DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 921

May 17, 1994

Generalization and Regularization in Nonlinear Learning Systems [1]

by

Grace Wahba

[1] To appear in the Handoobk of Brain Theory and Neural Networks, Michael Arbib, Ed. Prepared within the space and reference limitations of the Handbook

1 Introduction

In this article we will describe generalization and regularization from the point of view of multivariate function estimation in a statistical context. Multivariate function estimation is not, in principle, distinguishable from supervised machine learning. However, until fairly recently supervised machine learning and multivariate function estimation had fairly distinct groups of practitioners, and small overlap in language, literature, and in the kinds of practical problems under study.

In any case, we are given a *training set*, consisting of pairs of input (feature) vectors and associated outputs $\{\boldsymbol{t}(i), y_i\}$, for n training or example subjects, $i = 1, ...n$. From this data, it is desired to construct a map which *generalizes well*, that is, given a new value of $\boldsymbol{t}$, the map will provide a reasonable prediction for the unobserved output associated with this $\boldsymbol{t}$.

Most applications fall into one of two broad categories, which might be called nonparametric regression and classification. In nonparametric regression, y may be (any) real number or a vector of r real numbers. In classification y is usually represented as a q-dimensional vector of zeroes and ones, with a single 'one' in the kth position if the example (subject) came from category k. In some classification applications, the desired algorithm will, given $\boldsymbol{t}$, return a vector of zeroes and ones indicating a category assignment, ('hard' classification), In other applications, it may be desired to return a q-vector of probabilities (that is non-negative numbers summing to 1), which represent a forecast of the *probabilities* of an object with predictor vector $\boldsymbol{t}$ being in each of the q categories ('soft' classification).

In some problems the feature vector $\boldsymbol{t}$ of dimension d contains zeros and ones (for example as in a bitmap of handwriting), in others it may contain real numbers representing some physical quantities. In this article we will be generally concerned with the latter, since the ideas of generalization and regularization are easiest to discuss when there is a convenient topology (for example that determined by distance in Euclidean d-space) so that 'closeness' and 'smoothness' can be easily defined. *Regularization*, loosely speaking, means that some constraints are applied to the construction of the map with the goal of reducing the generalization error (see also REGULARIZATION THEORY AND LOW-LEVEL VISION). Ideally, these constraints embody *a priori* information concerning the true relationship between input and output; alternatively, various *ad hoc* constraints have sometimes been shown to work well in practice.

2 Generalization and Regularization in Non-Parametric Regression

2.1 Single Input Spline Smoothing

We will use Figure 2.1 to illustrate the ideas of generalization and regularization in the simplest possible nonparametric regression setup, that is, $d = 1$, $r = 1$, with $\boldsymbol{t} = t$ any real number in some interval of the real line. The boxed points (which are identical in each of the three panels of Figure 2.1) represent $n = 100$ (synthetically generated) input-output pairs $\{t(i), y_i\}$, generated according to the model

$$y_i = f_{TRUE}(t(i)) + \epsilon_i, \quad i = 1, ..., n, \tag{2.1}$$

where $f_{TRUE}(t) = 4.26(e^{-t} - 4e^{-2t} + 3e^{-3t})$, and the ϵ_i came from a pseudorandom number generator for normally distributed random variables with mean 0 and standard deviation 0.2. These figures are from Wahba and Wold (1975). Given this training data $\{t(i), y_i, i = 1, ..., n\}$, the learning problem is to create a map which, if given a new value of t, will predict the response $y(t)$. In this case, the data are noisy, so that even if the new t coincides with some predictor variable $t(i)$ in

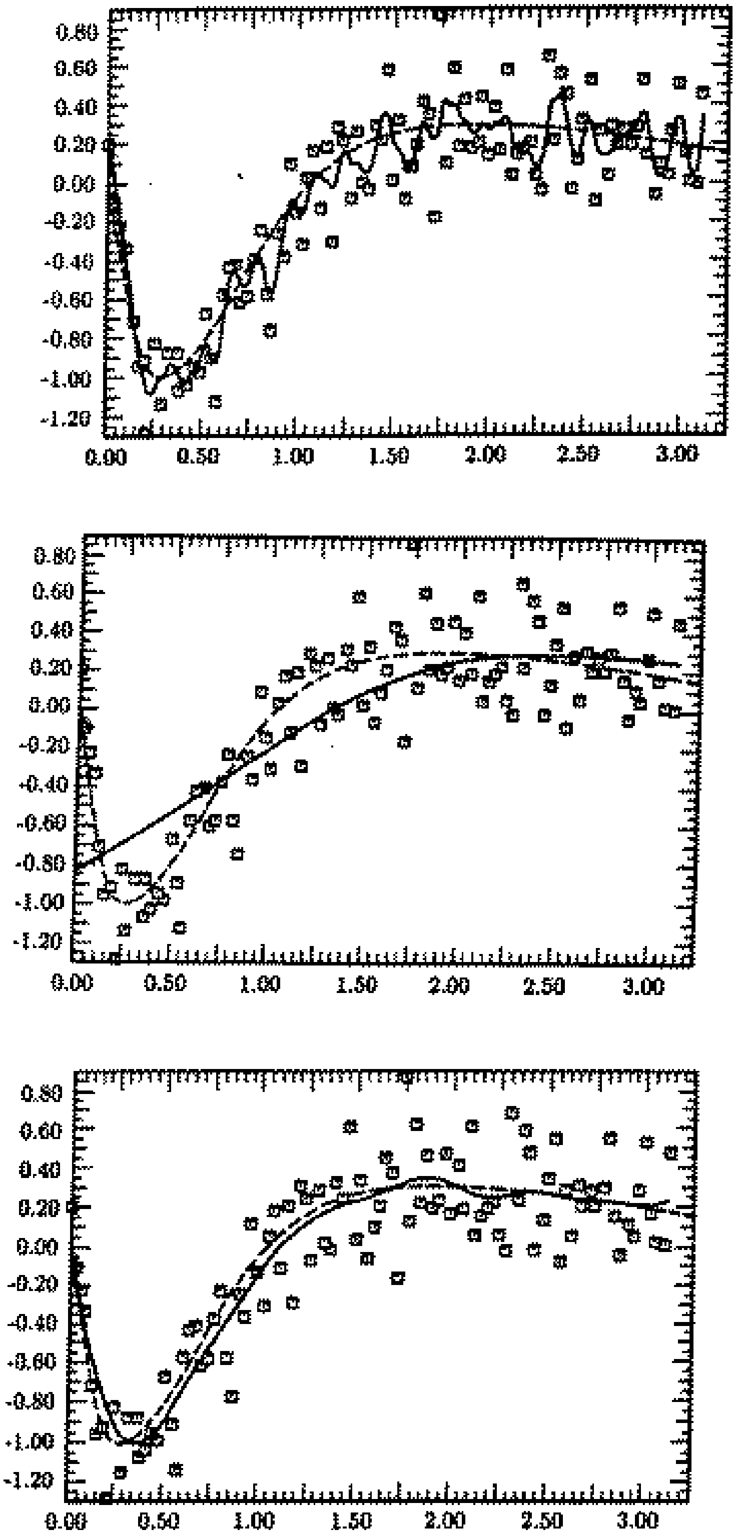

Figure 2.1: Training data (boxed points) have been generated by adding noise to $f_{TRUE}(t)$, shown by the dashed curve in each panel. All three panels have the same data. Top: Solid curve is fitted spline with λ too small. Middle: Solid curve is fitted spline with λ too large. Bottom: Solid curve is fitted spline with λ obtained by leaving-out-one cross validation.

the training set, merely predicting y as the response y_i is not likely to be satisfactory. Also, this does not yet provide any ability to make predictions when t does not exactly match any predictor values in the training set. It is desired to generate some sort of curve, which will allow a reasonable prediction of the response for any t within a reasonable vicinity of the set of training predictors $\{t(i)\}$. The dashed line in each panel of Figure 2.1 is $f_{TRUE}(t)$; the three solid black lines in the three panels of Figure 2.1 are three solutions to the variational problem: Find f in the [Hilbert] space W_2 of functions with continuous first derivatives and square integrable second derivatives which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t(i))^2 + \lambda \int (f^{(2)}(u))^2 du, \tag{2.2}$$

for three different values of λ. The parameter λ is known as the regularization or smoothing parameter. As $\lambda \rightarrow \infty$, f_λ tends to the least squares straight line best fitting the data, and as $\lambda \rightarrow 0$ the solution tends to that curve in W_2 which minimizes the penalty functional $J(f) = \int (f^{(2)}(u))^2 du$ subject to interpolating the data (provided the $\{t(i)\}$ are distinct). This latter interpolating curve is known as a cubic interpolating spline, and minimizers of (2.2) are known as smoothing splines. See Wahba (1990) and references cited there for further information concerning these and other properties of splines noted below, and further references. These splines have been studied at least since they were discussed by I. Schoenberg in the 1940's. Schoenberg gave the interpolating spline its name after the mechanical spline (a thin, flexible rod with weights attached at selected points) which was used by draftsmen for drawing smooth curves which were used to represent cross sections of ships hulls. In the top panel of Figure 2.1, λ has been chosen too small, and the wiggly solid line is attempting to fit the data too closely. It can be seen that using the wiggly curve in the top panel is not likely to give a good prediction of y, assuming that future predictor-response data is generated by the same mechanism as the training data. In the middle panel, λ has been chosen too large, the curve has been forced to flatten out, and again it can be seen that the heavy line will not give a good prediction of y. In the bottom panel, λ has been chosen by a leaving-out-one cross validation (see Wahba and Wold (1975)), and it can be seen that the λ obtained this way does a good job of choosing the right amount of smoothing to best recover f_{TRUE} of Equation (2.1). The f_{TRUE} of Equation (2.1) would provide the best predictor of the response in an expected mean square error sense if future data were generated according to Equation (2.1). The curve in the bottom panel has a reasonable ability to *generalize*, that is, to predict the response given a new value t of the predictor variable, at least if t is not too far from the training predictor set $\{t(i)\}$.

For each positive λ, there exists a unique $\kappa = \kappa(\lambda)$ so that the minimizer f_λ of (2.2) is also the solution to the problem: Find f in W_2 to minimize

$$RSS(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(t(i))^2 \tag{2.3}$$

subject to the condition

$$J(f) = \int (f^{(2)}(u))^2 du \leq \kappa. \tag{2.4}$$

As λ becomes large, the associated $\kappa(\lambda)$ becomes small, and conversely. In general, the term *regularization* refers to solving some problem involving best fitting, subject to some constraint(s) on the solution. These constraints may be of various forms. When they involve a quadratic penalty involving derivatives, like $J(f)$, the method is commonly referred to as Tikhonov regularization. The 'tighter' the constraints (i. e. the smaller κ, equivalently the larger λ) the further away the solution f_λ will generally be from the training data, that is, RSS will be larger. As the constraints get weaker and weaker then ultimately (if there are enough degrees of freedom in the method) the solution will interpolate the data. However, as is clear from Figure 2.1, a curve which runs through all the data points is *not* a good solution.

A fundamental problem in machine learning with noisy and or incomplete data, is to balance the 'tightness' of the constraints with the 'goodness of fit' to the data, in such a way as to minimize the 'generalization error', that is, the ability to predict the unobserved response for new values of t (or $\boldsymbol{t}$). This tradeoff is called the bias-variance tradeoff in the statistical literature, and has recently been discussed in the context of machine learning by Geman, Bienenstock and Doursat (1992). Numerous methods for curve fitting, other than smoothing splines, in the case $d = 1, r = 1$ have been proposed in the statistics literature. Popular methods include Parzen kernel estimates, nearest neighbor estimates, orthogonal series estimates, and least squares spline estimates. See Eubank (1988) and references cited there. Each method has one or more regularization parameters, either explicit or implicit, that control the bias-variance tradeoff.

2.2 Single Input, Single Hidden Layer Feed-Forward Neural Net

A single input, single hidden layer feed-forward neural net (NN) predictor for the learning problem of Section 2.2 output, is typically of the form

$$f_{NN}(t) = \sigma_0(b_o + \sum_{j=1}^{N} w_j \sigma_h(a_j t + b_j)) \tag{2.5}$$

where σ_h is the so-called 'activation function' of the hidden layer and σ_0 is the activation function for the output. σ_h is generally a sigmoidal function, for example, $\sigma_h(\tau) = e^{\tau}/(1 + e^{\tau})$, while σ_0 may be linear, sigmoidal or a threshold unit. In the learning problem of Section 2.1, best results would likely be obtained with σ_0 linear. Here N is the number of hidden units, and the w_j, a_j and b_j are 'learned' from the training data by some appropriate iterative descent algorithm that tries to steer these values towards minimizing some distance measure, typically $RSS = \sum_{i=1}^{n}(f_{NN}(t(i)) - y_i)^2$. It is clear that if N is sufficiently large, and the descent algorithm is run long enough, it should be possible to drive the RSS as close as one likes to 0. (In practice it is possible to get stuck in local minima.) However, it is also clear intuitively from Figure 2.1 that driving RSS all the way to zero is not a desirable thing to do. Regularization in this problem may be done by controlling the size of N, by imposing penalties on the w_j, by stopping the descent algorithm early, that is, not driving down RSS as far as it can go, or by various combinations of these strategies. Each will influence how closely f_{NN} will fit the data, how 'wiggly' it will be, and how well it will be able to predict unobserved data that is generated by a similar mechanism as the observed data. See Weigend (1993).

2.3 Multiple Input Single Hidden Layer Feedforward Neural Net

For d greater than 1, the single hidden layer feed-forward neural net with a d dimensional input and one dimensional output is of the form

$$f_{NN}(\boldsymbol{t}) = \sigma_0(b_o + \sum_{j=1}^{N} w_j \sigma_h(\boldsymbol{a}_j' \boldsymbol{t}(i) + b_j)) \tag{2.6}$$

where σ_0 and σ_h are as before, but now the $\boldsymbol{a}_j$ and $\boldsymbol{t}$ are d-vectors. All the remarks concerning the regularizing of this network by controlling N, penalizing the w_j and stopping short of driving RSS to a minimum noted in Section 2.2 apply here.

2.4 Multiple Input Radial Basis Function and Related Estimates

RADIAL BASIS FUNCTIONS are a popular method for nonparametric regression. We first describe a general form of nonparametric regression which will specialize to radial basis functions and other methods of interest. Let $R(\boldsymbol{s},\boldsymbol{t})$ be *any* symmetric, strictly positive definite function on $E^d \times E^d$. Here strictly positive definite means for any $K = 1, 2, ...$ the $K \times K$ matrix with j, kth entry $R(\boldsymbol{s}(j), \boldsymbol{s}(k))$ is strictly positive definite whenever the $\boldsymbol{s}(1), ..., \boldsymbol{s}(K)$ are distinct. (A symmetric $K \times K$ matrix M is said to be positive definite if for any K dimensional column vector x, $x'Mx$ is greater than or equal to 0, and is said to be strictly positive definite if $x'Mx$ is always strictly greater than 0.) Positive definiteness will play a key role in the discussion below because, (among other reasons) any positive definite matrix can be the covariance matrix of a random vector and any positive definite function $R(\boldsymbol{s},\boldsymbol{t})$ can be the covariance function of some stochastic process, $X(\boldsymbol{t})$. That is, there exists $X(\cdot)$ such that $Cov\ X(\boldsymbol{s})X(\boldsymbol{t}) = R(\boldsymbol{s},\boldsymbol{t})$. Given training data $\{\boldsymbol{t}(i), y_i\}$, it is always possible in principle to obtain a (regularized) input-output map from this data by letting the model $f_{R,\lambda}$ be of the form

$$f_{R,\lambda}(\boldsymbol{t}) = \sum_{j=1}^{N} c_j R(\boldsymbol{t}, \boldsymbol{s}(j)), \tag{2.7}$$

where the $\boldsymbol{s}(j)$ are $N \le n$ 'centers' which are placed at distinct values of the $\{\boldsymbol{t}(i)\}$ and $c = (c_1, ..., c_N)'$ is chosen to minimize $RSS(f) + \lambda J(f)$. Here

$$RSS(f_{R,\lambda}) = \sum_{i=1}^{n} (f_{R,\lambda}(\boldsymbol{t}(i)) - y_i)^2 \tag{2.8}$$

and the regularizing penalty $J(\cdot)$ is of the form

$$J(f_{R,\lambda}) = \sum_{j,k=1}^{N} c_j c_k J_{jk} \tag{2.9}$$

where J_{jk} are the entries of a non-negative definite quadratic form. The (strict) positive definiteness of R guarantees that

$$RSS(f_{R,\lambda}) + \lambda J(f_{R,\lambda}) \tag{2.10}$$

always has a unique minimizer in c, for any non-negative λ. This follows by substituting (2.7) into (2.10), and using the fact that the columns of the $n \times N$ matrix with i, j entry $R(\boldsymbol{t}(i), \boldsymbol{s}(j))$ are linearly independent since they are just N columns of the $n \times n$ positive definite matrix with i, j entry $R(\boldsymbol{t}(i), \boldsymbol{t}(j))$.

Radial basis function estimates are obtained for the special case where $R(\boldsymbol{s},\boldsymbol{t})$ is of the special form

$$R(\boldsymbol{s},\boldsymbol{t}) = r(\|W(\boldsymbol{s} - \boldsymbol{t})\|), \tag{2.11}$$

where W is some linear transformation on E^d and the norm is Euclidean distance. That is, $R(\boldsymbol{s},\boldsymbol{t})$ depends only on some generalized distance in E^d between $\boldsymbol{s}$ and $\boldsymbol{t}$. The regularization, that is, the effecting of the tradeoff between goodness of fit to the data and 'smoothness' of the solution, is performed by reducing N, and/or increasing λ. The choice of W will also affect the 'wiggliness' of $f_{R,\lambda}$ in the radial basis function case. Alternatively, a model can be obtained by choosing N small and minimizing $RSS(f)$. In that case N and W are the smoothing parameters.

In the special case $N = n$, $\boldsymbol{s}(i) = \boldsymbol{t}(i)$, the $f_{R,\lambda}$ can (for *any* positive definite R) be shown to be Bayes estimates, see Kimeldorf and Wahba (1971), Wahba (1990, 1992), Girosi, Jones and Poggio

(1993). Arguments can be given to show that if n is large and $N < n$ is not too small, then they are good approximations to Bayes estimates, see Wahba (1990, Chapter 7). In the special case $J_{i,j} = R(\boldsymbol{t}(i), \boldsymbol{t}(j))$, the Bayes model is easy to describe and we do it here; it is:

$$y_i = X(\boldsymbol{t}(i)) + \epsilon_i, \tag{2.12}$$

with $X(\boldsymbol{t})$ a zero mean Gaussian stochastic process with covariance $EX(\boldsymbol{s})X(\boldsymbol{t}) = bR(\boldsymbol{s}, \boldsymbol{t})$ and the ϵ_i independent zero mean Gaussian random variables with common variance σ^2, and independent of $X(\boldsymbol{t})$. In this case, the minimizer $f_{R,\lambda}$ of $RSS(f) + \lambda J(f)$, evaluated at $\boldsymbol{t}$, is the conditional expectation of $X(\boldsymbol{t})$, given $y_1, ..., y_n$ provided that λ is chosen as σ^2/nb. In general, pretending that one has a prior and computing the posterior mean or mode will have a regularizing effect. However, the degree of regularization (choice of λ here) may not be the same as what one would get by attempting to minimize the bias-variance tradeoff if the prior is not correct. See Section 4 below.

Thin plate splines in d variables (of order m) consist of radial basis functions plus polynomials of total degree less than m in d variables. ($2m - d > 0$ is required for technical reasons.) Letting $\boldsymbol{t} = (t_1, ..., d_d)$, the thin plate splines are minimizers (in an appropriate function space) of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{t}(i))^2 + \lambda \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial t_1^{\alpha_1}\cdots\partial t_d^{\alpha_d}}\right)^2 dt_1\cdots dt_d. \tag{2.13}$$

Setting $d = 1, m = 2$ gives the cubic spline case discussed earlier. Note that there is no penalty on polynomials of total degreee less than m, the thin plate splines with a particular choice of λ are Bayes estimates with an improper prior (that is, infinite variance) on the polynomials of total degree less than m, see Wahba (1990) and references cited there.

There are a number of variations on regularized estimates. Additive smoothing splines are of the form

$$f(\boldsymbol{t}) = \mu + \sum_{\alpha=1}^{d} f_\alpha(t_\alpha) \tag{2.14}$$

where μ and the f_α are the solution to a variational problem of the form: Find μ and $f_1, .., f_d$ in a certain function space to minimize

$$\sum_{i=1}^{n}(f(\boldsymbol{t}(i) - y_i)^2 + \sum_{\alpha=1}^{d}\lambda_\alpha J_\alpha(f_\alpha). \tag{2.15}$$

The J_α may be of the form of J in Equation (2.4). Here, there is a *regularization parameter* for each component. See Hastie and Tibshirani (1990), Wahba(1990). These additive models generalize to smoothing spline analysis of variance models (SS-ANOVA), whereby terms of the form $f_{\alpha\beta}, f_{\alpha\beta\gamma}$, etc., are added to the representation in Equation (2.14), and corresponding penalty terms with regularization parameters are added in Equation (2.15). The $f_\alpha, f_{\alpha\beta}$, etc, may be generalized to themselves being radial basis functions. See Gu and Wahba (1993).

Regression spline ANOVA models be obtained by setting the $f_\alpha, f_{\alpha\beta}$ etc. as linear combinations of a (relatively small) number of basis functions (usually splines). In this case the number of the basis functions is probably the most influential regularization parameter. See Hastie and Tibshirani (1991), Friedman (1991), and references cited there. These and similar methods again all have either explicit or implicit regularization parameters which govern the balance between the complexity of the model with the fit to the data - the bias-variance tradeoff. Most of them use some form of cross validation to make this tradeoff. Other references may be found in Wahba (1992).

3 Generalization and Regularization in Classification

Figure 3.1 is a scatterplot from Wahba *et al* (1994), of body mass index (`bmi`) and age (`age`) from 669 subjects from the Wisconsin Epidemiologic Study of Diabetic Retinopathy. Subjects who had diabetic retinopathy that progressed are indicated with a '+', (to be referred to as '1's), others are indicated by a $\cdot$, to be referred to as $0's$, thus y_i is 1 or 0. If the '+' and $\cdot$ were reasonably separable by some simple partition of the square in the top panel of Figure 3.1 the classification problem would be one of finding a reasonably simple description of this partition. Then a 'hard' classification would be made by assigning 1 or 0 to points according to which part of the partition they are in. In this kind of data, it is frequently of interest to make a 'soft' classification, that is, to estimate the *probability* $p(\boldsymbol{t})$ that a subject with predictor vector $\boldsymbol{t} =$(`bmi,age`) will become a 1. (This argument carries over to several classes, but for ease of discussion, we will only consider two classes.) In this case, some regularized (that is, 'smooth') estimate for $p(\boldsymbol{t})$ is desirable, since it would be highly unreasonable for the estimate to pass through the data, that is to take on the value 1 at the *'s and 0 at the $\cdot$'s. Regularized estimates can be obtained as follows. First, define

$$f(\boldsymbol{t}) = log[p(\boldsymbol{t})/(1-p(\boldsymbol{t}))]. \tag{3.1}$$

f is known in the statistics literature as the logit. Then $p(\boldsymbol{t})$ is a sigmoidal function of $f(\boldsymbol{t})$, that is $p(\boldsymbol{t}) = e^{f(\boldsymbol{t})}/(1+e^{f(\boldsymbol{t})})$. (Other sigmoidal functions may be used.) We will get a regularized estimate for f. $RSS(f)$ of Equation (2.3) will be replaced by an expression more suitable for $0-1$ data, by using the likelihood for this data. To describe the likelihood, note that if y is a random variable with $Prob\ [y=1] = p$ and $Prob\ [y=0] = (1-p)$, then the probability density (or likelihood) $P(y,p)$ for y when p is true, is just $P(y,p) = p^y(1-p)^{(1-y)}$, this merely says $P(1,p) = p$ and $P(0,p) = (1-p)$. Thus, the likelihood for $y_1, ...y_n$ (assuming that the y_i are independent), is

$$P(y_1,...,y_n; p(\boldsymbol{t}(1),...,p(\boldsymbol{t}(n)) = \Pi_{i=1}^n p(\boldsymbol{t}(i))^{y_i}(1-p(\boldsymbol{t}(i))^{(1-y_i)}. \tag{3.2}$$

Substituting f for p in (3.2), taking the negative logarithm, and suppressing the $\{y_i\}$ in the notation gives the negative log likelihood $L(f)$ in terms of f:

$$-\ logP(y_1,...,y_n; p(\boldsymbol{t}(1),...,p(\boldsymbol{t}(n)) \equiv L(f) = \sum_{i=1}^n [\log(1+e^{f(\boldsymbol{t}(i))}) - y_i f(\boldsymbol{t}(i))]. \tag{3.3}$$

It is natural for $L(f)$ to replace $RSS(f)$ of (2.3) when y_i is restricted to 0 or 1, since $RSS(f)$ is just (a linear function of) the negative log likelihood for y_i generated according to (2.1). A neural net implementation of soft classification would consist of finding $f_{NN}(\boldsymbol{t}) = logitp_{NN}(\boldsymbol{t})$ of the form of Equation (2.6) to minimize $L(f)$ of (3.3). If N is large enough, then, in principle, f_{NN} may be driven so that $p_{NN}(\boldsymbol{t}(i))$ is close to 1 if y_i is 1, and is close to 0 if y_i is 0. Again, it is intuitively clear that this is not desirable. As before, a regularized, or smooth f_{NN} can be obtained by controlling N, penalizing the w_i, stopping the iterative fitting early, or some combination of these. The bottom panel in Figure 3.1 gives the estimated probability of progression of diabetic retinopathy (estimated probability of a 1), as a function of `bmi` and `age`. This figure (from Wahba *et al* (1994)) is actually a cross section of a model of the form

$$f(\texttt{age},\texttt{gly},\texttt{bmi}) = \mu + f_1(\texttt{age}) + b\cdot\texttt{gly} + f_3(\texttt{bmi}) + f_{13}(\texttt{age},\texttt{bmi}) \tag{3.4}$$

where `gly` (glycosylated hemoglobin) was held fixed at the median value of the training set data for plotting purposes. μ, b, f_1, f_3 and f_{13} were obtained by finding f of the form of (3.4) to minimize

$$L(f) + \lambda_1 J_1(f_1) + \lambda_3 J_3(f_3) + \lambda_{13} J_{13}(f_{13}) \tag{3.5}$$

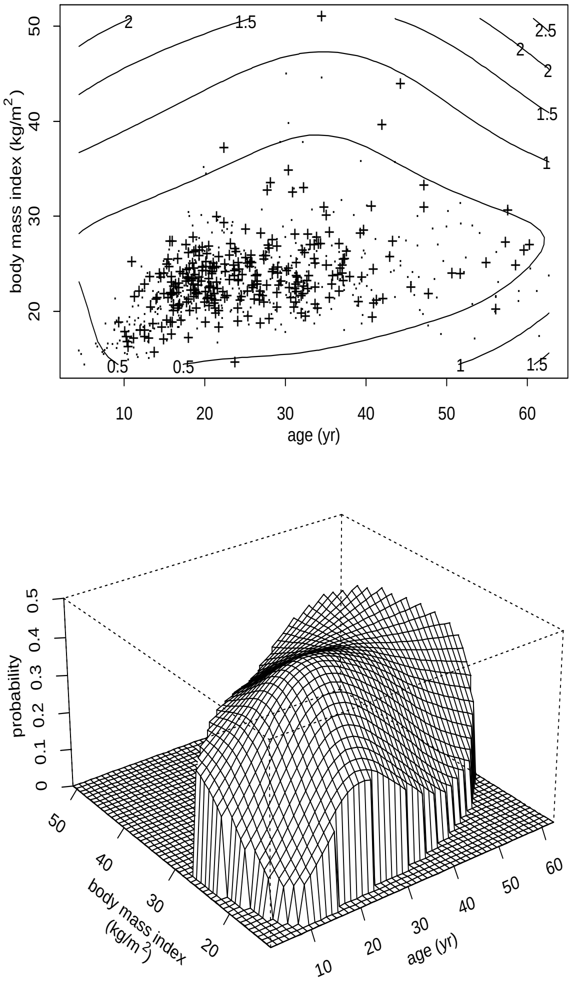

Figure 3.1: Top: Scatterplot of `bmi` and `age`. '+' indicates subjects whose diabetic retinopathy progressed, '·' are others. Bottom: Estimate of probability of progression, as a function of `bmi`, `age`.

in an appropriate function space. The component f's are smoothing splines, and the regularization or smoothing parameters λ_1, λ_3 and λ_{13} have been chosen according to a multivariate version of the iterative unbiased risk criteria in Gu (1992). Larger values of the λ's would have caused the plot to flatten out, and smaller values would have caused it to be 'wiggly'.

4 Choosing How Much to Regularize

At the time of this writing, it is a matter of lively debate and much research how to choose the various regularization parameters. Leaving out a large fraction of the training sample for this purpose and tuning the regularization parameter(s) to best predict the set-aside data (according to whatever criteria of best prediction is adopted) is conceptually simple, defensible, and widely used (this is called out-of-sample tuning). Successively leaving-out-one, successively leaving-out-10% , and generalized cross validation are all popular. If the variance of the observational error is known then unbiased risk estimates become available. See Li (1986), Wahba (1990) and references cited there, and Gu (1992) for these 'in-sample' tuning methods. When there is a Bayesian model behind the regularization procedure, then maximum likelihood estimates may be derived, see for example Wahba (1985), although in order for these and other Bayes estimates to do a good job of minimizing the generalization error in practice, it is usually necessary that the priors on which they are based are realistic.

5 Which method is best?

Feedforward neural nets, radial basis functions, and various forms of splines all provide regularized or regularizable methods for estimating 'smooth' functions of several variables, given a training set $\{\boldsymbol{t}(i), y_i\}$: Which approach is best? Unfortunately, there is not, nor is there likely to be, a single answer to that question. The answer most surely depends on the particular nature of the underlying but unknown 'truth' , the nature of any prior information that might be available about this 'truth' the nature of any noise in the data, the ability of the experimenter to choose the various smoothing or regularization parameters well, the size of the data set, the use to which the answer will be put, and the computational facilities available. From a mathematical point of view, the classes of functions well approximated by neural nets, radial basis functions and sums and products of splines (ANOVA splines) are not the same, although all of these methods have the capability of approximating large classes of functions. Of course, if a large enough data set is available, models utilizing all of these approaches may be built, and tuned, and then compared on data that has been set aside for this purpose.

6 Acknowledgments

This work was supported by NSF Grant DMS 9121003 and NIH Grant R01 EY09946.

References

*Eubank, R. (1989), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker.

Friedman, J. (1991), 'Multivariate adaptive regression splines', *Ann. Statist.* **19**, 1–141.

*Geman, S., Bienenstock, E. & Doursat, R. (1992), 'Neural networks and the bias/variance dilemma', *Neural Computation* **4**, 1–58.

Girosi, F., Jones, M. & Poggio, T. (1993), 'Priors, stabilizers and basis functions: from regularization to radial, tensor and additive splines. Artificial Intelligence Laboratory, M. I. T., A. I. Memo No. 1430.

Gu, C. (1992), 'Cross-validating non-Gaussian data', *J. Comput. Graph. Stats.* **1**, 169–179.

Gu, C. & Wahba, G. (1993), 'Semiparametric analysis of variance with tensor product thin plate splines', *J. Royal Statistical Soc. Ser. B* **55**, 353–368.

*Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman and Hall.

Kimeldorf, G. & Wahba, G. (1970a), 'A correspondence between Bayesian estimation of stochastic processes and smoothing by splines', *Ann. Math. Statist.* **41**, 495–502.

Li, K. C. (1986), 'Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing', *Ann. Statist.* **14**, 1101–1112.

*Ripley, B. (1994), 'Neural networks and related methods for classification', *J. Roy. Statist. Soc. Ser. B* **56** *to appear*

Wahba, G. (1985), 'A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem', *Ann. Statist.* **13**, 1378–1402.

*Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, Society for Industrial and Applied Mathematics.

Wahba, G. (1992), Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, *in* M. Casdagli & S. Eubank, eds, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII*, Addison-Wesley, pp. 95–112.

*Wahba, G. & Wold, S. (1975), 'A completely automatic French curve', *Commun. Stat.* **4**, 1–17.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1994), Structured machine learning for 'soft' classification with smoothing spline ANOVA and stacked tuning, testing and evaluation, *in* D. Wolpert, ed, *The Mathematics of Generalization, SFI Studies in the Sciences of Complexity, Proc. Vol XIX*, Addison-Wesley, to appear.

Weigend, A. (1993), On overfitting and the effective number of hidden units, *in* M. Mozer, P. Smolensky, D. Touretzky, J. Elman & A. Weigend, eds, *Proceedings of the 1993 Connectionist Models Summer School*, Erlbaum Associates, 335–342.