Additive models in high dimensions

Markus Hegland* Vladimir Pestov†

(Received 12 November 2004, revised 31 October 2005)

Abstract

Additive decompositions are established tools in nonparametric statistics and effectively address the curse of dimensionality. For the analysis of the approximation properties of additive decompositions, we introduce a novel framework which includes the number of variables as an ingredient in the definition of the smoothness of the underlying functions. This approach is motivated by the effect of concentration of measure in high dimensional spaces. Using the resulting smoothness conditions, convergence of the additive decompositions is established. Several examples confirm the error rates predicted by our error bounds. Explicit expressions for optimal additive decompositions (in an L_2 sense) are given which can be seen as a generalisation of multivariate Taylor polynomials where the monomials are replaced by higher order interactions. The results can be applied to the numerical approximation of functions with hundreds of variables.

*CMA, Mathematical Sciences Institute, Australian National University, Canberra, Australia. mailto:markus.hegland@anu.edu.au

†Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada. mailto:vpest283@uottawa.ca

See http://anziamj.austms.org.au/V46/CTAC2004/Hegl for this article, Published November 14, 2005. ISSN 1446-8735

Contents

1 Introduction

Additive approximations of real functions f defined on a domain $\Omega \subset \mathbb{R}^n$ have the form

$$\begin{aligned} f_{\text{add}}(x) \;=\; & f_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + \\ & f_{1,2}(x_1, x_2) + \cdots + f_{i_1,i_2}(x_{i_1}, x_{i_2}) + \cdots f_{n-1,n}(x_{n-1}, x_n) + \cdots + \\ & f_{i_1,i_2,\ldots,i_m}(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) + \cdots + f_{n-m+1,\ldots,n}(x_{n-m+1}, \ldots, x_n)\,. \end{aligned} \tag{1}$$

The terms of this decomposition, called *interactions*, are functions of at most m variables and the effectiveness of the approximation in dealing with the curse of dimensionality [1] is due to the choice $m \ll n$. We assume that Ω is endowed with a natural probability measure and we consider the approximation which is L_2 optimal with respect to the product measure defined by the marginal measures. This approximation is optimal for product measures (or independent random variables) and order optimality is shown for a slightly larger class of measures.

Additive approximations are widely used in statistics [2, 5, 13, 3], data mining [6] and in the theory of numerical quadrature [9, 7]. They are often

referred to as ANOVA decompositions as they generalise for real variables the methods which are used in the analysis of variance (ANOVA). It has been observed that often a choice of $m = 2, \ldots, 5$ appears to give reasonable approximations in practice. As the amount of data required to estimate interactions grows exponentially in m, this observation may reflect the fact that most currently available data sets are simply not large enough to identify higher order interactions and, consequently, only functions which are well approximated can be fitted from data. Here we show that smooth functions are indeed well approximated by ANOVA decompositions, and, consequently, can be fitted with the limited amount of data available.

In one of its forms, the phenomenon of concentration of measure [4, 8, 12] says that every Lipschitz function on a sufficiently high-dimensional domain is well-approximated by a *constant function*, that is, an additive function of the lowest possible order of interaction $m = 0$. However, as one would expect, a reasonably good approximation requires the intrinsic dimension of a dataset to be prohibitively high. For general functions one cannot expect an additive approximation to be better than the constant as will be seen later. Better approximations are obtained when a smoothness condition is invoked which holds uniformly over the dimension n. Our suggestion is to consider smooth functions and generalise the standard Lipschitz condition by requiring the L_2-norm of the vector of all mixed derivatives of order $k \leq m$ to be bounded above by a constant L_m, independent of the dimension of the domain Ω. In this case the method developed here is effective and the derived error rates are confirmed for a couple of examples.

In Section 2 we review the relevant concentration properties and introduce our smoothness assumptions. Section 3 provides the main approximations and error bounds and in Section 4 we discuss a couple of examples.

2 Concentration and continuity

The simplest class of additive functions is that of zeroth order of interaction, $m = 0$, in which case the approximating functions f_{add} are simply constants. It turns out that even the approximation by constants admits a substantial theory if the domain is high-dimensional. Such approximation improves as dimension grows, which observation is at the core of the *phenomenon of concentration of measure on high-dimensional structures.* The range of various manifestations of this phenomenon in mathematical sciences is extremely wide and includes results as diverse as the law of large numbers, blowing-up lemma in information theory, Dvoretzky's theorem on almost spherical sections of convex bodies, foundations of statistical physics, and so forth. (See [4, 8, 12, 10] and numerous references therein.)

The concentration phenomenon refers to the observation that for many 'natural' families of spaces, indexed by their dimension n, and endowed with a metric and a probability measure, the probability that a Lipschitz function $f(x)$ (defined on these spaces) differs from its expected value by less than $\varepsilon > 0$ is at least

$$1 - C_1 \exp(-C_2 \varepsilon^2 n)\,, \tag{2}$$

where $C_1, C_2 > 0$ are constants only depending on the family of spaces in question and the Lipschitz constant. Intuitively, it means that a 'nice' function on a space of high dimension 'concentrates' near one value. Such estimates with varying constants hold for the hypercubes (remember that the distance has to be appropriately normalised), the Euclidean spaces with the Gaussian measure, the Hamming cubes, the groups of unitary matrices.

A careful examination reveals that the cases where Lipschitz continuous functions are approximated well by constants occur for extremely high n way beyond $n \approx 100$, the case we are interested in. The next natural question is therefore: will the approximation error bounds improve automatically if one allows the approximation by additive functions of *higher interaction order than zero*?

It seems quite natural that by significantly relaxing the restrictions on the class of approximating functions one gets better approximation bounds. Rather surprisingly, it is not the case, as there exist functions on n-dimensional domains for which approximation by constants is the best possible among *all* additive functions with interaction orders k of up to $n-1$. A simple example of such a function is $f(\mathbf{x}) = x_1 x_2 \cdots x_n$ on the domain $[-1,1]^n$ with a uniform distribution.

In view of this, it seems unavoidable that one should impose additional restrictions on the functions f to obtain better bounds on higher-order approximations with additive functions. We will now put forward such restrictions as we find most natural.

In practical applications, the variables x_i denote features of an underlying object which may be a company, a biological cell or a river catchment. By increasing the number n of features one attempts to better characterise the objects. However, an increased number of features will typically lead to an increase of the (Euclidean) distance between the feature vectors. In order for the distance to sensibly model a distance between two objects one needs to scale the Euclidean distance as

$$d(\mathbf{x},\mathbf{y}) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}\,,$$

such that it no longer (on average) depends on the dimension for the case of independent uniformly distributed features $x_i \in [0,1]$. We will consider functions f which satisfy a Lipschitz condition $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\, d(\mathbf{x},\mathbf{y})$ with respect to this normalised norm. A simple example of such a function (where $L = 1$) is $f(\mathbf{x}) = \sum_{i=1}^{n} x_i/n$. For differentiable functions the Lipschitz condition with the scaled norm is equivalent to

$$\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \leq \frac{L^2}{n}\,. \tag{3}$$

Considering mth order differences and observing that they should be of order $\mathcal{O}(h^m)$ where $d(\mathbf{x},\mathbf{y}) \leq h$ for any two points $\mathbf{x}$ and $\mathbf{y}$ occurring in the difference one gets the smoothness condition:

$$\sum_{1\leq i_i<\cdots<i_m\leq n} \left(\frac{\partial^m f(\mathbf{x})}{\partial x_{i_1}\cdots\partial x_{i_m}}\right)^2 \leq \frac{L_m^2}{n^m}\,.$$

That this is indeed a smoothness condition which holds for practically important functions is supported by the examples $f(\mathbf{x}) = \exp\left(-\sum_{i=1}^n x_i^2/n\right)$ and $f(\mathbf{x}) = \sum_{i,j=1}^n q_{ij}\phi(x_i,x_j)$ with $\sum_{i,j=1}^n |q_{ij}| = 1$ (for example, the energy of n interacting particles with a fixed total mass or charge). These conditions given here assume that all the features are equally important or informative. An alternative smoothness assumption, where features are assigned weights of importance is given in [11]. Here we mainly consider the example of $\Omega = [0,1]^n$ but in a similar fashion, one can treat other important cases, such as the sphere and the normal distribution. In the first case no normalisation is required and the bounds on the derivatives are thus slightly different; however, the approximation results are basically the same and reflect the concentration property of high dimensional domains.

3 Additive approximation

Consider first the case of a probability distribution $p(x) = \prod_{i=1}^n p_i(x_i)$ and denote by E the expectation and by $E(f \mid x_{i_1},\ldots,x_{i_k})$ the usual conditional expectations. Using the operators D_i defined by

$$(D_i f)(x) = f(x) - E(f \mid x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)\,,$$

one obtains the "telescoping sum" using the independence assumption:

$$f(x) = E(f) + \sum_{i=1}^n D_i E(f \mid x_1,\ldots,x_i)\,. \tag{4}$$

Iterating this expansion for each term one gets the following decomposition.

Theorem 1 *For $f \in L_2$ and $1 \le m \le n$ one has*

$$f(x) = E(f) + \sum_{i=1}^{n} D_i E(f \mid x_i) + \sum_{1 \le i_2 < i_1 \le n} D_{i_2} D_{i_1} E(f \mid x_{i_2}, x_{i_1}) + \cdots$$
$$+ \sum_{1 \le i_{m-1} < \cdots < i_1 \le n} D_{i_{m-1}} \cdots D_{i_1} E(f \mid x_{i_{m-1}}, \ldots, x_{i_1})$$
$$+ \sum_{1 \le i_m < \cdots < i_1 \le n} D_{i_m} \cdots D_{i_1} E(f \mid x_1, x_2, \ldots, x_{i_m}, x_{i_{m-1}}, x_{i_{m-2}}, \ldots, x_{i_1}) .$$

Proof: One uses induction over m. The case $m = 1$ is just equation (4). One obtains the decomposition for the case of $m = k$ from the case $m = k-1$ by expanding the last term using the telescoping expansion again. ♠

A similar decomposition for the special case of $m = n$ has been proved in [2] where the theorem is called *Decomposition Lemma*.

Next we introduce the space of L_2 functions which are sums of functions each depending on k variables only, as:

$$L_{2,k} := \left\{ g(x) = \sum_{i_1,\ldots,i_k} g_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k}) \in L_2 \right\} .$$

(Note that $L_{2,k}$ is closed, which follows from Theorem 2 below.) Introduce the operator $P_m : L_2 \to L_{2,m}$ such that

$$(P_m f)(x) = E(f) + \sum_{i=1}^{n} D_i E(f \mid x_i) + \sum_{1 \le i_2 < i_1 \le n} D_{i_2} D_{i_1} E(f \mid x_{i_2}, x_{i_1}) + \cdots$$
$$+ \sum_{1 \le i_m < \cdots < i_1 \le n} D_{i_m} \cdots D_{i_1} E(f \mid x_{i_m}, \ldots, x_{i_1})$$

and the (remainder) operator $R_m : L_2 \to L_2$ with

$$(R_m f)(x) = \sum_{1 \le i_m < \cdots < i_1 \le n} D_{i_m} \cdots D_{i_1} E(f \mid x_1, \ldots, x_{i_m}, \ldots, x_{i_1}) \,.$$

From Theorem 1 one then gets $f = P_m f + R_{m+1} f$ and one shows that this is an orthogonal decomposition:

Theorem 2 *The operator P_m is an orthogonal projection, and*

$$E((f - P_m f)^2) \le E((f - g)^2) \,, \quad \textit{for all } g \in L_{2,m} \textit{ and } f \in L_2 \,.$$

Proof: Using the definition of D_i and the independence assumption one can see that $P_m f$ can be recast as a linear combination of terms of the form $E(f \mid x_{j_1}, \ldots, x_{j_t})$ for appropriately chosen $x_{j_1}, \ldots, x_{j_t}$. As for every square integrable g one has

$$\begin{aligned} &\int E(f \mid x_{j_1}, \ldots, x_{j_t}) g(\mathbf{x}) p(\mathbf{x}) \, d\mathbf{x} \\ &= \int E(f \mid x_{j_1}, \ldots, x_{j_t}) E(g \mid x_{j_1}, \ldots, x_{j_t}) p(\mathbf{x}) \, d\mathbf{x} \,, \end{aligned}$$

it follows that all operators producing the terms of the decomposition are self-adjoint and so P_m and R_{m+1} are self-adjoint as well.

Now consider any $f \in L_{2,m}$. As f only depends on m variables, one has $D_{i_1} \cdots D_{i_{m+1}} f = 0$ and so $R_{m+1} L_{2,m} = 0$. From this it follows that $P_m^2 = P_m$ and, as P_m is self-adjoint, it is an orthogonal projection onto $L_{2,m}$. ♠

In the sequel we will use the (marginal) cumulative distribution functions $P_i(x_i) = \int_{-\infty}^{x} p_i(s) \, ds$. Furthermore, let

$$G_i(t_1, t_2) := \min_{a,b=1,2} (P_i(t_a)(1 - P_i(t_b))$$

$$\text{and} \qquad \gamma := \int \max_i G_i(t,s)\, dt\, ds\,.$$

Integration by parts and the Cauchy–Schwarz inequality yield

$$\sum_{i=1}^{n} E((D_i f(x))^2) \le \gamma L^2$$

where L is the Lipschitz constant. For the error we will use the seminorm

$$|f|_m^2 := \sup_x \sum_{1\le i_1<\cdots<i_m\le n} \left(\frac{\partial^m f(x)}{\partial x_{i_1}\cdots\partial x_{i_m}}\right)^2 .$$

The approximation result for the optimal additive approximants is then

Theorem 3 *For any f with bounded seminorm $|f|_m$ the mean squared error of $P_m f$ is bounded by*

$$E((R_m f)^2) \le \gamma^m |f|_m^2\,. \tag{5}$$

Proof: One requires the kernel $k_i(x_i,t_i) = P_i(t_i) - H(t_i - x_i)$ where $H(x)$ is the Heaviside function, that is, $H(x) = 1$ for $x \ge 0$ and $H(x) = 0$ for $x < 0$. Using integration by parts one gets for differentiable f:

$$D_i f(x) = \int_{-\infty}^{\infty} k_i(x_i,t_i) \frac{\partial f}{\partial t_i}(x_1,\ldots,x_{i-1},t_i,x_{i+1},\ldots,x_n)\, dt_i\,. \tag{6}$$

Use $g_{i_1,\ldots,i_m}(t_{i_1},\ldots,t_{i_m}) := \partial^m E(f \mid x_1,\ldots,x_{i_m-1},t_{i_m},\ldots,t_{i_1})/\partial t_{i_1}\cdots\partial t_{i_m}$ and the fact that all the terms in the decomposition of $R_m f$ are orthogonal to get

$$E((R_m f)^2) = \sum_{i_1<\cdots<i_m} \int g_{i_1,\ldots,i_m}(t_{i_1},\ldots,t_{i_m}) g_{i_1,\ldots,i_m}(s_{i_1},\ldots,s_{i_m}) \times \prod_{j=1}^{m} G_{i_j}(s_{i_j},t_{i_j})\, ds\, dt\,.$$

With

$$\gamma_m := \int \max_{i_1<\cdots<i_m} \prod_{j=1}^{m} G_{i_j}(s_{i_j}, t_{i_j})\, ds\, dt$$

one gets $E((R_m f)^2) \le |f|_m^2\, \gamma_m$ and the bound follows from $\gamma_m \le \gamma^m$. ♠

If, as we suggested in the previous section, one has $|f|_m \le L_m/n^{m/2}$, one gets the error bound

$$E((R_m f)^2) \le \frac{\gamma^m L_m^2}{n^m}.$$

In particular, for the uniform distribution on $[-1,1]^n$ one has $\gamma = 1/3$ and so $E((R_m f)^2) \le L_m^2/(3^m n^m)$. In the case of the standard distribution one gets $\gamma \approx 0.516$ and so $E((R_m f)^2) \le 0.516^m L_m^2/n^m$.

So far, the given results concerned product measures only. Consider now an arbitrary probability distribution function $p(\mathbf{x})$ and let $p_i(x)$ be the marginal distributions of $p(\mathbf{x})$. Now let $P_m^{\otimes}$ denote the additive approximation P_m with respect to the measure defined by the product of the marginal distributions and let $R_{m+1}^{\otimes}$ be the corresponding error. In general, this approximation is not going to be close to optimal, in fact, one cannot expect to get a reasonable approximation in this way. (Consider, for example, the case where two variables are identical.)

The situation improves when the original measure and the product of the marginal measures are absolutely continuous with respect to each other, that is, when the measures have the same null sets. In this case the Radon–Nikodym theorem implies that there is a measurable function (the Radon–Nikodym derivative) $\psi(\mathbf{x})$ such that $p(\mathbf{x}) = \psi(\mathbf{x}) \prod p_i(x_i)$. If $E^{\otimes}$ is the expectation with respect to the product measure one has for any positive random variable Y the bound $E(Y) = E^{\otimes}(Y/\psi) \le E^{\otimes}(Y)/\operatorname{essinf}_{\mathbf{x}} \psi(\mathbf{x})$. Now one combines this with the approximation theorem for P_m to get a mean squared error bound for the approximation $P_m^{\otimes}$.

Theorem 4 *Let ψ be the Radon–Nikodym derivative of a distribution $p(\mathbf{x})$*

with respect to the distribution defined by the product of its marginals. Then the error of the optimal approximation with respect to the marginals

$$E((R_m^{\otimes} f)^2) \leq \gamma^m |f|_m^2 / \operatorname{essinf}_{\mathbf{x}} \psi(\mathbf{x}) \,. \tag{7}$$

By similar reasoning one can see that in this case the approximation rate of $P_m^{\otimes}$ is the best possible for additive approximations as:

Theorem 5

$$\min_{g \in L_{2,m}} E((f-g)^2) \leq E((R_m^{\otimes} f)^2) \leq \kappa \min_{g \in L_{2,m}} E((f-g)^2) \tag{8}$$

where $\kappa = \sup_{\mathbf{x}} \psi(\mathbf{x}) / \operatorname{essinf}_{\mathbf{x}} \psi(\mathbf{x})$.

A corollary of these two theorems is that the best additive approximation of order m does have the same approximation rate $\mathcal{O}(|f|_m)$ as the approximation with respect to the corresponding product measure if the two measures are equivalent.

4 Examples and summary

A first example is $f(x) = \exp(-\|x\|^2/n)$ and a uniform distribution in $[-1, 1]$. In this case the sum of the squared mth derivatives is bounded by $4^m/(n^m m!)$. The error bound from the previous section thus provides an expected squared error of $E(R_m f^2) \leq 4^m/(3^m m! n^m)$. The additive approximation can be determined explicitly using the formulas from the previous section and the error was estimated using a sample 1000 uniformly distributed points. The result is displayed in Figure 1) and while the actual bound is a bit pessimistic, the measured errors do display the predicted $n^{-m/2}$ behaviour as a function of n.

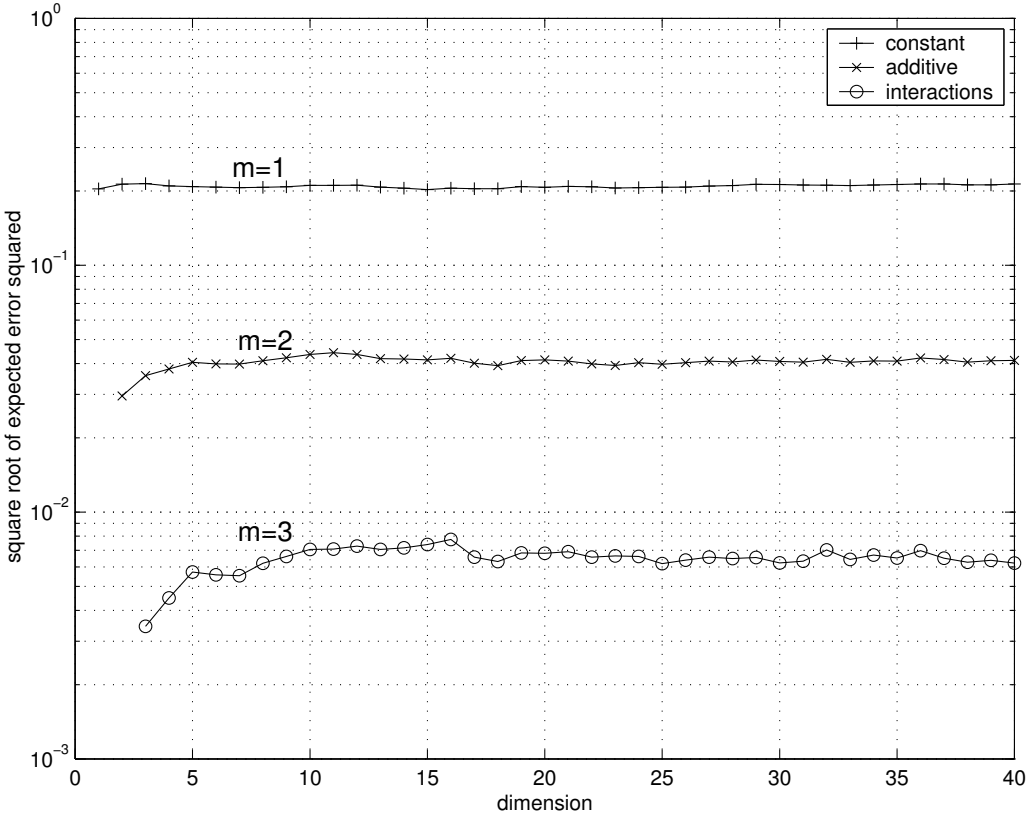

FIGURE 1: $n^{m/2}$ times RMS (root mean squared) errors for a constant, first and second order approximation for the function $f(x) = \exp(-\|x\|^2/n)$.

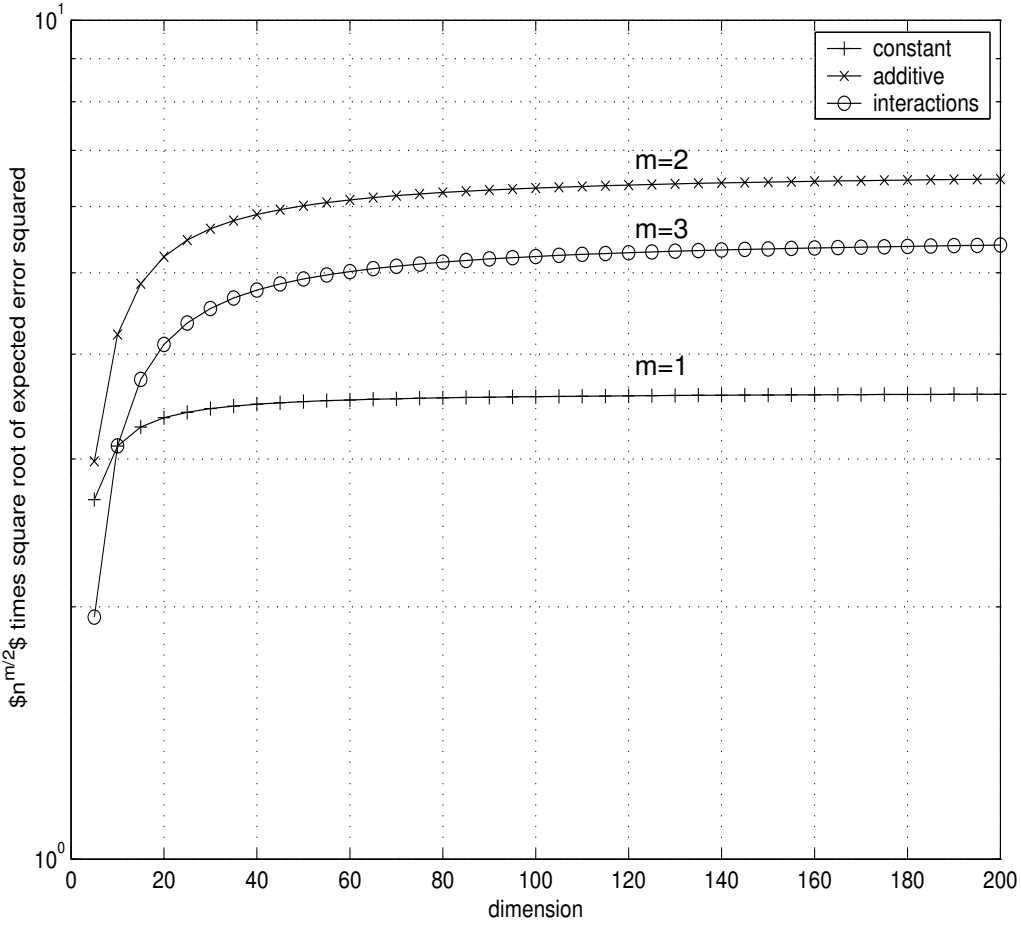

FIGURE 2: $n^{m/2}$ times RMS (root mean squared) errors for a constant, first and second order approximation for the function $f(x) = (1 + \sin(x_1))(1 + \sin(x_2))(1 + \sin(x_3))$.

A different case was considered in Figure 2). Here the chosen function was $f(x) = (1+\sin(x_1))(1+\sin(x_2))(1+\sin(x_3))$ and the data was normally distributed, so that the sum of the variances was independent of the dimension. One then gets the same error behaviour as for the first example which is confirmed by experiment.

Finally, we consider the MARS approximation using the algorithm proposed in [3] and 1000 random data points producing the errors displayed in Figure 3). The code allows the choice of the maximal order m of the interactions and one can see how by choosing higher m one gets better approximation. We conjecture that the drop in precision (especially for $m = 4$) relates to the fact that not enough data was available to provide a good estimate for the interaction terms.

In summary, a constructive approximation formula for best L_2 approximations is provided for product measures. It was shown that these approximation have error rates bounded by $\mathcal{O}(|f|_m)$ and the rates were confirmed in some experiments. These approximations converge with dimension for an appropriate scaling of the derivatives with dimension which is motivated by the scaling of the norm. In the case of general measures the method can be used to compute approximations with respect to the induced product measure defined by the product of the marginal measures. One gets order optimal approximations (in terms of the dimension) when the induced and the original measures are absolutely continuous with respect to each other.

Acknowledgments: At the initial stage research was supported by the Australian Cooperative Research Centre for Advanced Computational Systems (ACSys). Partial support was also provided by the Marsden Fund of the Royal Society of New Zealand, in particular towards a visit by the first named author (M. H.) to the Victoria University of Wellington in April 2002.

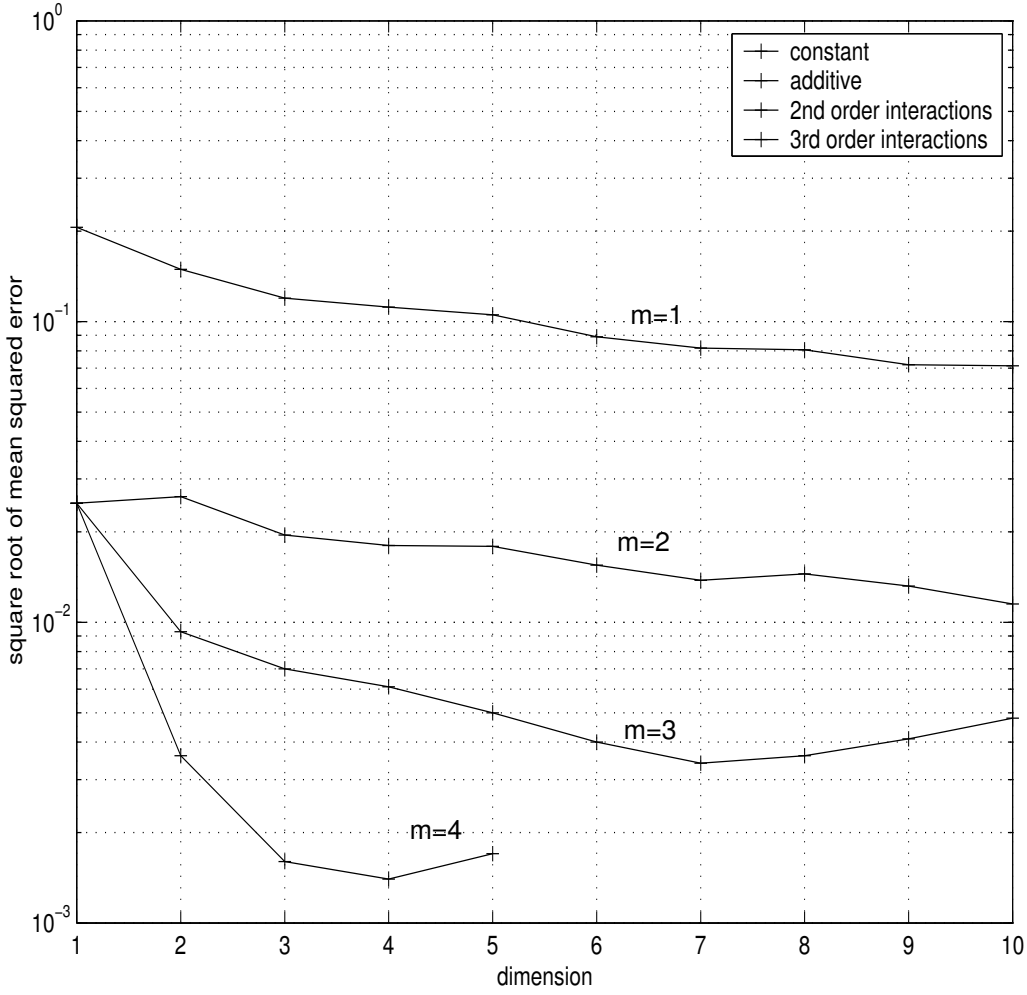

FIGURE 3: RMS errors for the MARS fit.

References

[1] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, 1961. C1206

[2] B. Efron and C. Stein, The jackknife estimate of variance, *Annals of Statistics* **9** (3), 1981, 586–596.
http://www.jstor.org/view/00905364/di983909/98p0170d/0
C1206, C1211

[3] J. H. Friedman, Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19** 1991, 1–141.
http://www.jstor.org/view/00905364/di983949/98p0128u/0
C1206, C1218

[4] M. Gromov, Metric Structures for Riemannian and Non-Riemannian Spaces, *Progress in Mathematics* **152**, Birkhäuser Verlag, 1999.
C1207, C1208

[5] T. J. Hastie and R. J. Tibshirani, Generalized Additive Models, *Monographs on Statistics and Applied Probability* **43**, Chapman & Hall, London a.o., 1990. C1206

[6] M. Hegland, Computational challenges in data mining, in: *Proceedings of the Computational Techniques and Applications Conference CTAC'99, ANZIAM Journal* **42** (2000–2001), Part C, pp. C1–C43.
http://anziamj.austms.org.au/V42/CTAC99/AHeg C1206

[7] F. J. Hickernell, Quadrature Error Bounds with Applications to Lattice Rules, *SIAM J. Num. Anal.* **33** (5) (1996), 1995–2016. C1206

[8] V. D. Milman, The heritage of P. Lévy in geometric functional analysis, *Astérisque* **157–158** (1988), 273–301. C1207, C1208

[9] A. B. Owen, Orthogonal Arrays for Computer Experiments, Integration and Visualisation, *Statistica Sinica* **2** (1992), 439–452. C1206

[10] V. Pestov, On the geometry of similarity search: dimensionality curse and concentration of measure, *Inform. Proc. Letters* **73** (2000), 47–51. C1208

[11] I. H. Sloan and H. Woźniakowski, When are Quasi-Monte Carlo algorithms efficient for high dimensional integrals, *J. Complexity* **14**, (1998), 1–33. C1210

[12] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, *Publ. Math. IHES* **81** (1995), 73–205. C1207, C1208

[13] G. Wahba, Spline models for observational data, *CBMS-NSF Reg. Conf. Ser. Appl. Math.* **59**, SIAM, 1990. C1206