STATISTICAL LIKELIHOOD REPRESENTATIONS OF PRIOR KNOWLEDGE IN MACHINE LEARNING

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

Leszek Plaskota
Department of Mathematics
Warsaw University
02-097 Warsaw, Poland
email: leszekp@hydra.mimuw.edu.pl

Andrzej Przybyszewski
McGill University, Montreal, Canada, and
University of Massachusetts, Worcester, MA
email: przy@ego.psych.mcgill.ca

ABSTRACT
We show that maximum a posteriori (MAP) statistical methods can be used in nonparametric machine learning problems in the same way as their current applications in parametric statistical problems, and give some examples of applications. This MAPN (MAP for nonparametric machine learning) paradigm can also reproduce much more transparently the same results as regularization methods in machine learning, spline algorithms in continuous complexity theory, and Baysian minimum risk methods.

KEY WORDS
machine learning, Bayesian statistics

1 Introduction

Machine learning and artificial neural network theory often deal with the problem of learning an input-output (i-o) function from examples, i.e., from partial information ([1], [2], [3], [4], [5],[6], [7]).

Given an unknown i-o function $f(\mathrm{x})$, along with examples $Nf = (f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n))$, the goal is to learn f. In this paper we describe an extension of standard maximum a posteriori (MAP) methods in parametric statistics to this (nonparametric) machine learning problem. Consider the problem ([6], [7]) of recovering f from a hypothesis space F of possible functions using information Nf , or more general information $Nf = (L_1 f, \ldots, L_n f)$, with L_i general functionals on f (e.g., Fourier coefficients). This problem occurs in machine learning, statistical learning theory ([6], [7])), information-based complexity, nonparametric Bayesian statistics ([8],[9]), optimal recovery [10],, and data mining [11].

Since we will assume little about the reader's knowledge of this area, we will include some basic examples and definitions.

To give an example of this type of nonparametric machine learning, we might be seeking a function $f(\mathrm{x})$ [12] which represents a relationship between inputs and outputs in a chemical mixture. Suppose we are building a control system in which homeostatic parameters, including temperature, humidity and amounts of chemical components of an industrial mixture can be controlled as input variables. Suppose we want to control the output variable y, which is the ratio of strength to brittleness of the plastic produced from the mixture. We want to build a machine which has the above input variables $\mathrm{x} = (x_1, x_2, ..., x_n)$ and whose output predicts the correct ratio y. The machine will use experimental data points $y = f(\mathrm{x})$ to learn from previous runs of the equipment. We may already have a prior model for f based on simple assumptions on the relationships of the variables. We then want to combine this prior information with that from the several runs we have made of our experiment.

Learning an unknown i-o function f from a high dimensional hypothesis space F is a nonparametric statistical problem – inference from the data Nf is done from a very large set of possibilities. We will show that standard parametric MAP estimation algorithms can be extended directly to a MAP for nonparametric machine learning (MAPN). The method presented here is simple and intuitively appealing in spite of the high dimensionality of the problems, and its estimates under standard hypotheses coincide with those obtained by other methods, e.g., optimal recovery [10], information based complexity [3], and statistical learning theory [7], as will be demonstrated below.

MAPN is a Bayesian learning algorithm which assumes prior knowledge given by a probability distribution μ for the unknown function $f \in F$, representing information about f (we assume here that F is a normed linear space). An example of Bayesian prior knowledge of the type mentioned above would be the stipulation that the probability distribution μ on F is a Gaussian measure centered at f. Examples of common a priori measures on a hypothesis space F include Gaussian and elliptically contoured measures ([3], [13], [8]).

The most important new element in this extension of MAP to hypothesis spaces of functions is a proof [14] that it is possible to define density functions $\varrho(f)$ corresponding to measures μ in a way analogous to how this is done in

finite dimensional estimation – see below.

The MAPN algorithm, as does MAP, will then use the density function $\rho(f)$ corresponding to this measure [14], and maximize it over the set $N^{-1}(z)$ of functions $f \in F$ consistent with the data z, yielding the MAPN estimate (or do the same with an assumption of some measurement error; see below).

The key issue in the development of the MAPN algorithm is that, as in MAP, we require existence of a density function ρ for μ. As is well known, such a density can exist for measures μ on finite dimensional F, i.e., when the number of parameters to be estimated is finite. In many data mining applications, however, as in the above example, an entire function f (i.e., an infinite number of parameters) must be extrapolated from data, and the corresponding infinite dimensional parameter space F presents an obstacle to extending MAP, since measures in infinite dimension do not admit density functions in the standard sense. This is because the density function $\rho(f)$ is the derivative (with respect to Lebesgue measure) of the a priori probability measure μ, and Lebesgue measure fails to exist in infinite dimension. Thus it has up to now been natural to assume that a probability density for f cannot be defined or maximized if F is a function space.

The method for defining a density function for a prior measure μ on an infinite dimensional F in fact can be accomplished in a way which is interestingly analogous to the finite dimensional case. In the latter situation,

$$d\mu(f) = \rho(f)\, df$$

with *df* Lebesgue measure (which exists only in finite dimension).

We show that it is possible to construct densities ρ for such measures also in infinite dimensional spaces F. More generally, we can show such densities can even exist for finitely additive measures, e.g., isonormal Gaussian measure on F with covariance operator $C = \frac{1}{2}I$.

Under finite dimensional Bayesian inference with prior density $\rho(f)$, the MAP estimate $\widehat{f}$ is the maximizer $\widehat{f}$ of $\rho(f)$, subject to the data z:

$$\widehat{f} = \arg\max_{f \in N^{-1}z} \rho(f).$$

Likelihood functions have some significant advantages, such as ease of use, ease of maximization, and ease of conditioning when further information (e.g., data $Nf = y$) becomes available.

As mentioned above, the lack of a density $\rho(f)$ in infinite dimensional hypothesis spaces F is based on the lack of a Lebesgue measure. However, Lebesgue measure is required only to define sets of "same size" at different locations (probabilities are then compared). This can be accomplished in other ways in infinite dimensional hypothesis spaces, as we show here.

We remark that the infinite dimensional nature of MAPN should be viewed as summarizing an inductive limit of finite dimensional approximation methods. The extent to which the algorithm is valid is determined by the validity of the same method for finite dimensional approximations. Here such approximations would entail approximating the space F of allowed i-o functions f as a finite dimensional space, say consisting of grid approximations of f. The validity of the MAPN procedure in infinite dimension states that there is a valid inductive limit of MAP algorithms for finite dimensional approximations of the desired f.

2 Invariant measures and density functions

Let μ be a probability measure on a normed linear space F. For estimation of an unknown $f \in F$ we will first consider how transformations of μ affect estimators. Consider the invariance properties of the measure μ under a transformation

$$T : F \to F\,.$$

If F is finite dimensional, we can define μ to be invariant with respect to T in two ways:

1. It can be *measure-invariant* with respect to T, so that

$$\mu(T^{-1}A) = \mu(A) \;\forall \text{ meas. } A$$

2. It can be *density-invariant* with respect to T, so that (at least if F finite dimensional)

$$\frac{d\mu}{d\lambda}(Tf) = \frac{d\mu}{d\lambda}(f) \text{ for all } x,$$

where $\rho(f) = \frac{d\mu}{d\lambda}(f) = \rho(f)$ is the density of f, with λ Lebesgue measure.

In a Bayesian setting, let μ denote the a priori distribution for the unknown f, and assume we have data $Nf = y$ regarding f. Then the expected error minimizing estimator of f is $\overline{f}_B = E(f|Nf = y)$ [3]. If T is *measure*-invariant then

$$T(\overline{f}_B) = \overline{T(f)_B},$$

i.e., the transform of estimator is the estimator of the transform. Thus average-case estimation ([3]; [15]) is invariant under T. However, this is not true for the MAP estimate.

If μ is density-invariant under T, then the MAP estimate is preserved under T, i.e.,

$$\arg\max \frac{d\mu}{d\lambda}(T^{-1}f) = T\left(\arg\max \frac{d\mu}{d\lambda}(f)\right).$$

Note that in finite dimension, if μ is density-invariant with respect to any T which preserves a norm $\|Ax\|$, then μ is *elliptically contoured* with respect A; essentially this means μ is a superposition of Gaussians whose covariance operators are scalar multiples of a single covariance (Traub, et al., 1987).

3 Definitions and basic results

Let F be finite dimensional, μ represent prior knowledge about $f \in F$, and define the density $\rho_\mu(f) = \frac{d\mu}{d\lambda}$ (with λ Lebesgue measure). Then we can also define ρ_μ (up to a multiplicative constant) by:

$$\frac{\rho_\mu(f)}{\rho_\mu(g)} = \lim_{\epsilon \to 0} \frac{\mu(B_\epsilon(f))}{\mu(B_\epsilon(g))}, \tag{1}$$

where $B_\epsilon(f)$ denotes the ϵ-ball centered at f. The first (derivative) definition does not extend to infinite dimension, but the second one on the right of 1 does. More generally, for cylinder (finitely additive) probability measures μ on F, we can define a density by $\frac{p_\mu(f)}{p_\mu(g)} = \lim_{R(N)\to\infty,\, \epsilon \to 0} \frac{\mu(B_{\epsilon,N}(f))}{\mu(B_{\epsilon,N}(g))}$, where $B_{\epsilon,N}(f)$ denotes the *epsilon-cylinder* at f of codimension n :

$$B_{\epsilon,N}(f) = \{f' |\ \|N(f - f')\|_2 \leq \epsilon\},$$

and N has finite rank, with $n = R(N)$ =rank(N) (with some technical conditions on the sequence N of finite rank operators).

Theorem (Kon, 2004): If μ_1 and μ_2 are two outer regular measures on F, and if $\frac{d\mu_2}{d\mu_1}(f)$ exists and is Lebesgue a.e. with respect to μ_1, then

$$r(f) \equiv \lim_{\epsilon \to 0} \frac{\mu_2(B_\epsilon(f))}{\mu_1(B_\epsilon(f))} \tag{2}$$

exists and is finite a.e. Then

$$r(f) = \frac{d\mu_2}{d\mu_1}(f),$$

almost everywhere.

Corollary: If μ has a density $\rho(f)$ and the derivative $\frac{d\mu(x-g)}{d\mu(x)}$ exists for all g, then

$$\frac{\rho(f_1)}{\rho(f_2)} = \left.\frac{d\mu(f - f_2 + f_1)}{d\mu(f)}\right|_{f=f_2}$$

For example, if μ is a Gaussian measure with positive covariance C, then $\rho(f) = e^{-\frac{1}{2}\|Af\|^2}$, where $C = A^{-2}$. By above, the case $C = I$ (cylinder measure only) is included as well.

4 Density viewpoint: density functions as a priori information

As mentioned earlier, it is sometimes attractive to define a measure μ representing prior knowledge about f which is only finitely additive, e.g. a cylinder measure, to properly reflect a priori information. As an example, consider a Gaussian μ, with covariance operator $C = I$. It is known that in infinite dimension, this is not a probability measure (i.e., it is not normalizable as a countably additive measure). Nevertheless, it has a density as a cylinder measure, as indicated above. This density $\rho(f) = e^{-\frac{1}{2}\|f\|^2}$ can reflect a priori knowledge about f in a precise way. Specifically, $\rho(f)$ reflects relative preferences for different choices of f. For example, the density $\rho(f) \equiv 1$, represents an a priori "Lebesgue measure" for f, i.e., indicating no preference, on any size space.

However, densities $\rho(f)$ as a priori information can incorporate other types of information, e.g., as *partial information* about a probability distribution or even non-countably additive probabilities e.g., cylinder measures (see e.g., [16]).

The use of a density which does not correspond to a measure can be illustrated with a simple example of inference with a priori information, which is in fact analogous to the situation in an infinite dimensional function space. Consider an unknown integer n. Suppose we only have the a priori knowledge that

$$P(n \text{ is even}) = 2P(n \text{ is odd})$$

Such information cannot be formulated in a single probability distribution on the integers, since such a distribution would have to be "uniform" on the evens and on the odds, and no such countably additive distribution exists. A likelihood function is needed to incorporate this information, i.e.,

$$\rho(n) = \begin{cases} 2 & \text{if } n \text{ even} \\ 1 & \text{if } n \text{ odd} \end{cases}$$

Likelihood functions play a similar role in an infinite dimensional space F, reflecting partial knowledge about f in a noncompact space where there is no "uniform" probability distribution. As an example in infinite dimension, suppose we know only that, given two choices f_1 and f_2 of f, the ratio of their probabilities is

$$\frac{e^{-\frac{1}{2}\|f_1\|^2}}{e^{-\frac{1}{2}\|f_2\|^2}},$$

it makes sense to define a likelihood function $e^{-\frac{1}{2}\|f\|^2}$, whether or not this corresponds to an actual measure. Thus likelihood functions are a more natural way than measures of incorporating a priori information in infinite dimensional Bayesian settings. This can also be seen in finite but high finite dimensional situations, in which a naive regularization approach to incorporating a priori information regarding an unknown f might be as follows.

Common methods for inference in statistical learning theory ([6], [7],[17]) involve regularization. Thus in addition to data $y = Nf$, a priori information might be: $\|Af\|$ should be small for a fixed linear A, where, e.g., A is a derivative operator, in which case $\|Af\|$ is a Sobolev norm.

For simplicity, assume $A = I$ is the identity and the norm is Euclidean. A naive approach might be to assume a prior distribution on the random variable $R = \|f\|$, say

$$\rho_R(R) = \frac{2}{\sqrt{\pi}} e^{-R^2} \ \ (R > 0). \tag{3}$$

Note this marginal for R corresponds to an n dimensional distribution of the form

$$\rho_f(f) = C_1 \frac{1}{\|f\|^{n-1}} e^{-\|f\|^2}, \tag{4}$$

if we assume $\rho_f(f)$ to be radially symmetric in f. This likelihood function is singular at the origin and clearly vanishes as $n \to \infty$ (so that it has no infinite dimensional limit), seemingly implying no likelihood methods in infinite dimension.

Compare this to the present likelihood function approach: we start with likelihood function $\rho_f(f) = C_2 e^{-\|Af\|^2}$ (above, $A = I$), which directly expresses intuition that a function f_1 is preferable to a function f_2 by a likelihood factor $\frac{e^{-\|Af_1\|^2}}{e^{-\|Af_2\|^2}}$. An added feature is that this likelihood function is the density of a measure on F if and only if A^{-2} is trace class [3].

The point here is that working with standard probability measures μ can be misleading, while expressing a priori information in likelihood functions clarifies a priori assumptions in practical situations.

In the infinite dimensional case, if we assume a priori likelihood function, e.g.,

$$\rho(f) = e^{-\|Af\|^2},$$

even if $A = I$, (so ρ is the density of an isonormal cylinder Gaussian measure), we understand the connection of $\rho(f)$ (and hence underlying probabilities) with a priori knowledge about likelihoods (see below).

5 Applications

MAPN in Bayesian estimation:

As indicated above, in infinite dimensional Bayesian inference the MAP estimator can be as useful as in parametric statistics. An unknown $f \in F$ with a Bayesian prior distribution on F now has a likelihood function, as in parametric statistics.

Gaussian prior: For example consider an infinite dimensional Gaussian prior measure μ on a Hilbert space F with covariance C. Assume without loss that C has dense range. We need:

Defining $A = \sqrt{C}^{-1}$, μ can be shown to have density $\rho_\mu(f) = e^{-\frac{1}{2}\|Af\|^2}$ [14]. Suppose we are given standard information $y = Nf = (f(x_1), \ldots, f(x_n))$. By the above, the conditional density $\rho_\mu(f|y)$ is the restriction of $\rho_\mu(f)$, so the MAPN estimate of f is:

$$\widehat{f} = \underset{Nf=y}{\arg\max} \quad e^{-\frac{1}{2}\|Af\|^2} = \underset{Nf=y}{\arg\min} \|Af\|$$

Note that this corresponds to the spline estimate of f [3], as well as the regularization theory estimate of f for exact (i.e., error-free) information [17], and Bayesian minimum average error estimates based on Gaussian priors [15].

Gaussian prior with noisy information: For inexact information, assume a random independent error ϵ with density $\rho_\epsilon(y)$. The information model is

$$y = Nf + \epsilon.$$

Then the MAPN estimate is

$$\widehat{f} = \arg\max_f \rho(f|y).$$

Note

$$\rho(f|y) = \frac{\rho_y(y|f)\rho(f)}{\rho_y(y)}.$$

If we further assume that the density in $\mathbb{R}^n$ of ϵ is Gaussian, i.e.,

$$\rho_\epsilon(\epsilon) = C_3 e^{-\|B\epsilon\|^2},$$

with B linear and C_3 a constant, we conclude

$$\begin{aligned} \rho(f|y) &= C_4 \frac{e^{-\|B(N(f)-y)\|^2} e^{-\|Af\|^2}}{\rho_y(y)} \\ &= C_5 e^{-\|B(N(f)-y)\|^2} e^{-\|Af\|^2}, \end{aligned}$$

where C_5 can depend on y. MAPN yields $\widehat{f}$ as a maximum of $e^{-\|B(N(f)-y)\|^2 - \|Af\|^2}$, so

$$\widehat{f} = \underset{Nf=y}{\arg\min} \ \|Af\|^2 + \|B(Nf-y)\|^2.$$

This log likelihood function can be minimized e.g., using Lagrange multipliers. Note this is exactly the spline solution of the same problem, and also the minimum of the regularization functional

$$\|Af\|^2 + \|B(Nf-y)\|^2$$

appearing in regularization methods for solving $Nf = y$ ([7], [17]).

Note also that the same procedure can be used for estimates based on elliptically contoured measures, which for elliptically contoured measures is more difficult to do using minimum square error methods. If likelihood functions are used, operator methods of maximizing them which work in finite dimension also typically work in infinite dimensions, with matrices replaced by operators (e.g., the covariance a Gaussian becomes an operator).

Example: In the example of the chemical mixture mentioned earlier, suppose that we have an input vector $\mathbf{x} = (x_1, x_2, ..., x_{20})$ representing an input mixture, with the 20 parameters representing temperature, mixing strength, along with 18 numbers representing proportions of chemical components (in moles/Kg). The measured output y represents the measured ratio of strength to brittleness for the plastic product of the mixture. In this example, we have assumed $n = 400$ runs of the experiment (a relatively small number given the size of the space F of possible i-o functions). For simplicity we assume that all variables x_i have been re-scaled so that their experimental range is the

interval $[-1/2, 1/2]$. Thus the a priori space F of possible i-o functions will be all (square integrable) functions on the input space $X = [-1/2, 1/2]^{20}$.

As a first approximation we assume that the target $f \in F$ is smooth, and thus we assume a prior probability distribution of f to be a Gaussian, favoring functions which are smooth and (for regularity purposes) not large. In addition, it may be felt (from previous experience) that the input variables $x_1, ..., x_5$ are associated with more sharp variations in y than the other 15 variables, and we expect the variation of f in these directions to be more sensitive to data than to our a priori assumptions of smoothness. Finally, we will want the a priori smoothness and size assumptions to have more weight in regions where there are less data in a precise parametric way. To this end our a priori desideratum will initially require "smallness" of the regularization term $\|f(\mathbf{x})\|_*^2 \equiv \|(d(\mathbf{x})\,[1 + (\mathbf{a} \cdot \mathbf{D})]\,f(\mathbf{x}))\|^2$, where $\mathbf{a} = (.1, .1, ..., .1, .2, .2, ..., .2) \in \mathbb{R}^{20}$ has its first 5 components equal to .10 (reflecting greater desired dependence of f on the variations in the first five variables) and its last 15 components equal to .20 (reflecting smaller dependence on the last 15 variables); these parameters can be adjusted. We have also defined $\mathbf{D} = \left(\frac{d^{15}}{dx_1^{15}}, \frac{d^{15}}{dx_2^{15}}, ..., \frac{d^{15}}{dx_{20}^{15}}\right)$; the high order of this operator is necessary since its domain must consist of continuous functions on $\mathbb{R}^{20}$, i.e., functions which have more than 10 derivatives. Note that the norm $\|\cdot\|$ above is given by $\|f\|^2 = \int_{\mathbb{R}^{20}} f(\mathbf{x})^2 d\mathbf{x}$, and the associated inner product is $\langle f, g\rangle = \int_{\mathbb{R}^{20}} f(x)g(x)dx$

Above the function $d(\mathbf{x})$ is chosen to reflect the density $\delta(\mathbf{x})$ of the sample points $\{\mathbf{x}_k\}_{k=1}^n$ at the location $\mathbf{x}$. Note if $\delta(\mathbf{x})$ is large, then we wish our estimate to depend more on data and less on a priori assumptions, and that we want $d(\mathbf{x})$ to be small there; the opposite is desired when $\delta(\mathbf{x})$ is small. As an initial approximation we choose $d(x) = (1 + \delta(\mathbf{x}))^{-1}$, with the local linear density $\delta(\mathbf{x})$ defined by $\delta(\mathbf{x})^m = \sum_{k=1}^n e^{-(\mathbf{x}-\mathbf{x}_k)}$, where here the dimension $m = 20$.

We will need to define a domain with proper boundary conditions for the square of the above regularization operator in order to obtain a function class F with a unique Gaussian distribution. The simplest boundary conditions are Dirichlet B.C., i.e., the requirement that f vanish on the boundary. Since this is not a natural requirement for our function class, we extend the domain space F to be square integrable functions on $[-1, 1]^{20}$ in order to impose such conditions sufficiently far away from the "region of interest" (the valid domain of variation of the inputs, $[-1/2, 1/2]^{20} \subset [-1, 1]^{20}$). Thus F is square integrable functions in 20 variables, each in the interval $[-1, 1]$. We define the operator $\mathbf{aD} = \left(a_1 \frac{d^{15}}{dx_1^{15}}, a_2 \frac{d^{15}}{dx_2^{15}}, ..., a_{20} \frac{d^{15}}{dx_{20}^{15}}\right)$. The Dirichlet operator B defined by $\langle f, Bf\rangle = \|(d(\mathbf{x})\,[(1 + (\mathbf{aD})]\,f(\mathbf{x}))\|^2$ is given by

$$\begin{aligned} Bf &= (d(\mathbf{x})\,[(1 + (\mathbf{aD})])\,(d(\mathbf{x})\,[(1 + (\mathbf{aD})])^*\,f \\ &= d(\mathbf{x})\,[(1 + (\mathbf{aD})]\,[(1 - (\mathbf{aD})]\,d(\mathbf{x})f(\mathbf{x}) \\ &= d(\mathbf{x})(1 - (\mathbf{aD})^2)d(\mathbf{x})f(\mathbf{x}) \end{aligned}$$

with domain of B restricted to f satisfying $f(x) = 0$ on the boundary, i.e., when any $x_i = \pm 1$. This operator has a trace class inverse $B^{-1} = C$, which is the covariance operator of a Gaussian measure P on F. We define P to be the associated Gaussian distribution on the space F of functions $f : [-1, 1]^{20} \to \mathbb{R}$. Note that our data are restricted to the subset $[-1/2, 1/2]^{20} \subset [-1, 1]^{20}$.

The measure P is our a priori Gaussian measure on F. By the above analysis, given an unknown $f \in F$ and information $\mathbf{y} = Nf = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))$, then according to the above, the MAPN estimator of f is

$$\begin{aligned} \widehat{f} &= \arg \inf_{Nf=\mathbf{y}} \{\|f(\mathbf{x})\|_*^2 \equiv \|(d(\mathbf{x})\,[1 + (\mathbf{aD})]\,f(\mathbf{x}))\|^2 \\ &= \langle f, Bf\rangle \\ &= \left\langle f, d^2(\mathbf{x})f(\mathbf{x}) - d(\mathbf{x})(\mathbf{aD})^2 d(\mathbf{x})f(\mathbf{x})\right\rangle\}. \end{aligned}$$

Note that ([18], [7], [12]), the above minimizer is

$$\widehat{f}(\mathbf{x}) = \sum_{k=1}^n c_k G(\mathbf{x}, \mathbf{x}_k), \tag{5}$$

where $G(\mathbf{x}, \mathbf{x}')$ is the Green function for the differential operator B defined above. Thus G is the kernel of the covariance operator $B^{-1} = C$, which can be computed separately. Note $B^{-1}f = d(\mathbf{x})^{-1}(1-(\mathbf{aD})^2)^{-1}d(\mathbf{x})^{-1}f$. Thus

$$G(\mathbf{x}, \mathbf{x}') = d(\mathbf{x})^{-1}G_0(\mathbf{x}, \mathbf{x}')d(\mathbf{x}')^{-1},$$

where $G_0(\mathbf{x})$ is the kernel of

$$(1 - (\mathbf{aD})^2)^{-1} = \left(1 - \sum_{i=1}^{20} a_i^2 \frac{d^{30}}{dx_i^{30}}\right)^{-1} \tag{6}$$

on F, with vanishing boundary conditions. Since the boundaries of the domain $[-1, 1]^{20}$ of F are relatively far from the region of interest $[-.5, .5]^{20}$ in which the data lie, we approximate the Green kernel $G_0(x, x')$ with boundary conditions vanishing on the boundary of $[-1, 1]^{20}$, by the approximation $\widetilde{G}(x, x')$, the kernel of $(1 - (\mathbf{aD})^2)^{-1}$ on $\mathbb{R}^{20}$ with boundary conditions at ∞. This has the form

$$\widetilde{G}(\mathbf{x}, \mathbf{x}') = \widetilde{G}(\mathbf{x} - \mathbf{x}')$$

$$= \mathcal{F}\left(\left(1 - \sum_{i=1}^{20} a_i^2\,(i\xi_i)^{30}\right)^{-1}\right)(\mathbf{x} - \mathbf{x}'), \tag{7}$$

where $\mathcal{F}$ denotes Fourier transform and ξ_i is the variable dual to x_i. Thus our approximation of f is given by (5) (see, e.g., [12]), with $c = \{c_i\}_{i=1}^n = G^{-1}(Nf)^{\mathrm{T}}$, where

G is a matrix with $G_{ij} = G(x_i - x_j)$, with T denoting transpose. Explicitly,

$$\widehat{f} = \sum_{k=1}^{n} c_k G(\mathbf{x} - \mathbf{x}_k) \approx \sum_{k=1}^{n} \left(\mathbf{G}^{-1} N f^{\mathrm{T}}\right)_k \widetilde{G}(\mathbf{x} - \mathbf{x}_k),$$

where $\widetilde{G}(\mathbf{x} - \mathbf{x}_k)$ is computable from (7).

6 Conclusions

We have showed that it is possible to extend MAP estimates to high and infinite dimensional approximations. In applications we have constructed an approximation to an unknown function f from pointwise information $Nf = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))$ by incorporating prior information regarding smoothness of f , and adapting the solution in a way which depends on the local density δ of data points $\mathbf{x}_i$.

References

[1] T. Mitchell, *Machine Learning* (New York: McGraw-Hill, 1997).

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

[3] J. Traub, G. Wasilkowski, and H. Woniakowski, *Information-Based Complexity* (Boston: Academic Press, 1988).

[4] T. Poggio and S. Smale, The Mathematics of Learning: Dealing with Data, *Notices of the AMS,* 50, 2003, 537-544.

[5] T. Poggio and C. Shelton, Machine Learning, Machine Vision, and the Brain, *The AI Magazine*, 20, 1999, 37-55. URL: citeseer.nj.nec.com/poggio99machine.htm

[6] V. Vapnik, *The Nature of Statistical Learning Theory*, (New York: Springer, 1995).

[7] V. Vapnik, *Statistical Learning Theory* (New York: Wiley, 1998).

[8] G. Wahba, Generalization and regularization in nonlinear learning systems, in M. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*, (Cambridge: MIT Press, 1995) 426-430.

[9] G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, in B. Schoelkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods Support Vector Learning*, (Cambridge: MIT Press, 1999) 69-88.

[10] C. Micchelli and T. Rivlin, *Lectures on optimal recovery*, Lecture Notes in Mathematics 1129, (Berlin: Springer-Verlag, 1985) 21-93.

[11] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference and prediction* (Berlin: Springer-Verlag, 2001).

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

[13] J. Traub and H. Wozniakowski, Breaking intractability, *Scientific American,* 270, 1994, 102-107.

[14] M. Kon, Density functions for machine learning and optimal recovery, 2004, preprint..

[15] G. Wahba, Bayesian confidence intervals for the cross-validated smoothing spline, *Journal of the Royal Statistical Society B,* 45, 1983, 133-150.

[16] N. Friedman and J. Halpern, Plausibility measures and default reasoning, In *Thirteenth National Conf. on Artificial Intelligence* (New York: AAAI, 1996), 1297-1304.

[17] T. Evgeniou, M. Pontil, and T. Poggio, Regularization Networks and Support Vector Machines, *Advances in Computational Mathematics,* 13, 2000, 1-50.

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

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