is uniformly small on $\mathcal{H}$. In fact, if $|L_{\mathbf{Z}}|$ is less than ε on $\mathcal{H}$, $\mathcal{E}(f_{\mathbf{Z}})$ is within 2ε of its minimum value on $\mathcal{H}$. This is shown as follows: We let $f_{\mathcal{H}} := \arg\min_{f \in \mathcal{H}} \mathcal{E}(f)$. We have that

$$\begin{aligned} 0 \leq \mathcal{E}(f_{\mathbf{Z}}) - \mathcal{E}(f_{\mathcal{H}}) &\leq (\mathcal{E}(f_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(f_{\mathbf{Z}})) + (\mathcal{E}_{\mathbf{Z}}(f_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(f_{\mathcal{H}})) \\ &\quad + (\mathcal{E}_{\mathbf{Z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})) \leq 2\varepsilon, \end{aligned} \tag{2}$$

since the first and third terms on the right are each less than ε, and the middle term is non-positive.

Note that $\mathcal{E}_{\mathbf{Z}}$ and therefore $L_{\mathbf{Z}}$ are random variables. It may not be true that $|L_{\mathbf{Z}}| < \varepsilon$ for all possible values of $\mathbf{Z}$. But we are able to show that if $\mathcal{H}$ is small enough (in a sense to be made precise later), then with high probability, $\{\sup_{f \in \mathcal{H}} |L_{\mathbf{Z}}(f)| < \varepsilon\}$. Here is an outline of how this can be shown; a detailed exposition may be found in Cucker & Smale (2002b).

(1) For any fixed $f \in \mathcal{H}$, we can use either the Hoeffding or Bernstein "concentration inequalities" to show that with high probability $|L_{\mathbf{Z}}(f)|$ is small. These inequalities can be stated as follows: If $\xi_1, \ldots, \xi_m$ are *iid* RVs with $|\xi_i| \leq M$, with mean μ, and with variance σ^2, then it possible to get exponential bounds on the convergence of empirical means to expectations:

$$\text{Prob}\left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i - \mu \right| \geq \varepsilon \right\} \leq 2 \exp\left[-\frac{m\varepsilon^2}{2M^2} \right] \text{ (Hoeffding), and}$$

$$\text{Prob}\left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i - \mu \right| \geq \varepsilon \right\} \leq 2 \exp\left[-\frac{m\varepsilon^2}{2(\sigma^2 + (1/3)M\varepsilon)} \right] \text{ (Bernstein).}$$

We apply these inequalities with $\xi_i = (Y_i - f(X_i))^2$. For application of these inequalities, we need that $(Y - f(X))^2$ be bounded. Note

$$|Y - f(X)| \leq |Y - f_\rho(X)| + |f_\rho(X) - f(X)|$$
$$\leq |Y - f_\rho(X)| + |f_\rho(X)| + |f(X)|.$$

We make the assumptions that $\|Y - f_\rho(X)\|_\infty < \infty$, and that $\|f_\rho\|_\infty < \infty$. (Here $\|\cdot\|_\infty$ denotes the essential supremum with respect to the underlying probability measure ρ.) Let $M_1 := \|Y - f_\rho(X)\|_\infty + \|f_\rho\|_\infty$. By compactness of $\mathcal{H}$, there is a constant A such that for all $f \in \mathcal{H}$, $\|f\|_\infty \leq A$. Thus,

$$\|Y - f(X)\|_\infty \leq M_1 + A. \tag{3}$$

Therefore the variables $\xi_i = (Y_i - f(X_i))^2$ are bounded and we can apply the above concentration inequalities.

(2) For any given $\varepsilon > 0$, we can cover $\mathcal{H}$ by a finite number of ε-balls (defined using the metric obtained from the supremum norm on $\mathcal{C}(\mathcal{X})$). This is because $\mathcal{H}$ is compact as a subset of $\mathcal{C}(\mathcal{X})$. Suppose n ε-balls are required, and let $f_1, \ldots, f_n$ be their centres.

(3) Using the fact that the elements $\mathcal{H}$ are uniformly bounded (see point 1), it can be shown that $L_{\mathbf{z}}$ is a Lipschitz function on $\mathcal{H}$, i.e., there is a constant K such that for all $f, g \in \mathcal{H}$, $|L_{\mathbf{z}}(f) - L_{\mathbf{z}}(g)| \leq K\|f - g\|_\infty$. Thus, if $L_{\mathbf{z}}(f_i)$ is small, $L_{\mathbf{z}}(f)$ is small for all f in the ε-ball centred at f_i.

Putting the above points together, it follows that $\{\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| < \varepsilon\}$ with high probability. The precise statement that follows from using the Bernstein inequality is as follows:

$$\Pr\left\{\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| \leq \varepsilon\right\} \geq 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \exp\left[-\frac{m\varepsilon^2}{4(2\sigma^2 + (1/3)M^2\varepsilon)}\right], \tag{4}$$

where $M = \text{ess sup } \sup_{f\in\mathcal{H}} |y - f(x)|$, $\sigma^2 = \sup_{f\in\mathcal{H}} \sigma^2(l_f)$ and $\mathcal{N}(\mathcal{H}, \varepsilon)$, the covering number of $\mathcal{H}$, is the minimum number of balls of radius ε required to cover $\mathcal{H}$. This inequality tells us that the larger the covering number $\mathcal{N}(\mathcal{H}, \varepsilon)$, the harder it is for the empirical error to be uniformly close to the true error over $\mathcal{H}$. Thus the covering number may be interpreted as a measure of the complexity of $\mathcal{H}$. We saw that if $|L_{\mathbf{z}}| < \varepsilon$ uniformly on $\mathcal{H}$, then $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\varepsilon$. It follows that

$$\Pr\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\varepsilon\} \geq 1 - 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) \exp\left[-\frac{m\varepsilon^2}{4(2\sigma^2 + (1/3)M^2\varepsilon)}\right]$$

Note that the confidence term on the right hand side of the above inequality declines as the covering number $\mathcal{N}(\mathcal{H}, \varepsilon)$ increases, i.e., as complexity of $\mathcal{H}$ increases. Equivalently, for a fixed confidence $1 - \delta$, the bound on $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})$ increases with the complexity of $\mathcal{H}$.

Since our ultimate goal is to choose an element of $\mathcal{H}$ which has low error, we are interested in putting a bound on $\mathcal{E}(f_{\mathbf{z}})$ that holds with some high probability $(1 - \delta)$. Note that

$$\mathcal{E}(f_{\mathbf{z}}) = [\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})] + \mathcal{E}(f_{\mathcal{H}}).$$

The term $[\mathcal{E}(f_{\mathbf{Z}}) - \mathcal{E}(f_{\mathcal{H}})]$, which is a random variable, is known as the *sample error*. The term $\mathcal{E}(f_{\mathcal{H}})$ is known as the *approximation error*. The approximation error, for a given hypothesis space $\mathcal{H}$, is the minimum value of error that can be obtained with that hypothesis space. We expect that the larger the hypothesis, the better is the approximation to f_ρ that can be made using elements of that hypothesis space. Thus the approximation error is expected to decrease with the complexity of the hypothesis space. For the sample error, we have seen that for a given level of confidence $(1 - \delta)$, we have a bound on the sample error that increases with the complexity of $\mathcal{H}$.

Now, instead of considering a single hypothesis space, we can consider a nested sequence of hypothesis spaces of increasing complexity,

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \mathcal{H}_2 \subset \dots .$$

As complexity increases, the approximation error decreases, whereas the sample error is expected to increase. Thus, it can be expected that there is an optimal complexity at which we get the lowest possible probabilistic upper bound on the total error $\mathcal{E}(f_{\mathbf{Z}})$. This is the level of complexity that we need to use. This approach to choosing an optimal level of complexity is called *structural risk minimization* (SRM). Vapnik (1998) bases his approach to the learning problem on SRM, with the VC-dimension as the measure of complexity. Vidyasagar (1997) gives another exposition of SRM.

The discussion above was about how, in a given nested sequence of hypothesis spaces, we can choose one of the optimal level of complexity. A separate, but important question is how this hierarchy itself should be chosen. The answer is that we have to choose a hierarchy that is adapted to the particular system that we are modelling. The situation is analogous to choosing a suitable basis for representing a function. For example, for some functions, using a basis of sinusoids is efficient, whereas for others, a wavelet basis might be more efficient.

3. Complexity control via regularization

In the previous section, we considered a scenario in which we had a hierarchy of compact hypothesis spaces, and we had to use the hypothesis space of the optimal complexity. We shall now consider a slightly different version of the learning problem. This time, however, the hypothesis space is not a compact subset of $\mathcal{C}(\mathcal{X})$, but rather a vector subspace of $\mathcal{C}(\mathcal{X})$. We assume that $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, which gives us a norm $\| \cdot \|_{\mathcal{H}}$ on $\mathcal{H}$. We want to find $f \in \mathcal{H}$ which minimizes the error $\mathcal{E}$.

Given empirical data $\mathbf{Z} = (Z_1, \dots, Z_m)$, we cannot compute $\mathcal{E}$, but we can compute $\mathcal{E}_{\mathbf{Z}} := (1/m) \sum_{i=1}^{m} (Y_i - f(X_i))^2$. Our learning algorithm is as follows: We choose $f \in \mathcal{H}$ that minimizes not $\mathcal{E}_{\mathbf{Z}}$, but a function $\phi_{\gamma,\mathbf{Z}} : \mathcal{H} \to \mathbb{R}$ defined by

$$\phi_{\gamma,\mathbf{Z}}(f) = \mathcal{E}_{\mathbf{Z}}(f) + \gamma \| f \|_{\mathcal{H}}^2.$$

Here γ is a positive constant, called the *regularization parameter*. Our solution is $f_{\gamma,\mathbf{Z}} = \arg\min_{f \in \mathcal{H}} \phi_{\gamma,\mathbf{Z}}(f)$. Analogous to the development in §2 , our goal is to make $\mathcal{E}(f_{\gamma,\mathbf{Z}})$, the error associated with the function $f_{\gamma,\mathbf{Z}}$, small.

In §2, we minimized the empirical error over a compact hypothesis space, whereas here, we have a non-compact hypothesis space, but what we are minimizing is the empirical error plus a *complexity penalty* of the form $\gamma \| f \|_{\mathcal{H}}^2$. The norm $\| \cdot \|_{\mathcal{H}}$ is chosen in a way that allows us to interpret $\| f \|_{\mathcal{H}}^2$ as a measure of the complexity of a function f. For example, by using a

Sobolev space norm, which penalizes the derivatives of f, functions that are less smooth would be penalized more. The size of this complexity penalty is controlled by constant γ. Note that as γ increases, the more likely is it that our algorithm yields a function f of small norm. Thus, loosely speaking, by increasing γ, we decrease the effective size of the hypothesis space. In §2, we used SRM to choose a hypothesis space of the optimal complexity. Analogously, in this setting our goal is to choose an optimal value of the regularization parameter γ.

Formulating the learning problem in the manner we have done above has two important advantages: First, the presence of the complexity penalty makes the determination of the function $f_{\gamma,\mathbf{z}}$ a numerically well-posed problem. Second, we can often even get explicit expressions for $f_{\gamma,\mathbf{z}}$. These features are illustrated in the examples that we work out later in this section.

We define another function $\phi_\gamma : \mathcal{H} \to \mathbb{R}$ by

$$\phi_\gamma(f) = \int_{\mathcal{Z}} (y - f(x))^2 \rho(\mathrm{d}z) + \gamma \|f\|_{\mathcal{H}}^2 = \|f - f_\rho\|^2 + \gamma \|f\|_{\mathcal{H}}^2 + \sigma_\rho^2,$$

and we let $f_\gamma = \arg\min_{f\in\mathcal{H}} \phi_\gamma(f) = \arg\min_{f\in\mathcal{H}} \|f - f_\rho\|^2 + \gamma \|f\|_{\mathcal{H}}^2$. ϕ_γ and f_γ cannot be computed, since the underlying probability distribution ρ is unknown.

We can bound $\mathcal{E}(f_{\gamma,\mathbf{z}})$ as follows:

$$\mathcal{E}(f_{\gamma,\mathbf{z}}) \leq |\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)| + \mathcal{E}(f_\gamma).$$

We now define the sample error as $|\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)|$, and the approximation error as $\mathcal{E}(f_\gamma)$. Note that the sample error is a random variable. We fix some confidence level $(1-\delta)$, and we also fix the number of observations m. We would as before like to have a bound on the sample error that holds with some high probability $(1-\delta)$, and we would also like to have a bound on the approximation error. We then choose a value of γ that minimizes the sum of these bounds.

This is a general framework, applicable to different hypothesis spaces. We now present some theory that is applicable in this general framework. In the next two subsections we give specific details about two examples. In the first, the hypothesis space is finite dimensional, and in the second it is a (possibly infinite dimensional) *Reproducing Kernel Hilbert Space (RKHS)*. The former case is sometimes termed as "Ridge Regression". We give a more detailed description of this case. For the latter case, we give an outline, and refer the reader to Cucker & Smale (2002a) for some of the details.

Owing to the complexity penalty $\gamma \|f\|_{\mathcal{H}}^2$ in the definitions of $\phi_{\gamma,\mathbf{z}}$ and ϕ_γ, it is expected that $\|f_{\gamma,\mathbf{z}}\|_{\mathcal{H}}$ and $\|f_{\gamma,}\|_{\mathcal{H}}$ would decrease with increasing γ. We will assume that there is a closed ball $B(\gamma)$ of radius r_γ in $\mathcal{H}$ to which both $f_{\gamma,\mathbf{z}}$ and f_γ must belong. We further assume that $r_\gamma \leq C/\gamma$ for some constant C, and that $B(\gamma)$ is compact (either in $\mathcal{H}$ or as a subset of $\mathcal{C}(\mathcal{X})$). These assumptions hold in the examples that we discuss.

Bounding the sample error: It is not possible to directly use the theory of §2, since that relies on the compactness of the hypothesis space. However, the compactness of $B(\gamma)$ enables us to reduce to the case of compact hypothesis space and use previous results. The approach to bounding the sample error that we give below differs from that given in Cucker & Smale (2002a)[1]. Note that

$$\phi_{\gamma,\mathbf{z}}(f) - \phi_\gamma(f) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) = L_{\mathbf{z}}(f)$$

[1]The argument given in Cucker & Smale (2002a) is not valid due to a problem with the proof of proposition 2 on page 422: the norm of an $n \times n$ matrix of the form $(I + K)^{-1}$, where K is positive semi-definite, induced by the max norm on $\mathbb{R}^n$, is not necessarily less than 1

As in (4) (using the Hoeffding rather than the Bernstein inequality),

$$\Pr\left\{\sup_{f\in B(\gamma)}|L_{\mathbf{Z}}(f)|\leq\varepsilon\right\}\geq 1-2\mathcal{N}\left(B(\gamma),\frac{\varepsilon}{8M_\gamma}\right)\exp\left[-\left(\frac{m\varepsilon^2}{2M_\gamma^4}\right)\right].$$

Here $M_\gamma := \text{ess sup}\,[\sup_{f\in B(\gamma)}|y-f(x)|]$. Fix a confidence $(1-\delta)$. How large must ε be? We need

$$2\mathcal{N}(B(\gamma),(\varepsilon/8M_\gamma))\exp[-m\varepsilon^2/(2M_\gamma^4)]\leq\delta.$$

Assume we have an estimate $\mathcal{N}(B(\gamma),(\varepsilon/8M_\gamma))\leq E(\gamma,\varepsilon)$, where E is a decreasing function of γ and also a decreasing function of ε. For the hypothesis spaces that we consider, we may not be able to compute the covering numbers, but we have such estimates. It is enough that ε satisfy

$$2E(\gamma,\varepsilon)\exp[-(m\varepsilon^2)/(2M_\gamma^4)]\leq\delta.$$

This gives us a lower bound $\varepsilon_0(m,\gamma)$ on ε, which can be obtained by solving

$$2E(\gamma,\varepsilon)\exp[-(m\varepsilon^2)/(2M_\gamma^4)]=\delta.$$

As in (2), we get that $|\phi_\gamma(f_{\gamma,\mathbf{Z}})-\phi_\gamma(f_\gamma)|\leq 2\varepsilon_0$ with confidence $(1-\delta)$. As γ increases, r_γ and M_γ decrease, and $E(\gamma,\varepsilon)$ should decrease. Also, $E(\gamma,\varepsilon)$ should decrease with ε. Thus, the left hand side of the previous equation decreases as γ increases, and also decreases as ε increases. Thus, for fixed δ and m, ε_0 increases as γ decreases, i.e., as complexity increases.

We now derive a bound on the sample error. We first need a lemma:

Lemma 1. *For all* $f\in\mathcal{H}$,

$$\|f-f_\gamma\|^2+\gamma\|f-f_\gamma\|_{\mathcal{H}}^2\leq\phi_\gamma(f)-\phi_\gamma(f_\gamma)$$

Proof. By definition of f_γ, for all $f\in\mathcal{H}$ and for all $\lambda\in[0,1]$,

$$\phi_\gamma(\lambda f+(1-\lambda)f_\gamma)\geq\phi_\gamma(f_\gamma).$$

Recall

$$\phi_\gamma(f)=\mathcal{E}(f)+\gamma\|f\|_{\mathcal{H}}^2=\|f-f_\rho\|^2+\sigma_\rho^2+\gamma\|f\|_{\mathcal{H}}^2.$$

Thus we get

$$\|\lambda f+(1-\lambda)f_\gamma-f_\rho\|^2+\gamma\|\lambda f+(1-\lambda)f_\gamma\|_{\mathcal{H}}^2\geq\|f_\gamma-f_\rho\|^2+\gamma\|f_\gamma\|_{\mathcal{H}}^2.$$

After some manipulation, this gives us

$$\lambda^2[\|f-f_\gamma\|^2+\gamma\|f-f_\gamma\|_{\mathcal{H}}^2]+2\lambda[\langle f-f_\gamma,f_\gamma-f_\rho\rangle+\gamma\langle f-f_\gamma,f_\gamma\rangle_{\mathcal{H}}]$$
$$\geq 0 \text{ for all } \lambda\in[0,1].$$

Thus, the coefficient of 2λ must be non-negative. Therefore,

$$\begin{aligned}\|f - f_\rho\|^2 + \gamma\|f\|_{\mathcal{H}}^2 &= \|(f - f_\gamma) + (f_\gamma - f_\rho)\|^2 + \gamma\|(f - f_\gamma) + f_\gamma\|_{\mathcal{H}}^2 \\ &= \|f - f_\gamma\|^2 + \|f_\gamma - f_\rho\|^2 + \gamma\|f - f_\gamma\|_{\mathcal{H}}^2 + \gamma\|f_\gamma\|_{\mathcal{H}}^2 \\ &\quad + 2[\langle f - f_\gamma, f_\gamma - f_\rho\rangle + \gamma\langle f - f_\gamma, f_\gamma\rangle_{\mathcal{H}}] \\ &\geq \|f - f_\gamma\|^2 + \|f_\gamma - f_\rho\|^2 + \gamma\|f - f_\gamma\|_{\mathcal{H}}^2 + \gamma\|f_\gamma\|_{\mathcal{H}}^2.\end{aligned}$$

Now, rearranging terms yields the inequality in the statement of this lemma. ■

Now, we need to get a bound on the sample error $|\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)|$. Since $\phi_\gamma(f) = \mathcal{E}(f) + \gamma\|f\|_{\mathcal{H}}^2$,

$$\begin{aligned}\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma) &= \phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma) + \gamma[\|f_\gamma\|_{\mathcal{H}}^2 - \|f_{\gamma,\mathbf{z}}\|_{\mathcal{H}}^2] \\ &= \phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma) + \gamma[\|f_\gamma\|_{\mathcal{H}} + \|f_{\gamma,\mathbf{z}}\|_{\mathcal{H}}] \\ &\quad \times [\|f_\gamma\|_{\mathcal{H}} - \|f_{\gamma,\mathbf{z}}\|_{\mathcal{H}}] \\ &\leq \phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma) + 2\gamma r_\gamma\|f_{\gamma,\mathbf{z}} - f_\gamma\|_{\mathcal{H}}.\end{aligned}$$

By the above lemma, $\gamma\|f - f_\gamma\|_{\mathcal{H}}^2 \leq \phi_\gamma(f) - \phi_\gamma(f_\gamma)$. Thus

$$\|f_{\gamma,\mathbf{z}} - f_\gamma\|_{\mathcal{H}} \leq \gamma^{-1/2}(\phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma))^{1/2}.$$

Also, $r_\gamma \leq C/\gamma$. It thus follows that

$$|\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)| \leq [\phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma)] + 2C\gamma^{-1/2}(\phi_\gamma(f_{\gamma,\mathbf{z}}) - \phi_\gamma(f_\gamma))^{1/2}.$$

This gives the following probabilistic bound for the sample error:

Theorem 1. *With confidence* $(1 - \delta)$*, the sample error* $|\mathcal{E}(f_{\gamma,\mathbf{z}}) - \mathcal{E}(f_\gamma)|$ *is bounded by*

$$S(\gamma) = 2\varepsilon_0(m, \gamma) + 2C\gamma^{-1/2}(2\varepsilon_0(m, \gamma))^{1/2}, \tag{5}$$

where $\varepsilon_0(m, \gamma)$ *is obtained by solving*

$$2E(\gamma, \varepsilon)\exp[-(m\varepsilon^2)/(2M_\gamma^4)] = \delta, \tag{6}$$

and $E(\gamma, \varepsilon)$ *is an upper bound for the covering number* $\mathcal{N}(B(\gamma), (\varepsilon/8M_\gamma))$.

Since $\varepsilon_0(m, \gamma)$ increases with decreasing γ, it is clear that the sample error bound $S(\gamma)$ increases with decreasing γ, i.e., with increasing complexity.

Bounding the approximation error: The approximation error is $\mathcal{E}(f_\gamma) = \|f_\gamma - f_\rho\|^2 + \sigma_\rho^2$, where

$$f_\gamma = \arg\min_{f\in\mathcal{H}}\|f - f_\rho\|^2 + \gamma\|f\|_{\mathcal{H}}^2.$$

Since we are interested in how $\mathcal{E}(f_\gamma)$ varies with γ, the constant term σ_ρ^2 can be dropped. Thus it is enough to get a bound $A(\gamma)$ on $\|f_\gamma - f_\rho\|^2$.

By setting to zero the functional derivative w.r.t. f of the expression $\|f - f_\rho\|^2 + \gamma \|f\|^2_{\mathcal{H}}$, it is possible to get an expression for f_γ in terms of f_ρ. This expression is then used to get the bound $A(\gamma)$. This is illustrated in the examples that follow later in this section.

Getting the optimal complexity and the approximation to f_ρ: We saw that the sample error bound $S(\gamma)$ increases with increasing complexity. Also, it is clear that the the approximation error must decrease with increasing complexity. The optimum value of γ, which corresponds to the optimum complexity, is that which minimizes the quantity $S(\gamma) + A(\gamma)$. Finally, using the optimum value of γ, we calculate $f_{\gamma,\mathbf{Z}}$, which is our approximation to the regression function f_ρ. We are more specific in the examples, which is what we proceed to next.

3.1 *Ridge regression*

Consider the situation where $\mathcal{H}$ is an n-dimensional subspace of $\mathcal{C}(\mathcal{X})$, equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Our goal is to choose, using empirical data $\mathbf{Z} = (Z_1, \ldots, Z_m)$, a function $f \in \mathcal{H}$ which has the least possible error $\mathcal{E}(f)$.

As above, for any positive γ, we define

$$\phi_{\gamma,\mathbf{Z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (Y_i - f(X_i))^2 + \gamma \|f\|^2_{\mathcal{H}}.$$

We choose as our solution $f_{\gamma,\mathbf{Z}} := \arg\min_{f \in \mathcal{H}} \phi_{\gamma,\mathbf{Z}}(f)$. In Statistics, this is called "ridge regression". An exposition of ridge regression may be found in Hastie *et al* (2001).

We go through the process of determining the optimal γ. The steps are as follows: First we need r_γ, the radius in $\mathcal{H}$ of a ball $B(\gamma)$ containing both $f_{\gamma,\mathbf{Z}}$ and $f_\gamma := \arg\min_{f \in \mathcal{H}} \phi_\gamma(f)$. From this, we obtain $M_\gamma := \text{ess sup}\,[\sup_{f \in B(\gamma)} |y - f(x)|]$. The next step is to determine an upper bound, $E(\gamma, \varepsilon)$, for the covering number $\mathcal{N}(B(\gamma), (\varepsilon/8M_\gamma))$. $\varepsilon_0(m, \gamma)$ is obtained by solving

$$2E(\gamma, \varepsilon) \exp[-(m\varepsilon^2)/(2M_\gamma^4)] = \delta,$$

where $1 - \delta$ is the desired confidence. Finally, we obtain the bounds $S(\gamma)$ and $A(\gamma)$, and choose the γ that minimizes the sum $S(\gamma) + A(\gamma)$. We proceed now to give an outline of these steps for the current example.

Let us first determine $f_{\gamma,\mathbf{Z}}$. Let $\{\varphi_1, \ldots, \varphi_n\}$ be a basis for $\mathcal{H}$ that is orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on $\mathcal{H}$, and let $f = \sum_{i=1}^n c_i \varphi_i$. Let $c := (c_1, \ldots, c_n) \in \mathbb{R}^n$. Then $\|f\|^2_{\mathcal{H}} = \|c\|_2^2$. Note that

$$\phi_{\gamma,\mathbf{Z}}(f) = \frac{1}{m} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{n} c_j \varphi_j(X_i) \right)^2 + \gamma \|c\|_2^2 = \frac{1}{m} \|\mathbf{Y} - \Phi c\|^2 + \gamma \|c\|_2^2,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_m)$, and Φ is an $m \times n$ matrix with $\Phi_{ij} = \varphi_j(x_i)$. If $f_{\gamma,\mathbf{Z}} = \sum_{i=1}^n c_{\gamma,\mathbf{Z},i} \varphi_i$, then the coefficient vector $c_{\gamma,\mathbf{Z}} := (c_{\gamma,\mathbf{Z},1} \ldots, c_{\gamma,\mathbf{Z},n})$ minimizes $(1/m)\|\mathbf{Y} - \Phi c\|^2 + \gamma \|c\|_2^2$. It is easy to show that

$$c_{\gamma,\mathbf{Z}} = (\Phi^T \Phi + \gamma m I)^{-1} \Phi^T \mathbf{Y}.$$

As explained earlier (just before (3)),

$$|Y_j| \leq |Y_j - f_\rho(X_j)| + |f_\rho(X_j)| \leq M_1$$

With this, $\|\Phi^T \mathbf{Y}\|_2 \leq \sqrt{n} m M_1 \max_i(\|\varphi_i\|_\infty)$. Together with the fact that $\|(\Phi^T\Phi + \gamma m I)^{-1}\| \leq 1/(\gamma m)$, we get

$$\|f_{\gamma,\mathbf{z}}\|_{\mathcal{H}} = \|c_{\gamma,\mathbf{z}}\|_2 \leq K/\gamma, \tag{7}$$

where $K = \sqrt{n} M_1 \max_i(\|\varphi_i\|_\infty)$.

The next step is to determine $f_\gamma = \arg\min_{f\in\mathcal{H}} \|f - f_\rho\|^2 + \gamma\|f\|^2_{\mathcal{H}}$. As before, let $f = \sum_{i=1}^n c_i\varphi_i$. Equating the derivative with respect to f to zero, we get that $f_\gamma = \sum_{i=1}^n c_{\gamma,i}\varphi_i$, with the coefficient vector c_γ being given by

$$c_\gamma = (\gamma I + M)^{-1} b.$$

Here $M_{ij} = \langle\varphi_i, \varphi_j\rangle_{L^2_\rho(X)}$. Let us assume that the basis $\{\varphi_i\}$ is chosen such that M is diagonal. b_i is given by $\langle\varphi_i, f_\rho\rangle_{L^2_\rho(X)}$. Clearly,

$$|b_i| \leq \|\varphi_i\|_\infty \|f_\rho\|_\infty \leq (\max_i \|\varphi_i\|_\infty)\|f_\rho\|_\infty.$$

Thus,

$$\|f_\gamma\|_{\mathcal{H}} = \|c_\gamma\|_2 \leq \|(\gamma I + M)^{-1}\|\|b\|_2 \leq [\sqrt{n}(\max_i \|\varphi_i\|_\infty)\|f_\rho\|_\infty]\gamma \leq K/\gamma.$$

The conclusion is that both f_γ and $f_{\gamma,\mathbf{z}}$ are in a closed ball of radius $r_\gamma = K/\gamma$ in $\mathcal{H}$. For all f in this ball we have

$$\|f\|_\infty = \left\|\sum_i c_i\varphi_i\right\|_\infty \leq n(\max_i \|\varphi_i\|_\infty)\|c\|_2 = n(\max_i \|\varphi_i\|_\infty)\|f\|_{\mathcal{H}}$$
$$\leq n(\max_i \|\varphi_i\|_\infty) r_\gamma \leq [n^{3/2}(\max_i \|\varphi_i\|_\infty)^2 M_1]/\gamma.$$

As in (3), we have

$$|Y - f(X)| \leq M_1 + [n^{3/2}(\max_i \|\varphi_i\|_\infty)^2 M_1]/\gamma = M_1(1 + (A/\gamma)) =: M_\gamma,$$

where $A := n^{3/2}(\max_i \|\varphi_i\|_\infty)^2$.

We now determine an upper bound $E(\gamma, \varepsilon)$ for the covering number $\mathcal{N}(B(\gamma), (\varepsilon/8M_\gamma))$. We saw that if $f \in B(\gamma)$ then $\|f\|_\infty \leq n(\max_i \|\varphi_i\|_\infty) r_\gamma$. It is a standard result (see Cucker & Smale 2002b) that in an n-dimensional normed linear space, the number of a-balls required to cover the ball of radius r is upper bounded by $(4r/a)^n$. Thus,

$$\mathcal{N}(B(\gamma), (\varepsilon/8M_\gamma)) \leq ([4n(\max_i \|\varphi_i\|_\infty) r_\gamma]/[\varepsilon/8M_\gamma])^n =: E(\gamma, \varepsilon).$$

Now, ε_0 can be determined by solving

$$2E(\gamma, \varepsilon)\exp[-(m\varepsilon^2)/(2M_\gamma^4)] = \delta.$$

While we do not have an explicit expression for ε_0, it is not difficult to determine it numerically. Then, we can use (5) to get the sample error bound $S(\gamma)$.

Next, we need to obtain a bound which expresses the variation with respect to γ of the approximation error $\mathcal{E}(f_\gamma)$. If $\bar{f}_\rho$ is the projection of f_ρ on $\mathcal{H}$,

$$\mathcal{E}(f_\gamma) = \|f_\gamma - f_\rho\|^2 + \sigma_\rho^2 = \|f_\gamma - \bar{f}_\rho\|^2 + \|\bar{f}_\rho - f_\rho\|^2 + \sigma_\rho^2.$$

It is only the term $\|f_\gamma - \bar{f}_\rho\|$ that depends on γ, so we need only obtain a bound for that term. Note that

$$\begin{aligned} f_\gamma &= \arg\min_{f\in\mathcal{H}}\{\|f_\gamma - f_\rho\|^2 + \gamma\|f\|_{\mathcal{H}}^2\} \\ &= \arg\min_{f\in\mathcal{H}}\{\|f_\gamma - \bar{f}_\rho\|^2 + \gamma\|f\|_{\mathcal{H}}^2\}. \end{aligned}$$

Now, let $f_\gamma = \sum_i c_{\gamma,i}\varphi_i$, and $\bar{f}_\rho = \sum_i \bar{c}_i\varphi_i$. Let c_γ and $\bar{c}$ be the corresponding coefficient vectors. Thus,

$$c_\gamma = \arg\min_{c\in\mathbb{R}^n}\{(c-\bar{c})^T M(c-\bar{c}) + \gamma\|c\|_2^2\}. \tag{8}$$

We had assumed without loss of generality that the matrix M is diagonal. Let $M = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Note that $\|\sum_{i=1}^n c_i\varphi_i\|^2 = c^T Mc = \sum_{i=1}^n c_i^2\lambda_i$. Differentiating with respect to c the expression to be minimized in equation (8), we get

$$c_{\gamma,i} - \bar{c}_i = -\gamma\bar{c}_i/(\lambda_i + \gamma).$$

Thus,

$$\begin{aligned} \|f_\gamma - \bar{f}_\rho\|^2 &= (c_\gamma - \bar{c})^T M(c_\gamma - \bar{c}) \\ &= \sum_i \frac{\gamma^2\bar{c}_i^2\lambda_i}{(\lambda_i+\gamma)^2} \le \left(\max_i \frac{\gamma^2}{(\lambda_i+\gamma)^2}\right)\|\bar{f}_\rho\|^2 \\ &= \left(\frac{\gamma}{(\min_i \lambda_i + \gamma)}\right)^2 \|f_\rho\|^2 \le \left(\frac{\gamma}{(\min_i \lambda_i + \gamma)}\right)^2 \|f_\rho\|_\infty^2. \end{aligned}$$

Therefore, we can define

$$A(\gamma) := \left(\frac{\gamma}{\min_i \lambda_i + \gamma}\right)^2 B^2,$$

where B is an upper bound for $\|f_\rho\|_\infty$. As described above, the value of γ is chosen so as to minimize the sum $S(\gamma) + A(\gamma)$ of the upper bounds on the sample error and approximation error. We have the above explicit expression of $A(\gamma)$, whereas $S(\gamma)$ can easily be determined numerically. Choosing the optimal γ corresponds to using a hypothesis space of the optimal complexity.

3.2 *Learning with reproducing kernel hilbert spaces*

We saw that in the regularization approach, we have a subspace $\mathcal{H} \subset \mathcal{C}(\mathcal{X})$, and we need to find the function $f_{\gamma,\mathbf{z}} \in \mathcal{H}$ that minimizes

$$\phi_{\gamma,\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m (Y_i - f(X_i))^2 + \gamma\|f\|_{\mathcal{H}}^2.$$

$\mathcal{H}$ should be chosen to be an adequately rich space of functions for which the task for determining the function $f_{\gamma,\mathbf{z}}$ is easy. It turns out that our purposes are well served by the use of a *reproducing kernel Hilbert space* associated with a *Mercer Kernel*. We do not prov all the

results stated in this section. Unless stated otherwise, the proofs may be found in Cucker & Smale (2002b,a).

A Mercer kernel is a continuous, symmetric and positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Here positive definiteness means that for any $x_1, \ldots, x_n \in \mathcal{X}$, the symmetric $n \times n$ matrix $K[\mathbf{x}]$ whose i, j-th entry is $K(x_i, x_j)$ is positive semi-definite. As an example, the function $K(x, t) = \exp -(\|x - t\|^2)/c^2$ is a Mercer kernel on subsets of $\mathbb{R}^n$.

For a Mercer kernel K, we have an operator $L_K : L^2_\rho(\mathcal{X}) \to L^2_\rho(\mathcal{X})$ defined by

$$L_K(f)(x) = \int_{\mathcal{X}} K(x, t) f(t) \rho(dt).$$

It can be shown that L_K is symmetric, positive definite and compact.

Given a Mercer kernel K, and $x \in \mathcal{X}$, let $K_x : \mathcal{X} \to \mathbb{R}$ be defined by $K_x(y) := K(x, y)$. Let H_0 be the subspace $\text{span}\{K_x | x \in \mathcal{X}\}$ of $\mathcal{C}(\mathcal{X})$. It can be shown that there is a well-defined inner product $\langle \cdot, \cdot \rangle$ on H_0 such that $\langle K_x, K_y \rangle = K(x, y)$.

This inner product has an interesting *reproducing property*: If $f \in H_0$, $f = \sum_{i=1}^n a_i K_{x_i}$, then

$$\begin{aligned} \langle f, K_y \rangle &= \left\langle \sum_{i=1}^n a_i K_{x_i}, K_y \right\rangle = \sum_{i=1}^n a_i \langle K_{x_i}, K_y \rangle = \sum_{i=1}^n a_i K(x_i, y) \\ &= \sum_{i=1}^n a_i K_{x_i}(y) = f(y). \end{aligned}$$

Thus the function K_y is like a δ-function at y.

The reproducing kernel Hilbert space $\mathcal{H}$ associated with the kernel K is defined as the completion of H_0 (w.r.t. the inner product of H_0). The reproducing property in the above equation holds for all $f \in \mathcal{H}$. $\mathcal{H}$ is a subspace of $\mathcal{C}(\mathcal{X})$, and the inclusion $\mathcal{H} \to \mathcal{C}(\mathcal{X})$ is bounded. It is a fact that $\mathcal{H}$ is the image of $L_K^{1/2}$, and that $L_K^{1/2} : L^2_\rho(\mathcal{X}) \to \mathcal{H}$ is a Hilbert space isomorphism.

Consider now the function

$$\phi_{\gamma,\mathbf{Z}}(f) := \frac{1}{m} \sum_{i=1}^m V(Y_i, f(X_i)) + \gamma \|f\|^2_{\mathcal{H}}. \tag{9}$$

Note that we have replaced $(Y_i - f(X_i))^2$ be a more general loss $V(Y_i, f(X_i))$ which is not assumed to be differentiable.

We show that $f_{\gamma,\mathbf{Z}}$ is a linear combination of the functions K_{x_i}. The following nice argument is due to Schölkopf & Smola (2002): We write $\mathcal{H} = \mathcal{H}_\| \oplus \mathcal{H}_\perp$, where $\mathcal{H}_\|$ is the space spanned by $\{K_{x_i} \mid i = 1, \ldots, m\}$, and $\mathcal{H}_\perp$ is its orthogonal complement. Any $f \in \mathcal{H}$ can correspondingly be expressed as $f = f_\| + f_\perp$. Note that $f_\perp(x_i) = \langle f_\perp, K_{x_i} \rangle = 0$. Thus $f(x_i) = f_\|(x_i)$. If

$$\mathcal{E}_{\mathbf{Z}}(f) := \frac{1}{m} \sum_{i=1}^m V(Y_i, f(X_i)),$$

$\mathcal{E}_{\mathbf{Z}}(f) = \mathcal{E}_{\mathbf{Z}}(f_\|)$. Thus,

$$\phi_{\gamma,\mathbf{Z}}(f) = \mathcal{E}_{\mathbf{Z}}(f_\|) + \gamma(\|f_\|\|^2_{\mathcal{H}} + \|f_\perp\|^2_{\mathcal{H}}).$$

This makes it clear that if f minimizes $\phi_{\gamma,\mathbf{Z}}$, then $f_\perp$ must be zero. This means that $f_{\gamma,\mathbf{Z}}$ must be an element of $\mathcal{H}_\parallel$, i.e., a linear combination of the functions K_{x_i}. In general, finding the coefficients in this linear combination requires solving a convex optimization problem. With the further assumption that V is differentiable, we now show how the reproducing property allows us obtain explicit expressions for the coefficients of the minimizer $f_{\gamma,\mathbf{Z}}$.

Since $f(x_i) = \langle f, K_{x_i}\rangle$, the functional derivative of $\phi_{\gamma,\mathbf{Z}}$ with respect to f is:

$$\frac{\delta\phi_{\gamma,\mathbf{Z}}}{\delta f} = \frac{1}{m}\sum_{i=1}^{m} D_2V(y_i,\, f(x_i))K_{x_i} + 2\gamma f.$$

$f_{\gamma,\mathbf{Z}}$ obtained by setting this equal to zero, and solving for f. Thus, we get that

$$f_{\gamma,\mathbf{Z}} = \sum_{i=1}^{m} a_{\gamma,\mathbf{Z},i}K_{x_i}, \text{ where } a_{\gamma,\mathbf{Z},i} = -\frac{D_2V(y_i,\, f_{\gamma,\mathbf{Z}}(x_i))}{2m\gamma}.$$

Plugging the first expression above into the second gives us

$$a_{\gamma,\mathbf{Z},i} = -\frac{D_2V(y_i, \sum_{j=1}^{m} a_{\gamma,\mathbf{Z},j}K(x_i, x_j))}{2m\gamma},$$

which can be solved for the coefficients $a_{\gamma,\mathbf{Z},i}$. In the special case, where the loss $V(Y_i,\, f(X_i))$ is actually the square loss $(Y_i - f(X_i))^2$, this equation becomes

$$a_{\gamma,\mathbf{Z},i} = -\frac{Y_i - \sum_{j=1}^{m} a_{\gamma,\mathbf{Z},j}K(x_i, x_j)}{m\gamma},$$

which can be rewritten in vector form as:

$$(m\gamma I + K[\mathbf{x}])a_{\gamma,\mathbf{Z}} = \mathbf{Y}$$

Thus, we have an easy way to explicitly compute $f_{\gamma,\mathbf{Z}}$ in the case where V is differentiable. Even where V is not differentiable, we saw that $f_{\gamma,\mathbf{Z}}$ is of the form $f = \sum_{i=1}^{n} a_i K_{x_i}$, in which the coefficients a_i are obtainable by solving an optimization problem. However, the theory we have described for choosing an optimal regularization parameter assumes that the loss function V has the specific form $V(y,\, f(x)) = (y - f(x))^2$.

Recall that $f_\gamma = \arg\min_{f\in\mathcal{H}} \|f - f_\rho\|^2_{L^2_\rho(\mathcal{X})} + \gamma\|f\|^2_{\mathcal{H}}$. By setting to zero the functional derivative with respect to f of the quantity $\|f - f_\rho\|^2_{L^2_\rho(\mathcal{X})} + \gamma\|f\|^2_{\mathcal{H}}$, we can show that

$$f_\gamma = (I + \gamma L_K^{-1})^{-1} f_\rho.$$

As described at the beginning this section, determining the optimal level of complexity is tantamount to determining the optimal value of the parameter γ. The steps towards determining the optimal γ were also described earlier. In the present setting, they are as follows: Having obtained expressions for $f_{\gamma,\mathbf{Z}}$ and f_γ, we determine the radius r_γ of the ball $B(\gamma) \subset \mathcal{H}$, and then $M_\gamma := \text{ess sup}\sup_{f\in\mathcal{H}}|Y - f(X)|$. The next steps are to obtain the bound $E(\gamma, \varepsilon)$, find $\varepsilon_0(m, \gamma)$, and get the bounds $S(\gamma)$ and $A(\gamma)$ on the sample and approximation errors. Finally, the γ that minimizes $S(\gamma) + A(\gamma)$ represents the optimal complexity. Except for the determination of the sample error bound $S(\gamma)$, for which the procedure was described in theorem 1, the details of these steps may be found in Cucker & Smale (2002a).

Links to the support vector machine methodology: It is interesting that the support vector machine regression and classification methodologies can be viewed as regularized optimization problems of the type described in (9). Support vector machine regression corresponds to using the *ϵ-insensitive loss function*

$$V(y, f(x)) := \max\{|y - f(x)| - \epsilon, 0\},$$

whereas support vector machine classification corresponds to using the loss function

$$V(y, f(x)) = (1 - yf(x))_+.$$

In the case of classification, the experimental outputs Y_i take values in the set $\{-1, +1\}$, and the label assigned to an $x \in \mathcal{X}$ is determined by the sign of the function $f_{\gamma,\mathbf{Z}}$. Our theory for choosing the optimal value of the regularization parameter γ depends upon V having the specific form $V(y, f(x)) = (y - f(x))^2$, so it cannot be applied to support vector machines. Other means, such as cross-validation, are required for choosing the best regularization parameter. This theme is developed in Evgeniou *et al* (2000), Schölkopf & Smola (2002) and Vapnik (1998).

4. Conclusions and future directions

In this paper we have described a clear and simple framework for learning theory, which we use for controlling complexity by the method of regularization. We have looked at applications to ridge regression and to learning with RKHS's. We also briefly outlined the connection between support vector machines and complexity control by regularization.

Let us now give a brief summary of the learning procedure that we have outlined in §3. Given empirical data $\mathbf{Z} = (Z_1, \ldots, Z_m)$, we choose for our solution the function $f_{\gamma,\mathbf{Z}}$ in $\mathcal{H}$ that minimizes the function $\phi_{\gamma,\mathbf{Z}} : \mathcal{H} \to \mathbb{R}$ defined by

$$\phi_{\gamma,\mathbf{Z}}(f) = \mathcal{E}_{\mathbf{Z}}(f) + \gamma \|f\|^2_{\mathcal{H}},$$

where $\mathcal{E}_{\mathbf{Z}} := \frac{1}{m} \sum_{i=1}^m (Y_i - f(X_i))^2$. Here γ is a positive constant, called the *regularization parameter*. Our goal is to make $\mathcal{E}(f_{\gamma,\mathbf{Z}})$, the error associated with the function $f_{\gamma,\mathbf{Z}}$, small. We bound $\mathcal{E}(f_{\gamma,\mathbf{Z}})$ as follows:

$$\mathcal{E}(f_{\gamma,\mathbf{Z}}) \leq |\mathcal{E}(f_{\gamma,\mathbf{Z}}) - \mathcal{E}(f_\gamma)| + \mathcal{E}(f_\gamma),$$

where $f_\gamma = \arg\min_{f \in \mathcal{H}} \|f - f_\rho\|^2 + \gamma \|f\|^2_{\mathcal{H}}$. $|\mathcal{E}(f_{\gamma,\mathbf{Z}}) - \mathcal{E}(f_\gamma)|$ is called the sample error, and $\mathcal{E}(f_\gamma)$ is called the approximation error. For a confidence level $(1 - \delta)$, we derive a bound $S(\gamma)$ on the sample error that holds with some high probability $(1 - \delta)$, and we also derive a bound $A(\gamma)$ on the approximation error. We then choose a value of γ that minimizes the sum of these bounds. This amounts to choosing the optimal level of complexity.

There are many important and interesting questions that remain to be fully explored. Some of them are outlined below.

Make links with the Akaike information-theoretic criterion (AIC): The setting for the AIC is as follows (see Burnham & Anderson (2002) and Sin & White (1996)): ρ is a distribution on $\mathcal{Z}$ (which describes our system) having density f and $\{g(\cdot|\theta) \mid \theta \in \Theta\}$ is a parameterized family of densities on $\mathcal{Z}$.

We want to choose a parameter $\theta \in \Theta$ which minimizes the Kullback–Leibler distance between f and $g(\cdot|\theta)$:

$$D(f\|g(\cdot|\theta)) = \int_{\mathcal{Z}} f \log(f/g).$$

Equivalently, we want to minimize the "error"

$$\mathcal{E}(\theta) := -E[\log g(\cdot|\theta)].$$

$\mathcal{E}(\theta)$ is the expected value of the loss function $l_\theta(z) := -\log g(z|\theta)$. Let θ_0 be the minimizer of $\mathcal{E}(\theta)$.

Given data $\mathbf{Z} = (Z_1, \ldots, Z_m)$, the maximum likelihood estimate $\theta_{\mathbf{Z}}$ is the minimizer of the empirical error $\mathcal{E}_{\mathbf{Z}}$:

$$\begin{aligned}\theta_{\mathbf{Z}} &= \arg\min_{\theta\in\Theta}\mathcal{E}_{\mathbf{Z}}(\theta)\\ &= \arg\min_{\theta\in\Theta}\left\{-\frac{1}{m}\sum_{i=1}^{m}\log g(Z_i|\theta)\right\}.\end{aligned}$$

We want to get an upper bound for

$$\mathcal{E}(\theta_{\mathbf{Z}}) = \underbrace{[\mathcal{E}(\theta_{\mathbf{Z}}) - \mathcal{E}(\theta_0)]}_{\text{Sample Error}} + \underbrace{\mathcal{E}(\theta_0)}_{\text{Approx Error}}.$$

The sample error is bounded as follows:

$$\begin{aligned}\mathcal{E}(\theta_{\mathbf{Z}}) - \mathcal{E}(\theta_0) \leq\ & (\mathcal{E}(\theta_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(\theta_{\mathbf{Z}}))\\ &+ (\mathcal{E}_{\mathbf{Z}}(\theta_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(\theta_0))\\ &+ (\mathcal{E}_{\mathbf{Z}}(\theta_0) - \mathcal{E}(\theta_0)).\end{aligned}$$

The second term is non-positive, and the third term is easy to bound by a Law of Large numbers argument. So, to bound the sample error it is actually not necessary to have $\mathcal{E} - \mathcal{E}_{\mathbf{Z}}$ to be uniformly small on Θ. We are only interested in $\mathcal{E}(\theta_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(\theta_{\mathbf{Z}})$ being small.

Note that bounding $\Delta := \mathcal{E}(\theta_{\mathbf{Z}}) - \mathcal{E}_{\mathbf{Z}}(\theta_{\mathbf{Z}})$ is not so easy because $\theta_{\mathbf{Z}}$ is a random variable. But this exactly the issue is addressed by the AIC:

Theorem 2. *$E[\Delta] = (k/m)$ + higher order terms, where k is the number of parameters for Θ, and the higher order terms tend to zero as $m \to \infty$ faster than $1/m$.*

Equivalently, $\mathcal{E}_{\mathbf{Z}}(\theta_{\mathbf{Z}}) + (k/m)$ is an asymptotically unbiased estimate for $\mathcal{E}(\theta_{\mathbf{Z}})$. Thus, the AIC focuses on exactly what we need, but this is only an asymptotic result. It would be very interesting to get versions of the AIC that are applicable when we have only a finite amount of data.

Dependent data: It is important to extend the theory to learning a function using non-*iid* observations. Some work in this direction has been done by Karandikar & Vidyasagar (2002). One would need to assume some *mixing conditions* which specify how the dependence between observations decreases to zero as the time interval between the observations increases.

Learning dynamics: This paper has focused on learning a *static* input-output map. An important extension would be to learning the behaviour of a dynamical system, on the basis of measurements (possibly noisy) of the input and output signals. Note that the samples of the output signal are not *iid*.

Learning time-varying systems: So far we have focused on learning the behaviour of a time-invariant system. But there are many applications in which we have to learn the behaviour of a time-varying system in an "on-line" manner. In this scenario, even if we get a new observation at each time step, *effectively* we still have a limited number of observations due to the time-varying nature of the system. Thus it is necessary for us to have some sort of complexity control. The optimal level of complexity probably depends upon how fast the system can vary.

A very interesting question would be whether this is applicable to adaptive filtering. Current adaptive filtering algorithms (such as Kalman filters, LMS filters) do not incorporate any complexity control.

As the above cited examples indicate, Learning theory offers many interesting and challenging problems to work on, and has the potential to make a significant impact in many important application areas.

This work was supported in part by a DRDO-IISc Programme and by DRDO and ISRO through the Nonlinear Studies Group, IISc.

References

Burnham K P, Anderson D 2002 *Model selection and multi-model inference* (Springer-Verlag)

Cucker F, Smale S 2002a Best choices for regularization parameters in learning theory: On the bias-variance problem. *Found. Comput. Math.* 2: 413–428

Cucker F, Smale S 2002b On the mathematical foundations of learning. *Bull. Am. Math. Soc.* 39: 1–49

DeVito E, Rosasco L, Caponnetto A, DeGiovannini U, Odone F 2005 Learning from examples as an inverse problem. *J. Mach. Learn. Res.* 6: 883–904

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

Hastie T, Tibshirani R, Friedman J 2001 *The elements of statistical learning: data mining, inference and prediction* (New York: Springer-Verlag)

Karandikar R L, Vidyasagar M 2002 Rates of uniform convergence of empirical means with mixing processes. *Stat. Probab. Lett.* 58: 297–307

Schölkopf B, Smola A 2002 *Learning with kernels: Support vector machines, regularization, optimization and beyond* (Cambridge, MA: MIT Press)

Sin C-Y, White H 1996 Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* 71: 207–225

Smale S, Zhou D 2005 Learning theory estimates via integral operators and their approximations. Preprint available at http://www.tti-c.orgsmale_papers/sampIII5412.pdf

Vapnik V 1998 *Statistical learning theory* (New York: John Wiley & Sons)

Vidyasagar M 1997 *A theory of learning and generalization* (Berlin: Springer-Verlag)