The Rôle of the Loss Function in Probabilistic Function Approximation

Tony J. Dodd[1], Robert F. Harrison[1] and Chris J. Harris[2]
[1]Department of Automatic Control and Systems Engineering
The University of Sheffield, Sheffield S1 3JD, UK
e-mail: {t.j.dodd, r.f.harrison}@shef.ac.uk
[2]Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK

Research Report No. 820
April 2002

Abstract

Generalising results on time series estimation it is natural to consider function approximation with finite data observations in a probabilistic setting. The function is treated as a stochastic process where for each point in the functions domain the function is a random variable. Equivalently the function can be considered as a single random variable whose range is a space of functions. In this paper two results, well known within the context of time series estimation and stochastic control, are generalised to probabilistic function approximation problems. Under mild conditions on the space of functions it is shown that the optimal function estimate corresponds, for all reasonable symmetrical loss functions, to the pointwise conditional expectation given the observed data. Further, in the case where the space of functions belongs to the class of Gaussian processes the optimal estimate is the conditional expectation even for asymmetric loss functions.

1 Introduction

Generalising results on time series estimation it is natural to consider function approximation with finite data observations in a probabilistic setting. The function is treated as a stochastic process where for each point in the function domain the function is a random variable. Equivalently the function can be considered as a single random variable whose range is a space of functions. In describing the space of functions as a stochastic process we make use of the finite dimensional distributions of points on the function. For example, the Gaussian case leads to a particularly rich theory. The Gaussianity of the space of functions refers to the correlation properties of points on the function surface. So for smooth functions we express a preference for close points in parameter (input) space to also be close in output space. We then have an appropriate probability space of functions.

Our problem is then, given such a probability space of functions and a set of observations of the function at known points, to infer the value of the function at unknown points. Although we may interpret the function as a single random variable we do not use the observations to infer directly a single function from this space. Instead, the function is inferred indirectly by estimating values for specific points on the function - there is no initial step where a functional relationship is formed. Obviously the whole function can be inferred by estimating values over the whole input space.

In order to distinguish between possible estimates it is necessary to compare the estimates using a loss function. Given the error in the estimate a loss function is defined which is positive and a nondecreasing function of the error. A natural estimate then corresponds to the minimum of the expected loss. Traditionally the loss function is chosen to be the mean-squared error which is simple to handle, optimal for Gaussian noise, and always gives the best linear estimate (although better nonlinear estimators may be possible). More often than not alternative loss functions are deemed too difficult to handle mathematically, although they may have practical merit.

In this paper two results, well known within the context of linear filtering, are generalised to probabilistic function approximation problems. Under mild conditions on the space of functions it is shown that the optimal function estimate corresponds, for all reasonable symmetrical loss functions, to the pointwise conditional expectation given the observed data. Further, in the case where the space of functions belongs to the class of Gaussian processes the optimal estimate is the conditional expectation even for asymmetric loss functions.

The issue of non-mean-squared error loss functions within linear filtering was first addressed by Benedict and Sondhi (1957) under the assumptions of Gaussianity and polynomial loss functions. It was shown that the optimum Wiener filter was identical to that found in the case of mean-squared error. More generally Sherman (1955, 1958) considered the case of symmetric loss functions and symmetric, unimodal, probability distributions. Brown (1962) specialised these results by assuming that the probability distribution is Gaussian. In this case a much wider, asymmetric class of loss functions is allowable. Generalising

in a different direction, if the loss function is also assumed to be symmetric and convex, the restriction on the unimodality of the conditional probability distribution can be relaxed (Hall and Wise 1991).

These results are well known in the context of linear filtering (Kalman 1960; Deutsch 1965; Jazwinski 1970) and have been applied to optimal control (Harris 1992). The purpose of this paper is to present the results in the more general context of probabilistic function approximation. The results are then directly applicable to the Gaussian process approach to function approximation (Williams 1999) and also to support vector machines (Vapnik 1998) in the presence of Gaussian noise.

In the next section the probabilistic setting for function estimation is described based on the idea of stochastic processes. The particular case of Gaussian stochastic processes is described in detail. In Section 3 optimal estimation is introduced and shown to correspond to the conditional expectation in the case of a squared error loss function. The key results of the paper relating to symmetric and asymmetric loss functions are then presented in Section 4. The particular example of polynomial loss functions for Gaussian processes is addressed in Section 5.

2 Probabilistic Function Estimation

We first give a concrete description of the class of problems to be solved. We assume there exists some (unknown) function (signal), $y(x)$, and noise, $n(x)$. Neither can be observed directly, instead we observe $z(x) = y(x)+n(x)$. Suppose we make N observations giving rise to the values of $z(x_1), \ldots, z(x_N)$. Given this information, the problem is then to make inferences about the (unobservable) values of the function, $y(x)$, for arbitrary x. As an example case, assume $x = t \in \mathbb{R}$ is time and the $z(t_1), \ldots, z(t_N), t_1 < \cdots < t_N$ constitute a time series of observations. Three estimation problems can then be envisaged: (i) if $t < t_N$ this is a smoothing problem; (ii) for $t = t_N$ it is called filtering; and (iii) when $t > t_N$ this is a prediction problem. More generally, for $x \in \mathbb{R}^N$, such ordering is not possible. The problem is then simply referred to as estimation, or approximation when we wish to emphasise that $y(x)$ is some unknown function.

A natural setting for the estimation problem is probability theory and statistics. The function, noise, and observations are then regarded as stochastic processes. Given a probabilistic description of these processes we can then determine the probability with which particular values of the function and noise will occur. We now review the definition and interpretations of stochastic processes.

2.1 Stochastic Processes

Formally, a stochastic process, $z(x)$, is a collection of random variables, defined on a common probability space, $(\Omega, \mathcal{F}, P)$, and indexed by the elements of a parameter set $\mathcal{X}$. In general, the stochastic process takes values in $\mathbb{R}^p$ (a vector valued process), however we consider only the scalar real-valued case. Common

examples of $\mathcal{X}$ include $\mathbb{R}, \mathbb{R}^+, \mathbb{Z}, \mathbb{Z}^+$, where the process is often referred to as a time series. We consider the case where $\mathcal{X} \subseteq \mathbb{R}^n$, for which the stochastic process is often called a random field.

Various interpretations of stochastic processes are meaningful (Lamperti 1977; Øksendal 1998). Most often, the process is regarded as a function on $\mathcal{X}$ such that for each $x \in \mathcal{X}$ the value of z is a random variable. Since, strictly $z = z(x, \omega)$ where $x \in \mathcal{X}, \omega \in \Omega$, this interpretation as a random variable means $\omega \to z(x, \cdot)$, where for each fixed x, the function $z(x, \cdot)$ is measurable with respect to $\mathcal{F}$. The notion of a sample function (trajectory or sample path) arises if, instead, we fix $\omega \in \Omega$ obtaining the function $z(\cdot, \omega) : \mathcal{X} \to \mathbb{R}$. For our purposes, though, it is also meaningful to consider z as a single random variable $z(x, \omega)$ whose range is a space of functions of $\mathcal{X}$. In this case we talk of a random function.

Many of the important properties of stochastic processes can be determined by the family of all finite-dimensional distributions. Let $x_1, \ldots, x_N \in \mathcal{X}$ then the collection

$$P(\xi) = P\{z(x_1) \leq \xi_1, \ldots, z(x_N) \leq \xi_N\} \tag{1}$$

as x ranges over all vectors of members of $\mathcal{X}$ of any finite length, is called the collection or family of finite-dimensional distributions of z. This family contains all the information which is available about z from the distributions of its constituent variables, $z(x)$. Given such a collection, Kolmogorov's extension theorem ensures that, under certain mild consistency conditions, a stochastic process exists having $P(\xi)$ as its finite-dimensional distributions (Lamperti 1977).

In this paper we will be concerned solely with the stationary case. For $\mathcal{X} = \mathbb{R}, \mathbb{R}^+, \mathbb{Z}$ or $\mathbb{Z}^+$, a stochastic process is stationary if the finite dimensional distributions of the process are invariant under translations of time (subsets of the real line). That is, z is stationary when

$$P\{z(x_1) \leq \xi_1, \ldots, z(x_N) \leq \xi_N\} = P\{z(x_1+\Delta) \leq \xi_1, \ldots, z(x_N+\Delta) \leq \xi_N\} \tag{2}$$

for every $x_1, \ldots, x_N, \Delta \in \mathcal{X}$. More generally, for $\mathcal{X} \subseteq \mathbb{R}^n$ the process is stationary if Eq. 2 holds for all n-vectors Δ in $\mathcal{X}$.

2.2 Gaussian Processes

Consider now a stochastic process $z(x)$ with $E\{|z(x)|^2\} < \infty$ for all $x \in \mathcal{X}$. Such stochastic processes are known as second order processes. We can then define the covariance function of the process

$$k(x, x') = E\{z(x)z(x')\} \tag{3}$$

for all $x, x' \in \mathcal{X}$.

Definition 2.1 *A function $k(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $x, x' \in \mathcal{X}$, is non-negative*

definite (or equivalently positive semi-definite) iff

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0$$

for any distinct $x_1, \ldots, x_N$ and scalars $c_1, \ldots, c_N$. If strict inequality holds unless all the c_i's are zero, k is said to be positive definite.

Theorem 2.1 *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance of some process $\{z(x)\}$ iff k is non-negative definite. k is positive definite iff the random variables $\{z(x)\}$ are linearly independent.*

Proof See, for example, Lamperti (1977). □

Definition 2.2 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable $z^N = [z(x_1), \ldots, z(x_N)] : \Omega \to \mathbb{R}^N$ has the multivariate normal distribution $N(\mu, K)$ if the distribution of z^N has a density of the form*

$$p(z^N) = p(z(x_1), \ldots, z(x_N)) = \frac{1}{\sqrt{(2\pi)^N |K|}} \exp\left\{-\frac{1}{2}(z^N - \mu)K^{-1}(z^N - \mu)^T\right\}$$

where K is a positive definite symmetric matrix.

Theorem 2.2 *If $[z(x_1), \ldots, z(x_N)]$ is $N(\mu, K)$ then*

1. *$E\{z(x_1), \ldots, z(x_N)\} = \mu$, which is to say $E\{y(x_i)\} = \mu_i$ for all i; and*
2. *the covariance matrix K is such that $[K]_{ij} = E\{(z(x_i) - \mu_i)(z(x_j) - \mu_j)\})$.*

Proof See, for example, Grimmett and Stirzaker (1992). □

Definition 2.3 *A stochastic process $z(x)$, $x \in \mathcal{X}$, is called Gaussian if every finite linear combination of the random variables $z(x)$ is normally distributed. Equivalently, every finite dimensional vector $[z(x_1), \ldots, z(x_N)]$ is multivariate normally distributed for all N.*

A Gaussian process is necessarily a second order process.

Theorem 2.3 *Suppose we have some function k, non-negative definite, such that for each finite set $x_1, \ldots, x_N \in \mathcal{X}$ the matrix $[K]_{ij} = k(x_i, x_j)$ is symmetric and non-negative definite. Then there exists a Gaussian process having zero mean and k for its covariance function.*

Proof This follows from Kolmogorov's extension theorem, see, for example, Lamperti (1977) and Grimmett and Stirzaker (1992). □

3 Optimal Estimation

Given a set of observations, $\eta(x_1), \ldots, \eta(x_N)$ of the stochastic process $z(x)$ the probability of occurrence of values $\xi(x)$ of the stochastic process $y(x)$ is given by the conditional probability distribution function

$$P[y(x) \leq \xi | z(x_1) = \eta(x_1), \ldots, z(x_N) = \eta(x_N)] = F(y). \tag{4}$$

Clearly, $F(y)$ embodies all the (statistical) information about $y(x)$ which is contained in the available observations.

A statistical estimate of the random function, $y(x)$, will be some function of the distribution function, $F(y)$, and therefore a function of the random observation variables, $z(x_1), \ldots, z(x_N)$. We denote this estimate by $y(x|Z^N)$ which is itself a stochastic process whose value is known whenever the values of $z(x_1), \ldots, z(x_N)$ are known. In general, the value of $y(x|Z^N)$ will be different from the (unknown) value of $y(x)$. We therefore need a criterion for assessing which is the best possible estimate. Define the error in the estimate by

$$e = y(x) - y(x|Z^N). \tag{5}$$

As a criterion we define a loss function $L(\cdot)$ on e which is (i) positive, and (ii) a non-decreasing function of e. A natural estimate $y(x|Z^N)$ of $y(x)$ is that which minimises the average or expected loss

$$E\{L[y(x) - y(x|Z^N)]\} = E[E\{L[y(x) - y(x|Z^N)]|z(x_1), \ldots, z(x_N)\}] \tag{6}$$

where the outer expectation on the RHS is over all possible observation sets. However, since this expectation does not depend on $y(x|Z^N)$ (which is already conditional on $z(x_1), \ldots, z(x_N)$), but only on $z(x_1), \ldots, z(x_N)$, then minimising Eq. 6 is equivalent to minimising

$$E\{L[y(x) - y(x|Z^N)]|z(x_1), \ldots, z(x_N)\}. \tag{7}$$

3.1 Least Squares Estimation

The basic, and simplest, case is to consider $L(e) = e^2$, known as least squares estimation, for which the optimal estimate can be found straightforwardly.

Theorem 3.1 *Assume that $L(e) = e^2$ and $y(x), z(x_1), \ldots, z(x_N)$ are any random variables with $E\{|y(x)|^2\} < \infty$ then the random variable $y^*(x|Z^N)$ which minimises the expected loss, Eq. 6, is the conditional expectation*

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}. \tag{8}$$

Proof We can write

$$E\{[y(x) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} = \\ E\{[y(x) - y^*(x|Z^N)] + [y^*(x|Z^N) - y(x|Z^N)]|z(x_1), \ldots, z(x_N)\}$$

which can be expanded thus

$$E\{[y(x) - y^*(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} + \\ 2E\{[y(x) - y^*(x|Z^N)][y^*(x|Z^N) - y(x|Z^N)]|z(x_1), \ldots, z(x_N)\} + \\ E\{[y^*(x|Z^N) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\}.$$

But $y^*(x|Z^N) - y(x|Z^N)$ is orthogonal to every function $y(x)$ which is measurable on the sample space $z(x_1), \ldots, z(x_N)$ and whose square is integrable (Doob 1953) (Theorem 8.3, p.22). Hence

$$E\{[y(x) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} = \\ E\{[y(x) - y^*(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} \\ + E\{[y^*(x|Z^N) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\}.$$

The first term on the RHS is unaffected by the choice of $y(x|Z^N)$, and for all $y(x|Z^N)$ is always positive. Hence the LHS is minimised by setting

$$y(x|Z^N) = y^*(x|Z^N). \tag{9}$$

□

A less rigorous argument follows by expanding the expected loss, Eq. 7, as

$$E\{[y(x) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} = \\ E\{y^2(x)|z(x_1), \ldots, z(x_N)\} - 2E\{y(x)y(x|Z^N)|z(x_1), \ldots, z(x_N)\} \\ + E\{y^2(x|Z^N)|z(x_1), \ldots, z(x_N)\}.$$

But $y(x|Z^N)$ is already conditioned on $z(x_1), \ldots, z(x_N)$ and therefore

$$E\{[y(x) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\} = \\ E\{y^2(x)|z(x_1), \ldots, z(x_N)\} - 2y(x|Z^N)E\{y(x)|z(x_1), \ldots, z(x_N)\} + y^2(x|Z^N).$$

Differentiating with respect to $y(x|Z^N)$ and equating to zero

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}. \tag{10}$$

3.2 Conditional Expectation with Gaussian Processes

Assume that $y(x)$, $n(x)$ and $z(x)$ are Gaussian. Consider the $N+1$ random variables, $z(x_1), \ldots, z(x_N), z(x)$, which have a joint Gaussian distribution with constant mean, μ, and covariance matrix, Σ. Let the $(N+1) \times (N+1)$ matrix Σ be partitioned as follows

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{11}$$

where $\Sigma_{11} \in \mathbb{R}^{N \times N}$, $\Sigma_{12} = \Sigma_{21}^T \in \mathbb{R}^N$ and Σ_{22} is a scalar.

The random variables $z^N = [z(x_1), \ldots, z(x_N)]^T$ and $z' = z(x) - \Sigma_{12}^T \Sigma_{11}^{-1} z^N$ are statistically independent and

$$E\{z'\} = \mu - \Sigma_{12}^T \Sigma_{11}^{-1} \mu_N, \quad E\{(z' - E\{z'\})(z' - E\{z'\})\} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \quad (12)$$

where $\mu_N = E\{z(x_1), \ldots, z(x_N)\}$, $\mu = E\{z(x)\}$. The proofs follow standard results (Mardia et al. 1979).

Since z' is independent of z^N, its conditional distribution for a given value of z^N is the same as its marginal distribution, Eq. 12. Rearranging, $z(x)$ is equal to $z' + \Sigma_{12}^T \Sigma_{11}^{-1} z^N$ where the second term is constant for given z^N. By substituting for z' and simplifying, the conditional mean of $z(x)$ given $z(x_1), \ldots, z(x_N)$ is given by

$$E\{z(x)|z(x_1), \ldots, z(x_N)\} = \mu + \Sigma_{12}^T \Sigma_{11}^{-1} (z^N - \mu_N) \quad (13)$$

and the conditional variance of $z(x)$ is the same as that of z', i.e.

$$E\{(z(x) - E\{z(x)|z(x_1), \ldots, z(x_N)\})^2 | z(x_1), \ldots, z(x_N)\} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}. \quad (14)$$

If we further assume that $n(x)$ is zero-mean and independent of $y(x)$ then $E\{z(x)|z(x_1), \ldots, z(x_N)\} = E\{y(x)|z(x_1), \ldots, z(x_N)\}$. Assume $n(x)$ is zero-mean with

$$E\{n(x)n(x')\} = \sigma_n^2 \delta(x - x'), \quad (15)$$

therefore $y(x)$ has the same mean as $z(x)$, i.e. μ, and

$$E\{(y(x) - E\{y(x)\})(y(x') - E\{y(x')\})\} = k(x, x'). \quad (16)$$

Then

$$E\{y(x)|z(x_1), \ldots, z(x_N)\} = \mu + k^T (K + \sigma_n^2 I)^{-1} (z^N - \mu_N) \quad (17)$$

and

$$E\{(y(x) - E\{y(x)|z^N\})^2 | z(x_1), \ldots, z(x_N)\} = k(x, x) + \sigma_n^2 - k^T (K + \sigma_n^2 I)^{-1} k \quad (18)$$

where $k = [k(x, x_1), \ldots, k(x, x_N)]^T$ and $[K]_{ij} = k(x_i, x_j)$.

4 Alternative Loss Functions

We now come to the main results of the paper. In this section more general classes of loss function are considered under different restrictions on the conditional probability distribution. For the given assumptions the optimal estimate always corresponds to the conditional expectation as would be found in the least squares case.

4.1 Symmetric Loss Functions

Definition 4.1 *The class of symmetric loss functions, $L(e)$, satisfy the following:*

$$L(0) = 0$$
$$L(e_2) \geq L(e_1) \geq 0 \text{ when } e_2 \geq e_1 \geq 0$$
$$L(e) = L(-e).$$

Examples include $L(e) = ae^2, ae^4, a|e|$ for $a \in \mathbb{R}^+$.

Theorem 4.1 *Assume that L satisfies Definition 4.1 and that the conditional distribution $F(y)$, defined by Eq.4, is:*

1. *symmetric about the mean $\bar{y}$:*

$$F(y - \bar{y}) = 1 - F(\bar{y} - y)$$

2. *convex for $y \leq \bar{y}$:*

$$F(\lambda y_1 + (1 - \lambda)y_2) \leq \lambda F(y_1) + (1 - \lambda)F(y_2)$$

for all $y_1, y_2 \leq \bar{y}$ and $0 \leq \lambda \leq 1$.

Then the random function, $y^*(x|Z^N)$, which minimises the expected loss, Eq. 6, is the conditional expectation

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}. \tag{19}$$

Proof See Sherman (1958) and Åström (1970). □

Note that conditions (1) and (2) of Theorem 4.1 are equivalent to the associated probability density function $p(y(x)|Z^N)$ being symmetric about the mean, $\bar{y}$, and unimodal.

Corollary 4.1 *If the stochastic processes $y(x)$, $n(x)$, and $z(x)$ are Gaussian, Theorem 4.1 holds.*

Proof Conditional distributions on a Gaussian process are Gaussian (Anderson 1984). Hence the requirements of Theorem 4.1 are always satisfied. □

4.2 Asymmetric Loss Functions

Consider further, the restriction whereby $y(x)$, $n(x)$, and $z(x)$ are Gaussian. We can then extend Theorem 4.1 to the class of asymmetric loss functions which are non-decreasing for $|e| \geq 0$ defined as follows.

Definition 4.2 *The class of asymmetric loss functions, $L(e)$, can be written as follows:*

$$L(e) = L_1(e) + L_2(e)$$

where

$$L_1(e) = 0 \text{ for } e \leq 0$$
$$0 \leq e_1 \leq e_2 \text{ implies } 0 \leq L_1(e_1) \leq L_1(e_2)$$

and

$$L_2(e) = 0 \text{ for } e \geq 0$$
$$e_1 \leq e_2 \leq 0 \text{ implies } 0 \leq L_2(e_2) \leq L_2(e_1).$$

We now present the analogous theorem for asymmetric loss functions with Gaussian conditional probability distributions. The proof is reproduced from in full owing to the importance of the result (Brown 1962).

Theorem 4.2 *Assume that L satisfies Definition 4.2, assume also that the stochastic processes $y(x)$, $n(x)$ and $z(x)$ are Gaussian. Then the random variable $Y^*(x|Z^N)$ which minimises the expected loss, Eq. 6, is the conditional expectation*

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}. \tag{20}$$

Proof Since $y(x)$ and $z(x)$ are Gaussian we must have $e = y(x) - y(x|Z^N)$ Gaussian and $E\{e\} = 0$ by hypothesis. Therefore the probability density function of e is given by

$$p(e) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{e^2}{2\sigma^2}\right\} \tag{21}$$

where, by definition,

$$\sigma = [E\{e^2\}]^{1/2} = [E\{[y(x) - y(x|Z^N)]^2|z(x_1), \ldots, z(x_N)\}]^{1/2}. \tag{22}$$

Now, by definition

$$E\{L(e)\} = \int_{-\infty}^{\infty} L(e)p(e)de. \tag{23}$$

Substituting Eq. 21 in Eq. 23

$$E\{L(e)\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} L(e) \exp\left\{-\frac{e^2}{2\sigma^2}\right\} de. \tag{24}$$

By replacing throughout e with σe and therefore de by σde (the limits of integration remain unchanged) we have

$$\begin{aligned} E\{L(e)\} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} L(\sigma e) \exp\left\{-\frac{\sigma^2 e^2}{2\sigma^2}\right\} \sigma de \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} L(\sigma e) \exp\left\{-\frac{e^2}{2}\right\} de \\ &= \frac{1}{\sqrt{2\pi}} \left[\int_0^{\infty} L_1(\sigma e) \exp\left\{-\frac{e^2}{2}\right\} de + \int_{-\infty}^{0} L_2(\sigma e) \exp\left\{-\frac{e^2}{2}\right\} de\right] \end{aligned}$$

where L_1 and L_2 are as given in Definition 4.2.

Now, consider σ_1 and σ_2 such that $0 \leq \sigma_1 \leq \sigma_2$. From the monotonicity requirements on L_1 and L_2 (Definition 4.2)

$$e \geq 0 \Rightarrow L_1(\sigma_1 e) \leq L_1(\sigma_2 e)$$

and

$$e \leq 0 \Rightarrow L_2(\sigma_1 e) \leq L_2(\sigma_2 e).$$

Therefore $0 \leq \sigma_1 \leq \sigma_2$ implies

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} L(\sigma_1 e) \exp\left\{-\frac{e^2}{2}\right\} de \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} L(\sigma_2 e) \exp\left\{-\frac{e^2}{2}\right\} de. \quad (25)$$

This states that, when considered as a function of σ,

$$E\{L(e)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} L(\sigma e) \exp\left\{-\frac{e^2}{2}\right\} de \quad (26)$$

is a nondecreasing function of σ and also, therefore, of σ^2. Since $E\{L(e)\}$ therefore varies monotonically with $E\{e^2\}$ it is clear that $y^*(x|Z^N)$ which minimises $E\{e^2\}$ will also minimise $E\{L(e)\}$. But

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}$$

minimises $E\{e^2\}$, therefore it also minimises $E\{L(e)\}$. □

5 Example: Polynomial Loss

As a particular example, consider the class of polynomial loss functions (Benedict and Sondhi 1957)

$$L(e) = |e|^\alpha \quad (27)$$

where $\alpha \in \mathbb{R}^+$. Again, assuming $y(x)$ and $z(x)$ are Gaussian we must have $e = y(x) - y(x|Z^N)$ Gaussian and $E\{e\} = 0$ by hypothesis, hence

$$\begin{aligned} E\{L(e)\} &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |e|^\alpha \exp\left\{-\frac{e^2}{2\sigma^2}\right\} de \\ &= \frac{2}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} |e|^\alpha \exp\left\{-\frac{e^2}{2\sigma^2}\right\} de \end{aligned}$$

by symmetry. This integral can be evaluated as (Barnett and Cronin 1986)

$$E\{L(e)\} = \frac{2^{\alpha/2}}{\sqrt{\pi}} \Gamma\left(\frac{\alpha+1}{2}\right) (\sigma^2)^{\alpha/2} \quad (28)$$

where, by definition

$$\sigma = [E\{e^2\}]^{1/2} = [E\{[y(x) - y(x|Z^N)]^2 | z(x_1), \ldots, z(x_N)\}]^{1/2}$$

and Γ is the usual Gamma function.

Differentiating Eq. 28 with respect to $y(x|Z^N)$ and equating to zero

$$\left[\frac{2^{\alpha/2}\alpha}{\sqrt{\pi}} \Gamma\left(\frac{\alpha+1}{2}\right) (\sigma^2)^{\alpha/2-1}\right] [-2E\{y(x)|z(x_1), \ldots, z(x_N)\} + 2y(x|Z^N)] = 0. \quad (29)$$

The first term will always be greater than zero as $\sigma^2 > 0$ and therefore a necessary and sufficient condition for a minimum is that the second term vanishes, i.e.

$$y^*(x|Z^N) = E\{y(x)|z(x_1), \ldots, z(x_N)\}. \quad (30)$$

6 Conclusions

A general framework for probabilistic function estimation from finite data has been described. The function is treated either as a stochastic process, where for each input point the function is a random variable, or equivalently as a single random variable whose range is a space of functions. In practise it is the former which is used. In order to distinguish between estimates it is necessary to compare them using a loss function. In the least squares case, and under no restrictions on the probability space of functions, the optimal estimate corresponds to the conditional expectation given the available observations. Further, it was shown that for symmetric loss functions and a symmetric (about the mean), unimodal, probability distribution of functions, the optimal estimate is still the conditional expectation. In the Gaussian case this result was shown to be further relaxed to include asymmetric loss functions. To demonstrate these results the particular case of a polynomial loss function and Gaussian space of functions was derived.

Acknowledgements

The authors would like to thank the UK EPSRC for their financial support under Grant No. GR/R15726/01.

References

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis* (Second ed.). John Wiley & Sons.

Åström, K. J. (1970). *Introduction to Stochastic Control Theory*, Volume 70 of *Mathematics in Science and Engineering*. Academic Press.

Barnett, S. and T. Cronin (1986). *Mathematical Formulae for Engineering and Science Students* (Fourth ed.). Addison Wesley Longman.

Benedict, T. and M. Sondhi (1957). On a property of Wiener filters. *Proceedings of the IRE 45*, 1021–1022.

Brown, J. (1962). Asymmetric non-mean-square error criterion. *IRE Transactions on Automatic Control 7*, 64–66.

Deutsch, R. (1965). *Estimation Theory*. Monographs and Textbooks in Pure and Applied Mathematics. Prentice-Hall.

Doob, J. (1953). *Stochastic Processes*. John Wiley & Sons.

Grimmett, G. and D. Stirzaker (1992). *Probability and Random Processes* (Second ed.). Clarendon Press.

Hall, E. B. and G. L. Wise (1991). On optimal estimation with respect to a large family of cost functions. *IEEE Transactions on Information Theory 37*(3), 691–693.

Harris, T. J. (1992). Optimal controllers for nonsymmetric and nonquadratic loss functions. *Technometrics 34*(3), 298–306.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*, Volume 64 of *Mathematics in Science and Engineering*. Academic Press.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering Series D 82*, 35–45.

Lamperti, J. (1977). *Stochastic Processes: A Survey of Mathematical Theory*, Volume 23 of *Applied Mathematical Sciences*. Springer-Verlag.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press.

Øksendal, B. (1998). *Stochastic Differential Equations: an Introduction with Applications* (Fifth ed.). Springer.

Sherman, S. (1955). A theorem on convex sets with applications. *Annals of Mathematical Statistics 26*, 763–767.

Sherman, S. (1958). Non-mean-square error criteria. *IRE Transactions on Information Theory 4*, 125–126.

Vapnik, V. (1998). *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons.

Williams, C. (1999). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning in Graphical Models*, pp. 599–621. The MIT Press.