Fitting multidimensional data using gradient penalties and the sparse grid combination technique

Jochen Garcke and Markus Hegland

October 28, 2008

Abstract

Sparse grids, combined with gradient penalties provide an attractive tool for regularised least squares fitting. It has earlier been found that the combination technique, which builds a sparse grid function using a linear combination of approximations on partial grids, is here not as effective as it is in the case of elliptic partial differential equations. We argue that this is due to the irregular and random data distribution, as well as the proportion of the number of data to the grid resolution. These effects are investigated both in theory and experiments. As part of this investigation we also show how overfitting arises when the mesh size goes to zero. We conclude with a study of modified "optimal" combination coefficients who prevent the amplification of the sampling noise present while using the original combination coefficients.

AMS subject classification: 62G08, 65D10, 62J02

Key words: sparse grids, combination technique, regression, high-dimensional data, regularisation

1 Introduction

In this paper we study the convergence behaviour of a grid based approach for the regression problem arising in machine learning. A set of (noisy) data points $\underline{x}_i$ in a d-dimensional feature space is given, together with associated values y_i. Often one assumes that a function f_* describes the relation between the predictor variables $\underline{x}$ and the response variable y. One now wants to (approximately) reconstruct the function f_* from the given data. This allows the prediction of the function value for any newly given data point for future decision-making. The dimension d of the attribute domain can be high and m, the number of data points, can be large, both pose a severe computational challenge.

We investigate a discretisation approach introduced by Garcke et.al. [9] for the regularised least squares ansatz [18]. It uses a gradient based regularisation and discretises the minimisation problem with an independent grid with associated local basis functions. The approach is similar to the numerical treatment of partial differential equations by finite element methods. This way the data information is transferred into the discrete function space defined by the grid and its corresponding basis functions.

To cope with the complexity of grid-based discretisation methods in higher dimensions the sparse grid combination technique [11] is applied to the regression problem [9]. Here, the regularised least squares ansatz is discretised and solved on a certain sequence of anisotropic grids, i.e., grids with different mesh sizes in each coordinate direction. The sparse grid solution is then obtained from the (partial) solutions on these different grids by their linear combination using combination coefficients which depend on the involved grids. The approach was shown to successfully treat machine learning problems in an intermediate dimensional regime, $d \lesssim 20$, with only a linear scaling in the number of data and therefore suitable for a large number of data [8, 9].

Following empirical results in [5], which show instabilities of the combination technique in certain situations, we investigate in this article the convergence behaviour of full and sparse grid discretisation of the regularised regression problem. There are two approaches to analyse the convergence behaviour of the combination technique. Originally, extrapolation arguments were used where a certain error expansion for the partial solutions is assumed [4, 11]. Alternatively, one can view the combination technique as an approximation of a projection into the underlying sparse grid space, which is exact only if the partial projections commute [14]. Both these assumptions do not hold for the regularised regression problem; the combination technique can actually diverge. While we investigated the projection view in detail in Hegland et. al. [14], we consider here the extrapolation view.

Of particular interest is the effect of the gradient based regularisation in higher dimensions since it does not enforce continuous functions as it only gives H^1-functions. Its systematic theoretical investigation is the main contribution of this paper.

Already in two dimensions empirical investigations show divergence of the combination technique depending on the number of data and the regularisation parameter. This is partly due to the the finiteness of the random sample. This effect vanishes asymptotically as the number of data points goes to infinity. Therefore the assumed error expansion needed for the extrapolation arguments for the convergence of the combination technique does not hold for a fixed number of data. Furthermore, after deriving a discrete version of the Sobolev inequality we formulate the overfitting effect in the data points of grid based approaches in higher dimensions and finish our theoretical investigations by showing how the solution between the data points converges to a constant.

It was seen that applying the optimised combination technique [13, 14] repairs the instabilities of the combination technique to a large extent. The combination coefficients now not only depend on the grids involved, but on the function to be reconstructed as well, resulting in a non-linear approximation approach. We show in numerical experiments the stable behaviour of this numerical scheme. Note that recent empirical studies show competetive results of this approach for a number of standard regression benchmarks in machine learning [6].

2 Regularised sparse grid regression and the combination technique

In this section we introduce the regularised least squares problem in a finite dimensional space V, establish the corresponding linear systems of equations and describe the sparse grid combination technique.

2.1 Regularised least squares regression

We consider the regression problem in a possibly high-dimensional space. Given is a data set

$$S = \{(\underline{x}_i, y_i)\}_{i=1}^m \quad \underline{x}_i \in \mathbb{R}^d,\ y_i \in \mathbb{R},$$

where we denote by $\underline{x}$ a d-dimensional vector or index with entries $x_1, \ldots, x_d$. We assume that the data has been obtained by sampling an unknown function f_* which belongs to some space V of functions defined over $\mathbb{R}^d$. The aim is to recover the function f_* from the given data as well as possible. To achieve a well-posed (and uniquely solvable) problem Tikhonov-regularisation theory [17, 18] imposes a smoothness constraint on the solution. We employ the gradient as a regularisation operator which leads to the variational problem

$$f_V = \operatorname*{argmin}_{f \in V} R(f)$$

with

$$R(f) = \frac{1}{m} \sum_{i=1}^{m} (f(\underline{x}_i) - y_i)^2 + \lambda ||\nabla f||^2, \tag{1}$$

where $y_i = f_*(\underline{x}_i)$. The first term in (1) measures the error and therefore enforces closeness of f to the labelled data, the second term $||\nabla f||^2$ enforces smoothness of f, and the regularisation parameter λ balances these two terms.

Let us define the following semi-definite bi-linear form

$$\langle f, g \rangle_{RLS} = \frac{1}{m} \sum_{i=1}^{m} f(\underline{x}_i) g(\underline{x}_i) + \lambda \langle \nabla f, \nabla g \rangle \tag{2}$$

and choose V so that $\langle \cdot, \cdot \rangle_{RLS}$ is a scalar product on it. With respect to this scalar product the minimisation (1) is an orthogonal projection of f_* into V [13], i.e., if $\|f - f_*\|_{RLS}^2 \leq \|g - f_*\|_{RLS}^2$ then $R(f) \leq R(g)$. As the point evaluations $f \mapsto f(\underline{x})$ are not continuous in the Sobolev space H^1 for $d \geq 2$ we do not get a H^1-elliptic problem. For the space V we thus suggest to choose a finite dimensional subspace $V \subset H^1$ of continuous functions containing the constant function.

In the following we restrict the problem explicitly to a finite dimensional subspace $V_N \subset V$ with an appropriate basis $\{\varphi_j\}_{j=1}^N$. A function $f \in V$ is then approximated by

$$f_N(\underline{x}) = \sum_{j=1}^{N} \alpha_j \varphi_j(\underline{x}). \tag{3}$$

We now plug this representation of a function $f \in V_N$ into (1) and obtain the linear system of equations

$$(\mathcal{B}^\top \mathcal{B} + \lambda m \cdot \mathcal{C})\, \underline{\alpha} = \mathcal{B}^\top \underline{y} \tag{4}$$

and therefore are able to compute the unknown vector $\underline{\alpha}$ for the solution f_N of (1) in V_N. $\mathcal{C}$ is a symmetric $N \times N$ matrix with entries $\mathcal{C}_{j,k} = \langle \nabla\varphi_j, \nabla\varphi_k \rangle$, and corresponds to the smoothness penalty. $\mathcal{B}^\top$ is a rectangular $m \times N$ matrix with entries $(\mathcal{B}^\top)_{j,k} = \varphi_j(\underline{x}_k)$ and transfers the information from the data into the discrete space, $\mathcal{B}$ correspondingly works in the opposite direction. The vector $\underline{y}$ contains the data labels y_i and has length m.

In particular we now employ a finite element approach, using the general form of anisotropic mesh sizes $h_t = 2^{-l_t}, t = 1, \ldots, d$ and number the grid points using the multi-index $\underline{j}, j_t = 0, \ldots, 2^{l_t}$. We use piecewise d-linear functions

$$\phi_{\underline{l},\underline{j}}(\underline{x}) \;:=\; \prod_{t=1}^{d} \phi_{l_t,j_t}(x_t), \quad j_t = 0, \ldots, 2^{l_t}$$

where the one-dimensional basis functions $\phi_{l,j}(x)$ are the so-called hat functions. We denote with V_n the finite element space which has the mesh size h_n in each direction.

2.2 Combination technique

The sparse grid combination technique [11] is an approach to approximate functions defined over higher dimensional spaces. Following this ansatz we discretise and solve the problem (1) on a sequence of small anisotropic grids $\Omega_{\underline{l}} = \Omega_{l_1,\ldots,l_d}$. For the combination technique we now in particular consider all grids $\Omega_{\underline{l}}$ with

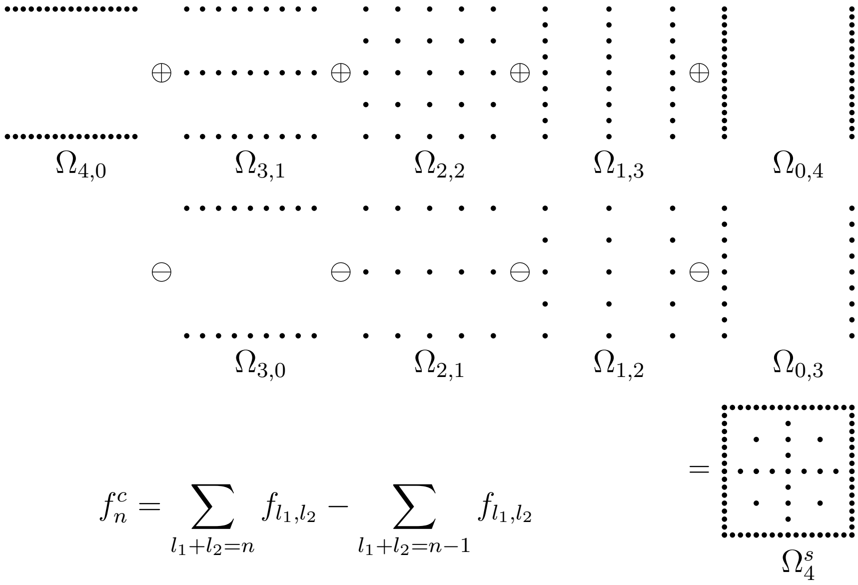
$$|\underline{l}|_1 := l_1 + ... + l_d = n - q, \quad q = 0, .., d-1, \quad l_t \geq 0,$$

set up and solve the associated problems (4). The original combination technique [11] now linearly combines the resulting discrete solutions $f_{\underline{l}}(\underline{x}) \in V_{\underline{l}}$ from the partial grids $\Omega_{\underline{l}}$ according to the formula

$$f_n^c(\underline{x}) := \sum_{q=0}^{d-1} (-1)^q \binom{d-1}{q} \sum_{|\underline{l}|_1 = n-q} f_{\underline{l}}(\underline{x}).$$

The function f_n^c lives in the sparse grid space

$$V_n^s := \bigoplus_{\substack{|\underline{l}|_1 = n-q \\ q = 0, ..., d-1 \quad l_t \geq 0}} V_{\underline{l}}.$$



$$f_n^c = \sum_{l_1+l_2=n} f_{l_1,l_2} - \sum_{l_1+l_2=n-1} f_{l_1,l_2}$$

Figure 1: Grids involved for the combination technique of level $n = 4$ in two dimensions.

The space V_n^s has dimension of order $\mathcal{O}(h_n^{-1}(\log(h_n^{-1}))^{d-1})$ in contrast to $\mathcal{O}(h_n^{-d})$ for conventional grid based approaches.

For the two-dimensional case, we display the grids needed in the combination formula of level 4 in Figure 1 and give the resulting sparse grid.

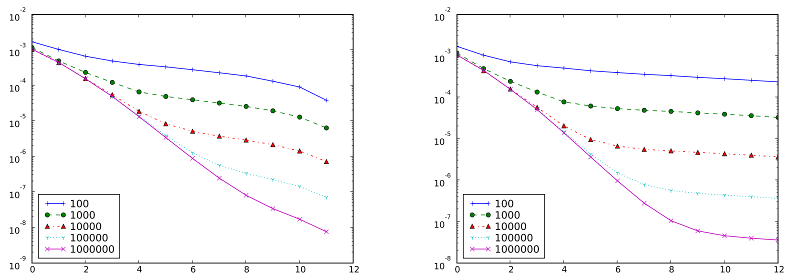
The approximation properties of the combination technique are connected to sparse grids in two different ways. First, using extrapolation arguments, it can be shown that the approximation property of the combination technique is of the order $\mathcal{O}(h_n^2 \cdot \log(h_n^{-1})^{d-1})$ as long as error expansions of the form

$$f - f_{\underline{l}} = \sum_{i=1}^{d} \sum_{j_1,\dots,j_m \subset 1,\dots,d} c_{j_1,\dots,j_m}(h_{j_1},\dots,h_{j_m}) \cdot h_{j_1}^2 \cdot \dots \cdot h_{j_m}^2 \tag{5}$$

for the partial solutions hold [11].

Second, viewing the minimisation of (1) as a projection, one can show that the combination technique is an exact projection into the underlying sparse grid space (and therefore of approximation order $\mathcal{O}(h_n^2 \cdot \log(h_n^{-1})^{d-1})$) only if the partial projections commute, i.e., the commutator $[P_{V_1}, P_{V_2}] := P_{V_1}P_{V_2} - P_{V_2}P_{V_1}$ is zero for all pairs of involved grids [14].

Figure 2: Convergence against highly refined solution measured using (1) for $\lambda = 0.01$. Left: full grid, Right: sparse grid combination technique.



3 Asymptotics for $h \to 0$

3.1 Empirical convergence behaviour

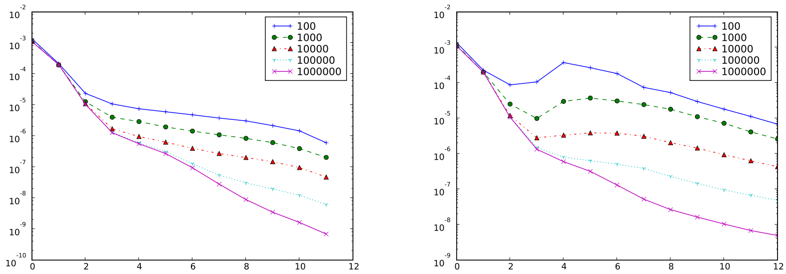
We first consider the convergence behaviour of full grid solutions for a simple regression problem, measured against a highly refined grid (due to the lack of an exact solution). As in [5] we consider the function

$$f(x, y) = e^{-(x^2+y^2)} + x \cdot y.$$

in the domain $[0, 1]^2$ where the data positions are chosen randomly. To study the behaviour with different number of data we take hundred, thousand, ten-thousand, hundred-thousand and one million data points. In Figure 2, left, we show the difference between a full grid solution of level l and one of level $n = 12$ using the functional (1) as a norm, in this experiment we use $\lambda = 0.01$. We see that the difference shows two different types of convergence behaviour, first it displays the usual h^2 convergence, but this convergence deteriorates when h gets smaller. Furthermore, the more data is used, the later this change in the reduction rate takes place. Qualitatively these observations do not depend on the regularisation parameter λ, in Figure 3, left, we provide results for a smaller regularisation parameter of $\lambda = 0.0001$.

A somewhat different picture arises if we employ the sparse grid combination technique. In Figure 2, right, we show the behaviour of the difference of the solution with the combination technique and the full grid solution of level $n = 12$, again using the functional (1) as a norm and $\lambda = 0.01$. We see a similar type of behaviour, although the change in convergence is more clearly present. On the other hand, using $\lambda = 0.0001$, shown in Figure 3, right, we

Figure 3: Convergence against highly refined solution measured using (1) for $\lambda = 0.0001$. Left: full grid, Right: sparse grid combination technique.



observe for the sparse grid combination technique on smaller data sets an intermediate rise of the difference to the highly refined full grid solution.

In a different experiment we consider the combination technique for the solution and record the following values: the least squares error on the given data $(\underline{x}_i, y_i), i = 1, \ldots, m$, the regularisation term, and the residual (1) of the approximation using $m = 1000$ data, $\lambda = 10^{-2}$, $\lambda = 10^{-3}$, $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$. These values are presented in Figure 4. We observe that for small regularisation parameters the errors increase, which cannot happen with a true variational discretisation ansatz for the residual. With larger regularisation parameters the now stronger influence of the smoothing term results in a (more) stable approximation method. Note that this instability is more common and significant in higher dimensions.

In the following we will present some theory which supports these observations.

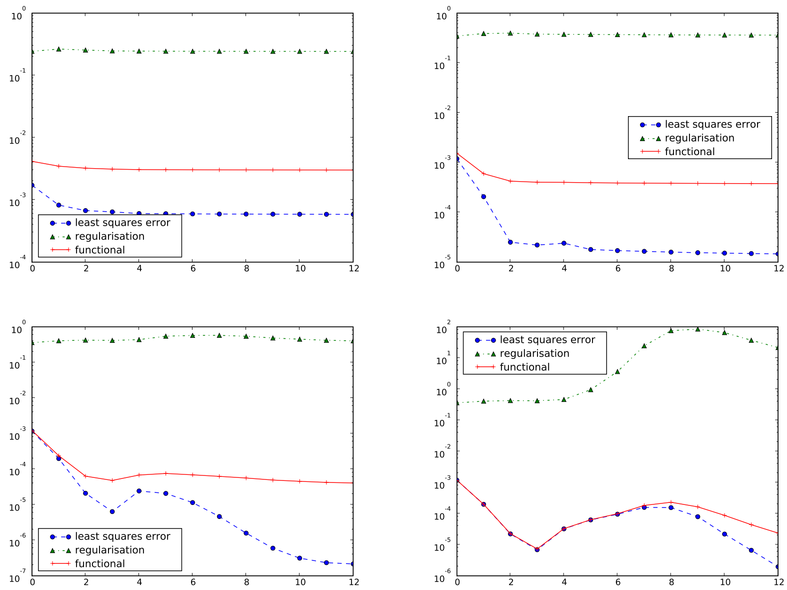
3.2 Case where $m = m(h)$ is large

If the number m of data points is large, the data term in $R(f)$ approximates an integral. For simplicity, we discuss only the case of $\Omega = (0, 1)^d$ and a uniform distribution of the data positions, however, most results hold for more general domains and probability distributions. Then, if $f_*(\underline{x})$ is a square integrable random field with $f_*(\underline{x}_i) = y_i$ and

$$J(f) = \lambda \int_\Omega |\nabla f(\underline{x})|^2 dx + \int_\Omega (f(\underline{x}) - f_*(\underline{x}))^2 dx \tag{6}$$

it follows that $J(f) \approx R(f)$ for large m.

Figure 4: Value of the least squares error, the regularisation term and the residual (1) for 1000 data using the combination technique. Top Left: with $\lambda = 10^{-2}$, Top Right: with $\lambda = 10^{-3}$. Bottom Left: with $\lambda = 10^{-4}$, Bottom Right: with $\lambda = 10^{-6}$.



Consider a finite element space $V_N \subset C(\Omega)$ with rectangular elements Q of side lengths $h_1, \dots, h_d$ and multilinear element functions. The number k of data points $\underline{x}_i$ contained in any element Q is a binomially distributed random variable with expectation $m \cdot h_1 \cdots h_d$. When mapped onto a reference element $I = [0, 1]^d$, the data points $\xi_1, \dots, \xi_k$ are uniformly distributed within I.

Let ϕ be a continuous function on Q with expectation $\overline{\phi} = \int_I \phi(\xi)d\xi$ and variance $\sigma(\phi)^2 = \int_I (\phi(\xi) - \overline{\phi})^2 d\xi$. By the central limit theorem, the probability that the inequality

$$\left| \int_I \phi(\xi)d\xi - \frac{1}{k}\sum_{i=1}^{m} \phi(\xi_i) \right| \leq \frac{c\sigma(\phi)}{\sqrt{m}}$$

holds for $m \to \infty$ is in the limit $\frac{1}{\sqrt{2\pi}} \int_{-c}^{c} e^{-t^2/2}dt$.

As we will apply the first lemma of Strang [1] on the bilinear forms corresponding to $J(f)$ and $R(f)$ we need this bound for the case of $\phi(\xi) =$

$u(\xi)v(\xi)$. Using a variant of the Poincaré-Friedrichs inequality [1] with the observation that the average of $w := \phi - \overline{\phi}$ equals zero, the product rule, the triangular inequality, and the Cauchy-Schwarz inequality we obtain

$$\sigma(\phi) \leq C\sqrt{\int_I |\nabla \phi(\xi)|^2 \, d\xi} \leq C\left(\|v\|\|\nabla u\| + \|u\|\|\nabla v\|\right) \leq C\|u\|_1\|v\|_1.$$

Transforming this back onto the actual elements Q, summing up over all the elements and applying the Cauchy-Schwarz inequality gives, with high probability for large m, the bound:

$$\left|\int_\Omega u(\underline{x})v(\underline{x})dx - \frac{1}{m}\sum_{i=1}^m u(\underline{x}_i)v(\underline{x}_i)\right| \leq \frac{c\|u\|_1\|v\|_1}{\sqrt{m}}.$$

A similar bound can be obtained for the approximation of the right hand side in the Galerkin equations. We now can apply the first lemma of Strang to get the bound

$$\|f - f_N\|_1 \leq C\left(\|f - f_N^{\text{best}}\|_1 + \frac{\|f\|_1}{\sqrt{k}}\right),$$

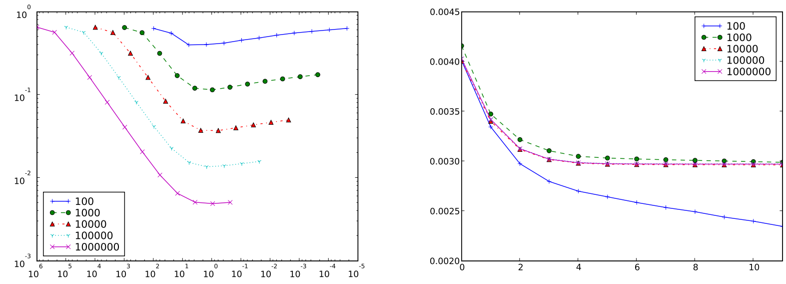
where f_N^{best} is the best approximation of f in the $\|\cdot\|_1$-norm.

This bound is very flexible and holds for any intervals I – it does not depend on the particular h_i just on the product. This is perfectly adapted to the situation of the sparse grid combination technique where one has on average $k_{\underline{l}} = 2^{-|\underline{l}|}m$ data points per element on level $|\underline{l}|$. It is known that the combination technique acts like an extrapolation method for the Poisson problem. This is not the case in the regression problem as there is no cancellation of the random errors. Assuming that the errors $e_{\underline{l}}$ are i.i.d. we conjecture that the error of an approximation using the sparse grid combination technique (for large enough k) satisfies a bound of the form

$$\|f - f_{\text{sg}}\|_1 \leq C\left(\|f - f_{\text{sg}}^{\text{best}}\|_1 + \frac{\|f\|_1\sqrt{\sum_{\underline{l}} c_{\underline{l}}^2 2^{|\underline{l}|}}}{\sqrt{m}}\right) \quad (7)$$

where, as usual, $c_{\underline{l}}$ are the combination coefficients and the summation is over all full subgrids of the sparse grid indexed by the vector of levels $\underline{l}$.

Figure 5: Left: H_1-seminorm difference of the solutions of $J(f)$ and $R(f)$ plotted against the number k of data points per cell. Right: Decrease of functional R.



To study this effect experimentally let us consider (6) with

$$\begin{aligned} f_*(x,y) &= -100\lambda \cdot \left((2x-1)\left(\frac{1}{4}y^4 - \frac{1}{3}y^3\right) + \left(\frac{1}{3}x^3 - \frac{1}{2}x^2\right)\left(3y^2 - 2y\right) \right) \\ & \quad +100 \cdot \left(\frac{1}{3}x^3 - \frac{1}{2}x^2\right)\left(\frac{1}{4}y^4 - \frac{1}{3}y^3\right). \end{aligned}$$

The function $f(x,y) = 100 \cdot \left(\frac{1}{3}x^3 - \frac{1}{2}x^2\right)\left(\frac{1}{4}y^4 - \frac{1}{3}y^3\right)$ is the solution of the resulting continuous problem.

As indicated, if we now assume that a Monte-Carlo approach is used to compute the integrals $\int_\Omega f(x)g(x)dx$ and $\int_\Omega f(x)f_*(x)dx$ in the Galerkin equations we obtain the regularised least squares formulation (1). For different data set sizes we measure the difference between the resulting discrete solutions using the sparse grid combination technique for the now fixed number of data and the above continuous solution. In Figure 5, left, we show the behaviour of the error for increasing discretisation levels measured in the H^1-seminorm. At first we have the "usual" decrease in the error, but after about one sample point per element the error increases instead due to the second term in (7).

3.3 Behaviour as $N \to \infty$ – overfitting in data points

The bound (7) holds only asymptotically in k and thus for a fixed number of data and very small mesh size it will break down. We give a bound for the residuals $|y_i - f_N(\underline{x}_i)|$ first for general reproducing kernel Hilbert spaces and then for the special case of regular grids. For a discussion of the framework of reproducing kernel Hilbert spaces, see [16, 18]. Let in the following V_0 always be the space of constant functions and let the L_2-orthogonal decomposition

$$V_N = V_0 \oplus V_N \cap V_0^{\perp}$$

hold, where $V_N \cap V_0^{\perp}$ is the space of functions with zero mean. The seminorm $\|\nabla f\|$ defines a norm on $V_N \cap V_0^{\perp}$. As every linear functional is continuous on a finite dimensional linear space there exists a reproducing kernel $k_{\underline{x}} \in V_N \cap V_0^{\perp}$ such that $f(\underline{x}) = \langle \nabla k_{\underline{x}}, \nabla f \rangle$ for $f \in V_N \cap V_0^{\perp}$, and one easily verifies that for $f \in V_N$ one gets

$$f(\underline{x}) = \langle 1, f \rangle + \langle \nabla k_{\underline{x}}, \nabla f \rangle, \quad f \in V_N.$$

Introducing this into the functional R gives

$$R(f) = \lambda \langle \nabla f, \nabla f \rangle + \frac{1}{m} \sum_{i=1}^{m} \left(\langle 1, f \rangle + \langle \nabla k_{\underline{x}_i}, \nabla f \rangle - y_i \right)^2 .$$

Due to the representer theorem [18] it follows that the minimiser of R is a linear combination of reproducing kernels

$$f_N(\underline{x}) = c_0 + \sum_{i=1}^{m} c_i k_{\underline{x}_i}(\underline{x}). \tag{8}$$

Inserting this in the formula for R gives

$$R(f) = \lambda\, \underline{c}^T \mathcal{K} \underline{c} + \frac{1}{m} (c_0 \underline{1} + \mathcal{K}\underline{c} - \underline{y})^T (c_0 \underline{1} + \mathcal{K}\underline{c} - \underline{y}), \tag{9}$$

where the kernel matrix $\mathcal{K}$ has entries $\mathcal{K}_{ij} = k_{\underline{x}_i}(\underline{x}_j)$ and $\underline{1}$ is the vector of ones.

The reproducing kernel depends on N and to make this clear we will denote k by k_N and $\mathcal{K}$ by $\mathcal{K}_N$ in the following, but also use the notation $k_N(\underline{x}_i, \underline{x}_j) = k_{\underline{x}_i}(\underline{x}_j)$.

Proposition 1. *Let $V_0 = \{f = const.\}, V_1, V_2, \ldots$ be a sequence of linear function spaces and $f_N \in V_N$ be the minimiser of $R(f)$ in V_N. Furthermore assume that $R(f_N) \le C$ for some $C > 0$ and all integers N. Then*

$$|y_i - f_N(\underline{x}_i)| \le \frac{2m \left(\lambda \|\underline{y}\|_\infty + \kappa C^{1/2} m^{1/2} \right)}{k_N(\underline{x}_i, \underline{x}_i)},$$

where k_N is the reproducing kernel of $V_N \cap V_0^{\perp}$ with respect to the H_1-seminorm and $\kappa = \sup_N \max_{i \ne j} |k_N(\underline{x}_i, \underline{x}_j)|$.

Proof. As $R(f_N)$ is bounded the sum of the squared residuals is bounded:

$$\frac{1}{m}\sum_{i=1}^{m}|y_i - f_N(\underline{x}_i)|^2 \le C.$$

Taking the gradient of (9) with respect to c_0 and multiplying with m gives for the minimum

$$mc_0 + \underline{1}^T\mathcal{K}_N\underline{c} - \underline{1}^T\underline{y} = 0.$$

Taking the gradient of (9) with respect to $\underline{c}$ and multiplying the result with m times the inverse of $\mathcal{K}_N$ gives

$$m\lambda\underline{c} + (c_0\underline{1} + \mathcal{K}_N\underline{c} - \underline{y}) = 0. \tag{10}$$

In block form this is

$$\begin{bmatrix} m & \underline{1}^T\mathcal{K}_N \\ \underline{1} & m\lambda\mathcal{I} + \mathcal{K}_N \end{bmatrix} \begin{bmatrix} c_0 \\ \underline{c} \end{bmatrix} = \begin{bmatrix} \underline{1}^T\underline{y} \\ \underline{y} \end{bmatrix}.$$

An elimination step leads then to the saddle point system

$$\begin{bmatrix} 0 & \underline{1}^T \\ \underline{1} & m\lambda\mathcal{I} + \mathcal{K}_N \end{bmatrix} \begin{bmatrix} c_0 \\ \underline{c} \end{bmatrix} = \begin{bmatrix} 0 \\ \underline{y} \end{bmatrix}. \tag{11}$$

From equations (8) and (10) one gets $y_i - f_N(\underline{x}_i) = m\lambda c_i$ and thus

$$|c_i| \le \sqrt{\sum_{i=1}^{m} c_i^2} \le \frac{1}{\lambda m^{1/2}}C^{1/2}.$$

Now let $\mathcal{K}_N = \mathcal{D} + \mathcal{H}$ where $\mathcal{D}$ contains the diagonal and $\mathcal{H}$ the offdiagonal components of the kernel matrix $\mathcal{K}_N$, respectively, and set $\underline{r} = -\mathcal{H}\underline{c}$. One then has

$$\|\underline{r}\|_\infty \le \frac{\kappa}{\lambda}C^{1/2}m^{1/2}.$$

Introducing $\underline{r}$ into equation (11) gives

$$\begin{bmatrix} \underline{0} & 1^T \\ \underline{1} & \mathcal{D} + m\lambda I \end{bmatrix} \begin{bmatrix} c_0 \\ \underline{c} \end{bmatrix} = \begin{bmatrix} 0 \\ \underline{y} + \underline{r} \end{bmatrix}.$$

Eliminating $\underline{c}$ gives

$$|c_0| = \left|\frac{\underline{1}^T(\mathcal{D} + m\lambda\mathcal{I})^{-1}(\underline{y} + \underline{r})}{\underline{1}^T(\mathcal{D} + m\lambda\mathcal{I})^{-1}\underline{1}}\right| \le \|\underline{y} + \underline{r}\|_\infty \le \|\underline{y}\|_\infty + \frac{\kappa}{\lambda}C^{1/2}m^{1/2}$$

as $\mathcal{D}+m\lambda\mathcal{I}$ has only positive elements and so c_0 is uniformly bounded in N. Now $(\mathcal{D}+m\lambda\mathcal{I})\underline{c} = \underline{y}+\underline{r}-c_0\underline{1}$ and one then gets

$$\begin{aligned} |c_i| &\leq (k_N(\underline{x}_i,\underline{x}_i)+m\lambda)^{-1}2\|\underline{y}+\underline{r}\|_\infty \\ &\leq k_N(\underline{x}_i,\underline{x}_i)^{-1}2\|\underline{y}+\underline{r}\|_\infty \\ &\leq 2k_N(\underline{x}_i,\underline{x}_i)^{-1}\left(\|\underline{y}\|_\infty + \frac{\kappa}{\lambda}C^{1/2}m^{1/2}\right). \end{aligned}$$

With that one achieves the desired bound as

$$|y_i - f_N(\underline{x}_i)| = m\lambda|c_i| \leq \frac{2m\left(\lambda\|\underline{y}\|_\infty + \kappa C^{1/2}m^{1/2}\right)}{k_N(\underline{x}_i,\underline{x}_i)}.$$

$\square$

In the following let V_N be the space of piecewise multilinear functions of d variables using isotropic rectangular grids with size $h = 1/(N^{1/d}-1)$. We will now formulate discrete versions of some well known analytical results. The foundation is the following lemma which could be called "discrete Sobolev inequality", it can be found in [2, 3] for the cases $d = 1, 2, 3$.

Lemma 1 (Discrete Sobolev inequality). *There are constants $C_d > 0$ such that for all $f \in V_N$*

1. $|f(x)| \leq C_d\|f\|_1$ *for* $d = 1$
2. $|f(x)| \leq C_d(1+|\log h|)\|f\|_1$ *for* $d = 2$
3. $|f(x)| \leq C_d h^{1-d/2}\|f\|_1$ *for* $d > 2$

where h is the grid size of V_N. These bounds are tight, i.e., they are attained for some functions $f \in V_N$, and, for $d > 2$ one has

$$C_d \geq \frac{1}{\sqrt{3d+1}}\left(\frac{3}{2}\right)^{d/2}.$$

Proof. For $d = 1$ the inequality is just the usual Sobolev inequality, and the inequality for $d = 2$ is known as discrete Sobolev inequality. The case of $d > 2$ uses the Sobolev embedding $H^1(\Omega) \hookrightarrow L^{2d/(d-2)}(\Omega)$ from which one gets

$$\|f\|_{L^{2d/(d-2)}}(\Omega) \leq A_d\|f\|_1$$

and the inverse inequality

$$\|f\|_{L^\infty(\Omega)} \le B_d h^{-d/p} \|f\|_{L^p(\Omega)}$$

for some constants A_d, B_d and $C_d = A_d B_d$.

To show the tightness for $d = 1$ one chooses $f = k_y$ (the reproducing kernel). For $d = 2$ this can be found in the literature. Here we show that this is also the case for $d > 2$. We consider for V_N piecewise multilinear functions with a regular isotropic grid with mesh size h. Choose as f a tent function which is one on an interior grid point and zero on all other grid points. In one of the cells touching the point where $f(\underline{x}) = 1$ one has, after translation,

$$f(\underline{x}) = \frac{x_1 \cdots x_d}{h^d}.$$

Since to have an interior point one requires $h \le 1/2$, and f behaves accordingly in all 2^d cells, one then has for the L_2 norm

$$\int_\Omega f(\underline{x})^2 \, dx = 2^d \left(\frac{h}{3}\right)^d$$

and for the L_2 norm of a derivative $\partial f / \partial x_i$:

$$\int_\Omega \left(\frac{\partial f(\underline{x})}{\partial x_i}\right)^2 dx = 2^d \left(\frac{h}{3}\right)^{d-1} \frac{1}{h}$$

and one gets

$$\|f\|_1^2 = \left(\frac{2}{3}\right)^d (h^d + 3dh^{d-2}).$$

Consequently, as $\|f\|_\infty = 1$ one has

$$\frac{\|f\|_\infty}{\|f\|_1} = h^{1-d/2} \left(\frac{3}{2}\right)^{d/2} \frac{1}{\sqrt{3d + h^2}}$$

from which the tightness of the bound and the lower bound for C_d follow. □

A direct consequence of this lemma is a discrete Poincaré inequality which is a generalisation of the inequality for two dimensions in [3].

Lemma 2 (Discrete Poincaré inequality)**.** *Let* $\underline{x}_0 \in \Omega$ *and* $f_0 = f(\underline{x}_0)$ *for* $f \in V_N$*. Then there exist constants* $C_d > 0$ *such that*

1. $\|f - f_0\| \le C_d \|\nabla f\|$ *for* $d = 1$
2. $\|f - f_0\| \le C_d(1 + |\log h|)\|\nabla f\|$ *for* $d = 2$
3. $\|f - f_0\| \le C_d h^{1-d/2}\|\nabla f\|$ *for* $d > 2$.

Proof. The proof follows the proof of the 2D version in [3]. First, by the triangle inequality one has

$$\|f - f_0\| \le \|f - \overline{f}\| + |\overline{f} - f_0|$$

where $\overline{f}$ is the mean value of f. The (standard) Poincaré inequality gives $\|f - \overline{f}\| \le C\|\nabla f\|$ for some $C > 0$. For the second term in the sum one uses the discrete Sobolev inequality to get a bound for $|\overline{f} - f_0| \le \|f - \overline{f}\|_\infty$ in terms of the Sobolev norm $\|f - \overline{f}\|_1$. One then inserts the (standard) Poincaré inequality to get a bound for the component $\|f - \overline{f}\|$ of the Sobolev norm and so gets the claimed inequalities. □

With these inequalities one can now derive bounds for the values of the energy norm and thus establish a "discrete V-ellipticity".

Proposition 2 (Discrete V-ellipticity)**.** *The energy norm for the penalised least squares problem on* V_N *satisfies the inequalities*

$$\alpha_{d,h}\|f\|_1 \le \|f\|_{RLS} \le \beta_{d,h}\|f\|_1, \quad f \in V_N$$

where there exist c_d *and* C_d *such that for all* N*,* m *and* λ *one has:*

1. $\alpha_{d,h} = (c_d\lambda^{-1/2} + \lambda^{-1/2} + 1)^{-1}$ *and* $\beta_{d,h} = C_d + \sqrt{\lambda}$ *for* $d = 1$
2. $\alpha_{d,h} = (c_d\lambda^{-1/2}(1+|\log h|)+\lambda^{-1/2}+1)^{-1}$ *and* $\beta_{d,h} = C_d(1+|\log h|)+\sqrt{\lambda}$ *for* $d = 2$
3. $\alpha_{d,h} = (c_d\lambda^{-1/2}h^{1-d/2} + \lambda^{-1/2} + 1)^{-1}$ *and* $\beta_{d,h} = C_d h^{1-d/2} + \sqrt{\lambda}$ *for* $d > 2$.

Proof. Consider first the upper bounds. One has

$$\|f\|_{\mathrm{RLS}} \le \|f\|_\infty + \sqrt{\lambda}\|f\|_1.$$

The upper bounds follow then directly from the discrete Sobolev inequality.

For the lower bounds observe that as $f \in V_N$ is continuous there exists an $\underline{x}_0 \in \Omega$ such that for $f_0 = f(\underline{x}_0)$ one has

$$f_0^2 = \frac{1}{m} \sum_{i=1}^{m} f(\underline{x}_i)^2.$$

We then have by the triangle inequality

$$\|f\| \leq \|f - f_0\| + |f_0|$$

and by the discrete Poincaré inequality

$$\|f\| \leq (c_d \lambda^{-1/2} h^{1-d/2} + 1)\|f\|_{\mathrm{RLS}}.$$

From this one gets the bound

$$\|f\|_1 \leq (c_d \lambda^{-1/2} h^{1-d/2} + \lambda^{-1/2} + 1)\|f\|_{\mathrm{RLS}}$$

for the case $d > 2$. The other two bounds follow by the same argument. $\square$

In the following we achieve specific bounds for $k_N(\underline{x}_i, \underline{x}_i)$ and get:

Proposition 3. *Let V_N be the spaces of piecewise multilinear functions of d variables and isotropic rectangular grids with size $h = 1/(N^{1/d} - 1) \leq h_0$ for some appropriately chosen $h_0 \leq 1/2$. Furthermore, let f_N be the minimiser of $R(f)$ in V_N and let $R(f_N) \leq C$ uniformly in N for some $C > 0$. Then*

$$|y_i - f_N(\underline{x}_i)| \leq C_d m(\lambda \|y\|_\infty + \kappa C^{1/2} m^{1/2}) h^{d-2}$$

for some $C_d \geq 0$ and $d \geq 3$ and

$$|y_i - f_N(\underline{x}_i)| \leq C_2 m(\lambda \|y\|_\infty + \kappa C^{1/2} m^{1/2})(1 + |\log h|)^{-2}$$

for some $C_2 \geq 0$ and $d = 2$.

Proof. We use the discrete Sobolev inequality (Lemma 1) to get a lower bound for $k_N(\underline{x}_i, \underline{x}_i) = k_{\underline{x}_i}(\underline{x}_i)$ and then apply Proposition 1. Here we explicitly show the case $d \geq 3$, the case $d = 2$ is similar.

Now let $\tilde{k}_{\underline{x}}$ be the reproducing kernel of V_N (which exists as V_N is finite dimensional) with respect to the H^1-norm $\|\cdot\|_1$ such that

$$f(x) = (\tilde{k}_{\underline{x}}, f)_1, \quad f \in V_N.$$

Using Cauchy-Schwarz it follows that

$$\frac{f(x)^2}{\|f\|_1^2} \le \|\tilde{k}_{\underline{x}}\|_1^2 = \tilde{k}_{\underline{x}}(x) = \tilde{k}(\underline{x}, \underline{x}), \quad f \in V_N.$$

By the discrete Sobolev inequality, and specifically its tightness, there exists an $f \in V_N$ such that

$$\frac{f(\underline{x})^2}{\|f\|_1^2} \ge C_d h^{2-d}$$

and consequently $\tilde{k}(\underline{x}, \underline{x}) \ge C_d h^{2-d}$.

Since the reproducing kernel for the direct sum of two perpendicular subspaces is the sum of the reproducing kernels [18] the previously introduced reproducing kernel $k_{\underline{x}}$ of $V_N \cap V_0^{\perp}$ with respect to the Sobolev seminorm $(\nabla\cdot, \nabla\cdot)$ satisfies

$$k_{\underline{x}} = \tilde{k}_{\underline{x}} - 1.$$

For small enough h one thus has

$$k(\underline{x}, \underline{x}) \ge C_d h^{2-d}$$

for some $C_d > 0$. The desired bound on the residual is then obtained by an application of Proposition 1. $\square$

Note that for the minimal distance of any two points of the given data it holds $|\underline{x}_i - \underline{x}_j| \ge \epsilon > 0$, therefore κ, the supremum over $k_N(\underline{x}_i, \underline{x}_j), i \ne j$ is bounded in this proposition. A proof for $d = 2$ can be found in a similar situation in Section 4.7 in [12], where $|k_N(\underline{x}_i, \underline{x}_j)| \le C(1 + |\log(|\underline{x}_i - \underline{x}_j| + h)|)$ is shown.

The case $d = 2$ is illustrated in Figure 5, right. While we have proved this result for a special V_N one can easily extend it to other spaces V_N.

3.4 Asymptotics between the data points

While the approximations f_N do converge on the data points they do so very locally. In an area outside a neighbourhood of the data points the f_N tend to converge to a constant function so that f_N may recover fast oscillations only if sufficient data is available and only close to the data points.

We have seen that the residuals $f_N(x_i) - y_i$ go to zero with increasing N, now we will show that even the whole penalised squared residual goes to zero for dimensions larger than 2.

Proposition 4. *The value of functional R converges to zero on the estimator f_N and*

$$R(f_N) \le Cm\lambda h^{d-2}$$

for some $C > 0$. As $\|\nabla f_N\| \le \sqrt{R(f_N)}$ it follows that $\|\nabla f_N\| \le C\sqrt{m\lambda h^{(d-2)}}$.

Proof. While we only consider regular partitioning with hyper-cubical elements Q, the proof can be generalised to other elements. First, let b_Q be a member of the finite element function space such that $b_Q(\underline{x}) = 1$ for $\underline{x} \in Q$ and $b_Q(\underline{x}) = 0$ for $\underline{x}$ in any element which is not a neighbour of Q. One can easily see that

$$\int_Q |\nabla b_Q|^2 dx \le Ch^{d-2}. \tag{12}$$

Choose h such that for the k-th component of $\underline{x}_i$ one has

$$|\underline{x}_{i,k} - \underline{x}_{j,k}| > 3h, \quad \text{for } i \ne j. \tag{13}$$

In particular, any element contains at most one data point. Let furthermore Q_i be the element containing $\underline{x}_i$, i.e., $\underline{x}_i \in Q_i$. Then one sees that g defined by

$$g(\underline{x}) = \sum_{i=1}^m y_i b_{Q_i}(\underline{x})$$

interpolates the data, i.e., $g(\underline{x}_i) = y_i$. Consequently,

$$R(g) = \lambda \int_\Omega |\nabla g|^2 dx.$$

Because of the condition on h one has for the supports $\operatorname{supp} b_{Q_i} \cap \operatorname{supp} b_{Q_j} = \emptyset$ for $i \ne j$ and so

$$R(g) = \lambda \sum_{i=1}^m y_i^2 \int_\Omega |\nabla b_{Q_i}|^2 dx$$

and, thus,

$$R(g) \le Cm\lambda h^{d-2}.$$

It follows that $\inf R(f) \le R(g) \le Cm\lambda h^{d-2}$. □

We conjecture that in the case of $d = 2$ one has $R(f_N) \le Cm\lambda/|\log h|$.

Proposition 5. *For $d \ge 3$ the solution f_N of (1) converges towards a constant function almost everywhere.*

Proof. Let $u_N := f_N - \int f_N dx$. Obviously $\int u_N dx = 0$ and with Proposition 4 it holds $\int |\nabla u_N|^2 dx \leq Cmh^{(d-2)}$. Using a suitable form of the Poincaré inequality, e.g., (7.45) from [10], one has the existence of a constant ω such that

$$\|u_N\|_1^2 \leq \omega \|\nabla u_N\|^2 \leq \omega Cmh^{(d-2)},$$

It follows $u_N \to 0$ in H^1 for $h \to 0$ and therefore $f_N = \int f_N dx$ almost everywhere. $\square$

In experiments for large N one sees that f_N varies only very close to the data points and is otherwise almost exactly constant. This suggests the following decomposition:

$$f_N = \overline{f} + \sum_{k=1}^{m} c_k g_k$$

where the values $g_k(x)$ are very small except in a neighbourhood of x_k respectively.

Let us assume the normalisation

$$g_k(\underline{x}_k) = 1$$

and that the bounds

$$2\lambda |\langle \nabla g_i, \nabla g_j \rangle| \leq \epsilon_h \tag{14}$$

and

$$|g_i(\underline{x}_j)|^2 \leq \epsilon_h \tag{15}$$

hold for some $\epsilon_h = \mathcal{O}(h^{2(d-2)})$. The existence of such g_i will be shown shortly and follows from the localisation of the variations of f_N close to the data points. The exact size of ϵ_h is of less importance as long as it is $o(R(f_N))$. Under this assumption one then gets for the penalised squared residual

$$R(f_N) = \frac{1}{m} \sum_{i=1}^{m} (\overline{f} + c_i - y_i)^2 + \lambda \sum_{i=1}^{m} c_i^2 \|\nabla g_i\|^2 + \mathcal{O}(\epsilon_h).$$

An approximation to f_N can then be obtained by minimising the approximate penalised squared residual to find functions g_i and coefficients c_i. Through the approximation the functions g_i are "decoupled" in R.

Working in the weak formulation one observes that the reproducing kernels $k_{\underline{x}_i}$ minimise the functional

$$\phi(u) = \frac{1}{2}\|\nabla u\|^2 - u(\underline{x}_i)$$

and that

$$g_i = \frac{k_{\underline{x}_i}}{\|\nabla k_{\underline{x}_i}\|^2}$$

minimises $\frac{1}{2}\|\nabla u\|^2$ under the constraint $u(\underline{x}_i) = 1$. Using results from the previous section one can show that furthermore the values

$$g_i(\underline{x}_j) = \langle \nabla g_i, \nabla k_{\underline{x}_j} \rangle = \|\nabla k_{\underline{x}_j}\|^2 \langle \nabla g_i, \nabla g_j \rangle = \langle \nabla k_{\underline{x}_i}, \nabla k_{\underline{x}_j} \rangle / \|\nabla k_{\underline{x}_i}\|^2$$

are very small and fulfil conditions (14) and (15). It follows thus that this choice of g_i leads to good candidates for f_N and it remains to determine the coefficients c_i and $\overline{f}$. Inserting these particular g_i one then gets for the approximate penalised squared residuals

$$R(f_N) = \frac{1}{m}\sum_{i=1}^{m}\left(\overline{f} + c_i - y_i\right)^2 + \lambda \sum_{i=1}^{m} \frac{c_i^2}{\|\nabla k_{\underline{x}_i}\|^2} + \mathcal{O}(\varepsilon_h).$$

Setting the gradient with respect to c_i to zero gives the formula

$$c_i = \frac{y_i - \overline{f}}{1 + m\lambda/\|\nabla k_{\underline{x}_i}\|^2},$$

while setting the gradient with respect to $\overline{f}$ to zero gives

$$\overline{f} = \frac{1}{m}\sum_{i=1}^{m}(y_i - c_i).$$

One now substitutes c_i using the formula given above and solves for $\overline{f}$ to get (after some algebraic manipulations):

$$\overline{f} = \frac{\sum_{i=1}^{m} y_i/(\|\nabla k_{\underline{x}_i}\|^2 + m\lambda)}{\sum_{i=1}^{m} 1/(\|\nabla k_{\underline{x}_i}\|^2 + m\lambda)}.$$

When $N \to \infty$ then all the $\|\nabla k_{\underline{x}_i}\|^2$ go to the same constant K which goes to infinity and it follows that

$$\overline{f} \to \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

Furthermore $c_i \to y_i - \overline{y}$ and one gets the approximation

$$f_N \approx \overline{y} + \sum_{i=1}^{m} (y_i - \overline{y}) \frac{k_{\underline{x}_i}}{\|k_{\underline{x}_i}\|^2}$$

and consequently one gets the asymptotic behaviour "between the points" as

$$f_N \to \overline{y}.$$

We are now going to show that the estimator f_N is bounded everywhere. First we give the following lemma, which is a variant of the Aubin-Nitsche-Lemma.

Lemma 3. *For f_N as above it holds for $d \geq 4$*

$$\left\| f_N - \int f_N \right\| \leq Ch \|\nabla f_N\|,$$

where C only depends on Ω, i.e., the dimension.

Proof. Let $u_N := f_N - \int f_N$. We define the function space

$$\widehat{H^1} := \left\{ f \in H^1 \mid \int_\Omega f(\underline{x}) = 0 \right\}.$$

For $g \in \widehat{H^1}$ let $\varphi_g \in \widehat{H^1}$ be

$$\langle \nabla \varphi_g, \nabla v \rangle = \langle g, v \rangle, \qquad v \in \widehat{H^1}.$$

It holds $u_N \in \widehat{H^1}$, therefore $\langle \nabla \varphi_g, \nabla u_N \rangle = \langle g, u_N \rangle$.

Since f_N fulfills the Galerkin equations and by definition of u_N

$$\langle \nabla u_N, \nabla v \rangle = \langle \nabla f_N, \nabla v \rangle = \frac{1}{\lambda m} \sum_{i=1}^{m} \left(y_i - \left(u_N(\underline{x}_i) + \int f_N \right) \right) v(\underline{x}_i), \qquad \forall v \in V_N,$$

and so

$$\langle u_N, g \rangle = \langle \nabla u_N, \nabla(\varphi_h - v) \rangle + \frac{1}{\lambda m} \sum_{i=1}^{m} \left(y_i - u_N(\underline{x}_i) - \int f_N \right) v(\underline{x}_i), \qquad \forall v \in V_N \cap \widehat{H^1}.$$

It follows that

$$\|u_N\| = \sup_{g\in\widehat{H^1}} \frac{\langle u_N, g\rangle}{\|g\|} \leq \|\nabla u_N\| \sup_{g\in\widehat{H^1}} \left(\frac{\inf_v \left\{ \|\nabla(\varphi_g - v)\| \mid v \in V_N \cap \widehat{H^1}, v(\underline{x}_i) = 0 \right\}}{\|g\|} \right).$$

Now let $v_0 \in V_N \cap \widehat{H^1}$ be the best approximant of φ_g. We now define v_1 (for fine enough meshes) such that

$$\begin{aligned} v_1(\underline{x}) &= 0, && \underline{x} \text{ in cells outside the neighbourhood of all } \underline{x}_i \\ v_1(\underline{x}) &= -v_0(\underline{x}), && \underline{x} \text{ in cells containing an } \underline{x}_i \end{aligned}$$

and set $v := v_0 + v_1$, it follows directly that $v(\underline{x}_i) = 0\,\forall i$. In this case

$$\|\nabla(\varphi_g - v)\| \leq \|\nabla(\varphi_g - v_0)\| + \|\nabla v_1\|.$$

Note that $\varphi_g \in H^2$ due to the regularity of the Neuman boundary value problem on the cube [3, 10]. Standard approximation theory (e.g., section 4.4. in [3]) now gives

$$\|\nabla(\varphi_g - v_0)\| \leq Ch|\varphi_g|_{H^2} \leq Ch\|g\|.$$

For the second expression it follows after (12) in the proof of Proposition 4 that

$$\|\nabla v_1\| \leq Ch^{d/2-1}\|g\|.$$

Putting these together gives for $d \geq 4$

$$\|\nabla(\varphi_g - v)\| \leq Ch\|g\|$$

and therefore

$$\|u_N\| \leq Ch\|\nabla u_N\|$$

which is the desired result. □

Lemma 4. *For f_N as above it holds for $d \geq 4$ that $\|f_N\|_\infty$ is uniformly bounded in h.*

Proof.

$$\begin{aligned} \left\| f_N - \int f_N \right\|_\infty &\leq Ch^{-d/2} \left\| f_N - \int f_N \right\| && \text{inverse inequality, e.g Chapter 4 in [3]} \\ &\leq Ch^{-d/2+1}\|\nabla f_N\| && \text{Lemma 3} \\ &\leq Ch^{-d/2+1}h^{d/2-1} \leq C && \text{Proposition 4} \end{aligned}$$

The bound follows since

$$\begin{aligned}|f_N(x)| &\leq |f_N(x_i)| + \left|f_N(x_i) - \int f_N(x)\right| + \left|f_N(x) - \int f_N(x)\right| \\ &\leq Cm\lambda h^{d/2-1} + 2C \\ &\leq Cm\lambda + 2C \leq C'\end{aligned}$$

□

4 Projections and the combination technique

In the previous section we have seen that due to the discrete data term the usual convergence behaviour for the numerical solution of partial differential equations does not exist in the case of regression. Therefore the assumed error expansion (5) for the approximation property of the combination technique [11] does not hold.

The combination technique was studied as a sum of projections into partial spaces [14]. We reproduce in the following some of the work presented there.

It is well known that finite element solutions of V-elliptic problems can be viewed as Ritz projections of the exact solution into the finite element space satisfying the following Galerkin equations:

$$\langle f_N, g\rangle_{RLS} = \langle f_*, g\rangle_{RLS}, \quad g \in V_N.$$

The projections are orthogonal with respect to the energy norm $\|\cdot\|_{RLS}$. Let $P_{\underline{l}} : V \to V_{\underline{l}}$ denote the orthogonal projection with respect to the norm $\|\cdot\|_{RLS}$ and let P_n^S be the orthogonal projection into the sparse grid space $V_n^S = \sum_{|\underline{l}|\leq n} V_{\underline{l}}$. If the projections $P_{\underline{l}}$ form a commutative semigroup, i.e., if for all $\underline{l}, \underline{l}'$ there exists a $\underline{l}''$ such that $P_{\underline{l}}P_{\underline{l}'} = P_{\underline{l}''}$ then there exist $c_{\underline{l}}$ such that

$$P_n^S = \sum_{|\underline{l}|\leq n} c_{\underline{l}} P_{\underline{l}}.$$

We have seen in the previous section why the combination technique may not provide good approximations as the quadrature errors do not cancel in the same way as the approximation errors. The aspect considered here is that the combination technique may break down if there are angles between spaces which are sufficiently smaller than $\pi/2$ and for which the commutator may not be small.

In numerical experiments we estimated the angle of two spaces under the scalar product $\langle \cdot, \cdot \rangle_{RLS}$ from the regularised regression setting [14]. In a first example, the data points are chosen to be the four corners of the square $\Omega = [0,1]^2$. In this case, the angle turns out to be between 89.6 and 90 degrees. Lower angles corresponded to higher values of λ. In the case of $\lambda = 0$ one has the interpolation problem at the corners. These interpolation operators, however, do commute. In this case the penalty term is actually the only source of non-orthogonality. A very similar picture evolves if one chooses the four data points from $\{0.25, 0.75\}^2$. The angle is now between 89 and 90 degrees where the higher angles are now obtained for larger λ and so the regulariser improves the orthogonality.

A very different picture emerges for the case of four randomly chosen points. In our experiments we now observe angles between 45 degrees and 90 degrees and the larger angles are obtained for the case of large λ. Again, the regulariser makes the problem more orthogonal. We would thus expect that for a general fitting problem a choice of larger α would lead to higher accuracy (in regard to the sparse grid solution) of the combination technique. In all cases mentioned above the angles decrease when smaller mesh sizes h are considered. This gives another explanation for the divergence behaviour of the combination technique observed in Figure 4.

4.1 Optimised combination technique

In [13] a modification of the combination technique is introduced where the combination coefficients not only depend on the spaces as before, which gives a linear approximation method, but instead depend on the function to be reconstructed as well, resulting in a non-linear approximation approach. In [14] this ansatz is presented in more detail and the name "opticom" for this optimised combination technique was suggested there.

Assume in the following that the generating subspaces of the sparse grid are suitably numbered from 1 to s. To compute the optimal combination coefficients c_i one minimises the functional

$$\theta(c_1, \ldots, c_s) = |Pf - \sum_{i=1}^{s} c_i P_i f|^2_{RLS},$$

where one uses the scalar product corresponding to the variational problem $\langle\cdot,\cdot\rangle_{RLS}$, defined on V to generate a norm. By simple expansion and by differentiating with respect to the combination coefficients c_i and setting each of these derivatives to zero we see that minimising this norm corresponds to finding c_i which have to satisfy

$$\begin{bmatrix} \|P_1 f\|^2_{RLS} & \cdots & \langle P_1 f, P_s f\rangle_{RLS} \\ \langle P_2 f, P_1 f\rangle_{RLS} & \cdots & \langle P_2 f, P_s f\rangle_{RLS} \\ \vdots & \ddots & \vdots \\ \langle P_s f, P_1 f\rangle_{RLS} & \cdots & \|P_s f\|^2_{RLS} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \|P_1 f\|^2_{RLS} \\ \|P_2 f\|^2_{RLS} \\ \vdots \\ \|P_m f\|^2_{RLS} \end{bmatrix}. \tag{16}$$

The solution of this small system creates little overhead. However, in general an increase in computational complexity is due to the need for the determination of the scalar products $\langle P_i f, P_j f\rangle_{RLS}$. Their computation is often difficult as it requires an embedding into a bigger discrete space which contains both V_i and V_j. Note that in the case of regularised regression the computation of the scalar product can be achieved efficiently [6].

Since $P_l f$ is a Galerkin-solution in V_l it follows that

$$\|P_l f\|^2_{RLS} = \frac{1}{m}\sum_{i=1}^{m}(P_l f(\underline{x}_i), P_l f(\underline{x}_i)) + \lambda\langle\nabla f_l, \nabla P_l f\rangle_2 = \frac{1}{m}\sum_{i=1}^{m} P_l f(\underline{x}_i) y_i.$$

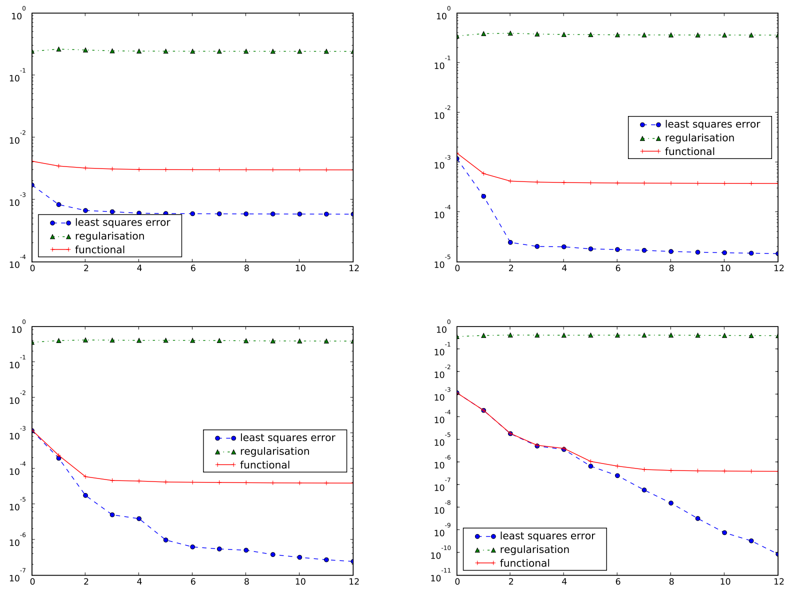
We therefore can interpret the optimised combination technique (i.e., the sum of projections into the partial spaces with the opticom coefficients) as a Galerkin formulation which uses the partial solution $P_l \hat{f}$ as ansatz functions. This way one can formulate an optimised combination technique for problems where the projection arguments do not hold and are replaced by Galerkin conditions. This happens for example for eigenvalue problems [7].

Using these optimal coefficients c_i the combination formula is now

$$f_n^c(\underline{x}) := \sum_{q=0}^{d-1} \sum_{|\underline{l}|_1 = n-q} c_{\underline{l}} f_{\underline{l}}(\underline{x}). \tag{17}$$

In Figure 6 we give results for the optimised combination technique on the data used in Figure 4 for the standard combination technique. The rise of the residual observed before is not present anymore since the optimised combination technique repairs the instabilities. The graph for the smallest regularisation parameter suggests potential overfitting, the error in regard to

Figure 6: Value of the least squares error, the regularisation term and the residual (1) for 1000 data using the optimised combination technique. Top Left: with $\lambda = 10^{-2}$, Top Right: with $\lambda = 10^{-3}$. Bottom Left: with $\lambda = 10^{-4}$, Bottom Right: with $\lambda = 10^{-6}$.



the given data always gets smaller with finer discretiations. Using larger regularisation parameters the influence of the smoothing term prevents results too close to the training data.

An empirical comparison of the optimised combination technique with several regression methods on typical benchmark data shows an excellent performance of the procedure [6].

5 Conclusions

Here we consider a generalisation of the usual kernel methods used in machine learning as the "kernels" of the technique considered here have singularities on the diagonal. However, only finite dimensional approximations are considered. The overfitting effect which occurs for fine grid sizes is investigated. We found that the method (using the norm of the gradient as a penalty) did asymptotically (in grid size) overfit the data but did this very locally only close to the data points. It appeared that the information in the data was concentrated on the data points and only the null space of the penalty operator (in this case constants) was fitted for fine grids. Except for the overfitting in the data points one thus has the same effect as when choosing very large regularisation parameters so that the overfitting in the data points does arise together with an "underfitting" in other points away from the data. Alternatively, one could say that the regularisation technique acts like a parametric fit away from the data points for small grid sizes and overall for large regularisation parameters. A different view is regarding the discretisation as a regularisation [15], so to avoid over-fitting the grid size needs to be limited.

The effect of the data samples is akin to a quadrature method if there are enough data points per element. In practise, it was seen that one required at least one data point per element to get reasonable performance. In order to understand the fitting behaviour we analysed the performance both on the data points and in terms of the Sobolev norm. The results do not directly carry over to results about errors in the sup norm which is often of interest for applications. However, the advice to have at least one data point per element is equally good advice for practical computations. In addition, the insight that the classical combination technique amplifies the sampling errors and thus needs to be replaced by an optimal procedure is also relevant to the case of the sup norm.

The method considered here is in principle a "kernel method" [16] when combined with a finite dimensional space. However, the arising kernel matrix does have diagonal elements which are very large for small grids and, in the limit is a Green's function with a singularity along the diagonal. It is well known in the machine learning literature that kernels with large diagonal elements lead to overfitting, however, the case of families of kernels which approximate a singular kernel is new.

References

[1] D. Braess. *Finite elements.* Cambridge University Press, Cambridge, second edition, 2001.

[2] J. H. Bramble and J. Xu. Some estimates for a weighted L^2 projection. *Math. Comp.*, 56(194):463–476, 1991.

[3] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[4] H.-J. Bungartz, M. Griebel, and U. Rüde. Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems. *Comput. Methods Appl. Mech. Eng.*, 116:243–252, 1994.

[5] J. Garcke. *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern*. Doktorarbeit, Institut für Numerische Simulation, Universität Bonn, 2004.

[6] J. Garcke. Regression with the optimised combination technique. In W. Cohen and A. Moore, editors, *Proceedings of the 23rd ICML '06*, pages 321–328, New York, NY, USA, 2006. ACM Press.

[7] J. Garcke. An optimised sparse grid combination technique for eigenproblems. In *Proceedings of ICIAM 2007*, volume 7 of *PAMM*, pages 1022301–1022302, 2008.

[8] J. Garcke and M. Griebel. Classification with sparse grids using simplicial basis functions. *Intelligent Data Analysis*, 6(6):483–502, 2002. (shortened version appeared in KDD 2001, Proc. of the Seventh ACM SIGKDD, F. Provost and R. Srikant (eds.), pages 87-96, ACM, 2001).

[9] J. Garcke, M. Griebel, and M. Thess. Data mining with sparse grids. *Computing*, 67(3):225–253, 2001.

[10] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001.

[11] M. Griebel, M. Schneider, and C. Zenger. A combination technique for the solution of sparse grid problems. In P. de Groen and R. Beauwens, editors, *Iterative Methods in Linear Algebra*, pages 263–281. IMACS, Elsevier, North Holland, 1992.

[12] W. Hackbusch. *Elliptic differential equations*, volume 18 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1992.

[13] M. Hegland. Additive sparse grid fitting. In *Proceedings of the Fifth International Conference on Curves and Surfaces, Saint-Malo, France 2002*, pages 209–218. Nashboro Press, 2003.

[14] M. Hegland, J. Garcke, and V. Challis. The combination technique and some generalisations. *Linear Algebra and its Applications*, 420(2–3):249–275, 2007.

[15] F. Natterer. Regularisierung schlecht gestellter Probleme durch Projektionsverfahren. *Numer. Math.*, 28:329–341, 1977.

[16] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.

[17] A. N. Tikhonov and V. A. Arsenin. *Solutions of ill-posed problems.* W.H. Winston, Washington D.C., 1977.

[18] G. Wahba. *Spline models for observational data*, volume 59 of *Series in Applied Mathematics.* SIAM, Philadelphia, 1990.