Complexity of Regularization RBF Networks

Mark A. Kon
Department of Mathematics and Statistics
Boston University
Boston, MA 02215
mkon@bu.edu

Leszek Plaskota
Institute of Applied Mathematics
University of Warsaw
02-097 Warsaw, Poland
leszekp@hydra.mimuw.edu.pl

Abstract

We attempt to unify and compare a class of algorithms for learning input-output (i-o) functions f from examples. Our general approach involves parsing information about f into a priori and a posteriori information, with each represented by a probability measure on the space F of candidate functions. A consequence is that complexities of different approximation algorithms for the same problems will be possible to compare, and optimal algorithms will be possible to identify. We illustrate this by formulating an information complexity theory for regularization radial basis function (RBF) networks. We show the ϵ-complexity of approximating f using regularization is equivalent to the ϵ-complexity of approximating f using any consistent Bayesian approach. In particular, a Gaussian prior distribution may be assumed for correct computation of all complexities.

1. Introduction

There are currently many areas of mathematics, statistics, and computer science which deal with learning theory, which effectively involves the extrapolation of functions from partial information or examples.

The theory of learning in neural networks has the goal of extrapolating an input-output (i-o) function f from partial information consisting of examples of f. Given a set of data points, the many ways of extrapolating a function $f(x)$ from these imply a need for a study of how such methodologies fit into a larger framework. In this paper we will discuss approaches to this problem, and attempt to place several approximation procedures into a wider context. Through this we hope to develop an approach to comparing approximation errors of different methodologies, and hence the complexities of finding ϵ-approximations of i-o functions. A sub-goal is the formulation of a normative index of approximation complexities, and more direct comparisons of different approximation methods.

One solution of the learning problem is provided by radial basis function (RBF) neural networks [3,4], which extrapolate f from examples by effectively assuming a prescribed smoothness, or possibly other a priori information. An information complexity theory for RBF networks estimates how many examples are needed to approximate f to a given tolerance. Such a theory has been studied in [1, 2], in the context of a worst-case formulation of error, with an assumption that f is known a priori to be contained in a given set F_1.

A more general formulation of the learning problem can be formulated as follows. Assume that there is a function $f : X \to Y$ to be learned through examples

$$Nf = (f(x_1), \ldots, f(x_k))$$

If random errors are involved so that $f(x_i)$ is replaced by $f(x_i) + \epsilon_i$, the relationship between X and Y becomes probabilistic, and a probability density μ on $X \times Y$ is a better description of the relationship of input and output. We will assume here for simplicity that the relationship between X and Y is deterministic, so that $\epsilon_i = 0$.

Most approaches to estimating f have in common the fact that there is information provided *a priori* regarding f, to be used in learning f. In addition, it is assumed some additional data about f are given (generally in the form of examples Nf). These data are *a posteriori* information. The learning algorithm must make the best guess at f from the two types of information. More generally, we can formulate *a priori* information as some kind of prior knowledge about $f(x)$. This prior knowledge can be thought of as a probability distribution μ_{pr} on the space F of all possible f.

On the other hand, an *a posteriori* probability distribution μ_{po} on F is implied by the information Nf. The density $\mu_{\text{po}}(f)$ will be assumed to depend only on the finite collection of numbers $f(x_1), \ldots, f(x_n)$. A distribution of this type, which depends on a finite number

of variables, is known as a *cylinder* distribution (or cylinder measure).

The job of an inference engine is to combine the probability distributions μ_{pr} and μ_{po} in the best way in order to estimate f. To estimate f it is also necessary to assume a penalty for errors. The most general assumption is that this has a form $U(f, \overline{f})$ where f is the sought function, and $\overline{f}$ is the estimate of f. A reasonable algorithm for estimating f will minimize the expected risk, which must be carefully defined, given that two equally valid probability measures are at work.

2. The parsing of approximation methods: some examples

Learning theory includes many different and very effective approaches. An important goal is to classify these within a coherent framework, and to analyze and compare their complexities of approximation within this framework. We include some examples of successful learning algorithms and indicate their parsing of a priori and a posteriori information. Since the latter generally consists of examples Nf, it is not specified here unless necessary. In all of the examples below, it is possible to identify reasonable μ_{pr} and μ_{po}, though this is not done here for brevity.

1. *Interpolatory approach (worst case approach) in continuous complexity theory*: Here generally, a priori information consists of the fact that $f \in F_1$, where F_1 is a balanced convex set of functions in a normed linear space. Optimal algorithms approximate center of set $F_1 \cap N^{-1}y$ through an algorithm $\phi(Nf) \approx f$.

2. *Average case approach in continuous complexity*: The a priori information is that the function f to be learned is drawn from a given probability distribution $d\mu_{\text{pr}}(f)$ on a normed linear space F. The algorithm for choosing f selects f to be the average of $d\mu_{\text{pr}}(f)$ conditioned on $Nf = y$. If μ_{pr} is a Gaussian measure then an optimal learning algorithm is the *spline algorithm* [5], which gives an approximation to f of the form

$$\overline{f} = \phi(Nf) = \sum_j f(x_j) C_\mu K(x, x_j)$$

where $C_\mu = A^{-2}$ is the covariance operator of the Gaussian (a generalization of the covariance matrix). Here, $K(x, x_j)$ is a radial basis function, i.e., the reproducing kernel for the space with the norm $\|Af\|$.

3. *Feedforward Neural Network Algorithms*: These learn from examples $Nf = (f(x_1), \ldots, f(x_n))$ of an unknown i-o function $f(x)$, and form an approximation

$$\overline{f} = \sum_j c_j G_j(x), \tag{1}$$

where G_j might be ridge functions, radial basis functions or other functions computed by the hidden layer, with the coefficients computed using backpropagation or other algorithms. An implicit a priori assumption is that $\overline{f}$ can be written in the form (1). This is often effectively a smoothness assumption, since a finite number of neurons with a smooth activation function H will produce smooth approximations $\overline{f}$. Indeed, if $G_j(x)$ are RBF's, we know optimal approximations (1) explicitly minimize Sobolev norm.

4. *Maximum likelihood approaches*: These estimate f using an approximation $\overline{f}$ which is consistent with the information Nf, and whose probability is the largest under the a priori measure $d\mu_{\text{pr}}(f)$ restricted to $\{f : Nf = y\}$.

5. *Regularization (radial basis function) approaches*: Here the a priori information consists, for example, of the fact that f is smooth, i.e., that its Sobolev norm $\|Af\|$ is small (here $A = -\Delta + 1$, with $-\Delta$ the Laplacian).

The algorithm for choosing f involves minimization of a weighted combination

$$H_\lambda(\overline{f}) = \|N\overline{f} - y\|^2 + \lambda \|A\overline{f}\|,$$

where $Nf = y$ is the a posteriori information (data) from the f to be learned. The minimizing $\overline{f}$, under some standard assumptions, is evaluated by a neural network which computes a linear combination of radial basis functions (RBF's) of the form

$$\overline{f} = \sum_j c_j K(x, x_j).$$

This method can also be reformulated probabilistically in the context of the prior and post probability densities μ_{pr} and μ_{po}.

6. Vapnik-Chervenenkis (V-C) approach: Given nested family $\{V_\lambda\}$ of candidate *a priori* spaces increasing in size (and complexity) with λ, one can take the following approach. If λ is small then the candidate set $N^{-1}y \cap V_\lambda$ small, so that the selection of approximation

$$\tilde{f} \in N^{-1}y \cap V_\lambda$$

is from reasonable sized set. The method is to choose a sufficiently small λ that the set V_λ is too small to overfit the data.

7. Adaptive resonance theory (ART) algorithms: There is a dynamic neural network weight allocation in this procedure for classifying input vectors **x**. This is a feedforward

network with a second competitive processing stage where the hidden neuron $y_i = G_i(x)$ with highest activation suppresses all other neurons $y_j \neq y_i$. Thus, an ART network computes the function

$$f^*(x) = (0, \ldots, 0, G_j(x), 0, \ldots, 0), \tag{2}$$

where the right hand side represents the choice of the $G_j(x) = y_j$ with the maximum value indicating membership of the input in a class C_j. A priori information consists of the fact that the i-o function f^* is approximable in the form (2). Choice of a smooth $G_i(x)$ is effectively an a priori assumption on smoothness of the separators of classes C_i above. Note we are ignoring here the dynamics of programming an ART network

Other methodologies not mentioned above include:

- Computational learning theory
- Regression methods in statistics
- Maximum entropy methodologies
- Decision tree methodologies in artificial intelligence

After placing different algorithms into a single context, a second step should be to identify and compare ϵ-complexities of different algorithms. Given $\epsilon > 0$, we wish to compute the information complexity (the number k of pieces of information in Nf) required to approximate f within ϵ. A further goal should be to form a normative index of such methods according to their (now comparable) optimality properties. This step might involve precisely defining optimality within such classes of approaches, and identifying optimal algorithms from among differing approaches. This by no means obviates the need for varied methodologies, but rather allows them to be compared to each other on the same problems.

In the next section we illustrate a methodology for comparing complexities. We formulate a relationship between RBF regularization network techniques for learning, and Bayesian learning methods used in the average case setting of continuous complexity theory. We believe these results are interesting because they formulate a consistent complexity theory for RBF regularization networks. In particular it follows that the complexity depends only on the RBF network itself, through its regularization operator A.

Our results are obtained through consideration of all possible Bayesian prior distributions consistent with the regularization operator A of the RBF network. To define average complexity (number of examples) required to approximate f within error $\epsilon > 0$, we will first define the average case error. For this an a priori probability distribution $d\mu_{\text{pr}}$ on the set of possible f is required. If the approximation complexity is dependent on the assumed μ_{pr}, there will be several different complexities consistent with a single given RBF regularization network, and no complexity theory will be possible.

Using probabilistic methods, we will show that all μ_{pr} consistent with the regularization operator A yield a single complexity order, yielding a unique complexity theory for the RBF network with regularization operator A.

We also will show that all Bayesian models which are consistent with the regularization RBF network yield the same complexity of approximation using RBF algorithms. This complexity is the same one obtained by assuming a Gaussian prior distribution whose covariance operator is A^{-2}.

3. A formulation of regularization networks in Bayesian terms

Let H be a Hilbert space, and let $f \in H$ be an unknown function we seek to identify. Assume we have information $Nf = (f(x_1), \ldots, f(x_k)) = y$, (sometimes called standard information).

We assume we have the priori information that f has small norm with respect to some operator, e.g.,

$$\|Af\| = \text{small}$$

(for example, $A = -\Delta + 1$ leads to the assumption of smoothness for f; see above).

We then minimize a regularization functional to approximate f:

$$f = \arg\min\{\|Nf - y\|^2 + \lambda \|Af\|^2\}. \tag{3}$$

We assume the information

$$Nf = y \tag{4}$$

is exact, and that y is given. Then (4) becomes a constraint, and through regularization theory it can be shown that we should take the $\lambda \to 0$ limit in the optimization problem (3), which yields

$$f = \arg\min\{\|Af\|^2 : \ Nf = y\}. \tag{5}$$

This use of a priori information implies a Bayesian viewpoint: there exists an a priori probability measure ν on H whose density at $f \in H$ is a "function of" $\|Af\|$ and decreases monotonically with $\|Af\|$.

A first guess at an a priori density would be

$$d\nu(f) = h(\|Af\|)\, df, \tag{6}$$

with df Lebesgue measure on H (at least if H is finite dimensional). A measure $d\nu$ consistent with this definition does not always exist in infinite dimension. However, in all dimensions, if A is invertible and A^{-1} is trace class, there is a Gaussian measure on H consistent with (6). If H is finite dimensional with dimension d, this measure has the form

$$d\mu_A(f) = \frac{1}{(2\pi)^{d/2}\det(A)} e^{-\frac{1}{2}\|Af\|^2} df.$$

The covariance operator of this Gaussian measure is A^{-2}. In infinite dimensions such measures are discussed in, e.g., [5].

In the general (non-Gaussian) case, we want a measure $\nu(f)$ which is *consistent* with the regularization operator A in the sense that it "depends" only on $\|Af\|$.

A general a priori measure ν satisfying this condition can be constructed as follows. Define the sets

$$H_c = \{f \in H : \|Af\| = c\}.$$

Measure theoretically, we have

$$H = \bigcup_{c\geq 0} H_c = H_1 \times \mathbb{R}^+, \tag{7}$$

with $\mathbb{R}^+$ the nonnegative reals and the subscript on the right denoting $c = 1$. We want measures ν which are "constant" on each H_c. For any Borel measure ν supported on the domain of A, define

$$\nu_c = \nu(\cdot\,|H_c)$$

to be the conditional measure of ν on H_c.

Note that the Gaussian measure μ_A with covariance A^{-2} has the property (as in the finite dimensional case) that its conditional density μ_c is the unique uniform measure on each H_c.

By (7) (analogously to the finite dimensional case), any measure $d\nu(f)$ whose density depends only on $\|Af\|$ has uniform conditional measures $\nu_c = \mu_{A_c}$ on H_c, and some marginal measure ν_m on $\mathbb{R}^+$. The measure ν_m (together with the uniform measures ν_c) uniquely define the measure ν on H, which is the general form of an a priori measure on H is consistent with the regularization assumptions. In addition, it is consistent with the regularity assumptions stipulate that ν_m be decreasing (since we want probability to decrease with increasing values of $\|Af\|$).

Henceforth we assume the measure ν_m has a finite mean. We can now exactly compute the information complexity $\text{comp}_\nu(\epsilon)$ of approximating f to error $\epsilon > 0$. This is defined as the minimum number k of examples $(x_1, \ldots, x_k)$ which yield an average error of less than ϵ (see [5, 6]).

Specifically, for given

$$y = N_k f = (f(x_1), \ldots, f(x_k)) \tag{8}$$

of cardinality k, we define the best estimate $\overline{f}(N_k, y)$ of f by

$$\overline{f}(N_k, y) = \arg\inf_{\overline{f}} (E_{f\in N_k^{-1}y}\|f - \overline{f}\|^2),$$

i.e., the function $\overline{f}$ which minimizes the squared error. Above, $E_{f\in N_k^{-1}y}$ denotes the average over all $f \in H$ consistent with the information y, with respect to the (conditional) measure ν.

We emphasize that the results here assume that the information Nf has the standard form (8) (i.e., consists of pointwise evaluations of f only).

We then define the average case error to be

$$e(N_k, y) = E_f(\|f - \overline{f}(N_k, y)\|^2)^{1/2},$$

where the expectation is taken over all f with respect to ν. Finally, for given k we choose the information operator N_k of cardinality k which gives the smallest error for the worst y, and measure the error:

$$e(k) = \inf_N \sup_y e(N_k, y).$$

This represents the minimum error $e(k)$ of estimating the worst function f, using the best information operator N_k consisting of k examples $f(x_1), \ldots, f(x_k)$.

The ϵ-complexity $\text{comp}(\epsilon)$ is defined as the minimum number of examples required to obtain an error less than ϵ:

$$\text{comp}_\nu(\epsilon) = \inf\{k \in \mathbb{N}|\, e(k) \leq \epsilon\}.$$

Now we relate $\text{comp}_\nu(\epsilon)$ with the theory of regularization networks. Assume we are given a regularization operator A and a regularized estimate $\overline{f}_{\text{reg},k}$ of f obtained from (5) above, using the best information operator $N = N_k$. For a given a priori measure $\mu_{\text{pr}} = \nu$, we define the RBF regularization network ϵ-complexity by

$$\text{comp}_{\text{reg}}(\epsilon) = \inf\{k \in \mathbb{N}|E_f(\|\overline{f}_{\text{reg},k} - f\|^2)^{1/2} \leq \epsilon\}.$$

We wish to compare this ϵ-complexity with the ϵ-complexity obtained from the average case (Bayesian) approach with continuous complexity theory, using any priori probability density ν consistent with the regularization assumptions arising from the operator A above. Thus the a priori probability distribution ν will be assumed to have uniform density on the sets

$$H_c = \{f|\,\|Af\| = c\},$$

with a decreasing marginal density ν_m (see above) which has a finite mean.

With this as the general form of an a priori measure ν, we have the following theorem, whose proof is omitted for brevity:

Theorem: The ϵ-complexity of the regularization approximation problem with regularization operator A is equal to the average case (Bayesian) ϵ-complexity for any measure ν which is consistent with A. This in turn is equivalent to the average case ϵ-complexity for the

Gaussian measure μ_A with covariance operator A^{-2}. Specifically, we have

$$\text{comp}_{\text{reg}}(\epsilon) = O(\text{comp}_{\nu}(\epsilon)) = O(\text{comp}_{\mu_A}(\epsilon)) \quad (9)$$

We remark that the last complexity in (9) can be computed using known techniques in continuous complexity [5].

4. Conclusions

We conclude that information ϵ-complexities can be computed for regularization networks on the basis of any a priori assumption μ_{pr} consistent with the regularization operator A, and that the complexities which obtain are independent of this choice of μ_{pr}.

This allows the definition of a "regularization complexity" $\text{comp}_{\text{reg}}(\epsilon)$ which depends only on the regularization operator A.

Equivalently, what is necessary to define global information complexities is the regularization assumption that "$\|Af\|$ should be small"; all complexities consistent with this assumption can be computed from the average case setting using a Gaussian prior.

Finally, we conclude with

Proposition: Under the same measure ν, maximum likelihood estimation also gives the same ϵ-complexities as in the theorem.

Acknowledgments:

The first author's research was partially supported by the U.S. National Science Foundation.

The second author's research was partially supported by the Polish-American Fulbright Foundation and the National Committee for Scientific Research of Poland.

References:

[1] M. Kon and L. Plaskota, "Information complexity of neural networks", *Neural Networks* **13,** 2000, pp. 365-376.

[2] M. Kon and L. Plaskota, "Complexity of neural network approximation with limited information: a worst-case approach", to appear, *J. Complexity.*

[3] C. A. Micchelli, "Interpolation of scattered data: Distance matrices and conditionally positive definite functions", *Constructive Approximation* **2**, 1986, pp.11-22.

[4] C.A. Micchelli and M. Buhmann, "On radial basis approximation on periodic grids", *Math. Proc. Camb. Phil. Soc.* **112**, 1992, pp. 317-334.

[5] Traub J., G. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*, Academic Press, Boston, 1988.

[6] Traub, J. and H. Wozniakowski, *A General Theory of Optimal Algorithms*, Academic Press, New York, 1980.