1

Fitting multidimensional data using gradient penalties and combination techniques

Jochen Garcke and Markus Hegland

Mathematical Sciences Institute
Australian National University
Canberra ACT 0200
jochen.garcke,markus.hegland@anu.edu.au

Summary. 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 allows the approximation of the sparse grid fit with a linear combination of fits 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. The application of modified "optimal" combination coefficients provides an advantage over the ones used originally for the numerical solution of PDEs, who in this case simply amplify the sampling noise. As part of this investigation we also show how overfitting arises when the mesh size goes to zero.

1.1 Introduction

In this paper we consider the regression problem arising in machine learning. A set of data points $\underline{x}_i$ in a d-dimensional feature space is given, together with an associated value y_i. We assume that a function f_* describes the relation between the predictor variables $\underline{x}$ and the response variable y and want to (approximately) reconstruct the function f_* from the given data. This allows us to predict the function value of any newly given data point for future decision-making.

In [4] a discretisation approach to the regularised least squares ansatz [10] was introduced. An independent grid with associated local basis functions is used to discretise the minimisation problem. This way the data information is transferred into the discrete function space defined by the grid and its corresponding basis functions. Such a discretisation approach is similar to the numerical treatment of partial differential equations by finite element methods.

To cope with the complexity of grid-based discretisation methods in higher dimensions Garcke et.al. [4] apply the sparse grid combination technique [5] to the regression problem. 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 employed grids. The curse of dimensionality for conventional "full" grid methods affects the sparse grid combination technique much less; currently up to around 20 dimensions can be handled.

Following empirical results in [3], 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. The convergence behaviour of the combination technique can be analysed using extrapolation arguments where a certain error expansion for the partial solutions is assumed. Alternatively one can view the combination technique as approximation of a projection into the underlying sparse grid space, which is exact only if the partial projections commute.

We will study how both these assumptions do not hold for the regularised regression problem and how the combination technique can actually diverge. Applying the optimised combination technique, introduced in [7], repairs the resulting 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.

1.2 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 with $\underline{x}$ 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 [9, 10] 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

$$R(f) = \frac{1}{m}\sum_{i=1}^{m}(f(\underline{x}_i) - y_i)^2 + \lambda||\nabla f||^2, \tag{1.1}$$

where $y_i = f_*(\underline{x}_i)$. The first term in (1.1) 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

$$\langle f, g \rangle_{RLS} = \frac{1}{m} \sum_{i=1}^{m} f(\underline{x}_i) g(\underline{x}_i) + \lambda \langle \nabla f, \nabla g \rangle \tag{1.2}$$

and choose V so that $\langle \cdot, \cdot \rangle_{RLS}$ is a scalar product on it. With respect to this scalar product the minimisation (1.1) is an orthogonal projection of f_* into V [7], i.e. if $\|f - f_*\|^2_{RLS} \leq \|g - f_*\|^2_{RLS}$ than $R(f) \leq R(g)$. As the point evaluations $f \rightarrow f(\underline{x})$ are not continuous in the Sobolev space H^1 for $d \geq 2$ we do not get a H^1-elliptic problem. We 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

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

We now plug the representation (1.2) of a function $f \in V_N$ into (1.1) and obtain the linear equation system

$$(\mathcal{B}^\top \mathcal{B} + \lambda m \cdot \mathcal{C})\alpha = \mathcal{B}^\top y \tag{1.3}$$

and therefore are able to compute the unknown vector α for the solution f_N of (1.1) in V_N. $\mathcal{C}$ is a symmetric $N \times N$ matrix with entries $\mathcal{C}_{j,k} = \langle \nabla \varphi_j, \nabla \varphi_k \rangle$, $j, k = 1, \ldots, N$ and corresponds to the smoothness operator. $\mathcal{B}^\top$ is a rectangular $m \times N$ matrix with entries $(\mathcal{B}^\top)_{j,k} = \varphi_j(\underline{x}_k)$, $j = 1, \ldots, N$, $k = 1, \ldots, m$ and transfers the information from the data into the discrete space, $\mathcal{B}$ correspondingly works in the opposite direction. The vector 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 $h_t = 2^{-l_t}, t = 1, \ldots, d$ the grid points are numbered 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.

1.3 Combination technique

The sparse grid combination technique [5] is an approach to approximate functions in higher dimensional spaces. Following this ansatz we discretise and solve the problem (1.1) on a sequence of small anisotropic grids $\Omega_{\underline{l}} = \Omega_{l_1,...,l_d}$. 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 (1.3). The original combination technique [5] 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

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

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.

Using extrapolation arguments it can be shown that the approximation property of the combination technique is of the order $O(h_n^2 \cdot \log(h_n^{-1})^{d-1})$ as long as error expansions of the form

$$f - f_{\underline{l}} = \sum_{i=1}^{d} \sum_{j_1,...,j_m \subset 1,...,d} c_{j_1,...,j_m}(h_{j_1}, \ldots, h_{j_m}) \cdot h_{j_1}^p \cdot \ldots \cdot h_{j_m}^p$$

for the partial solutions hold [5].

Viewing the minimisation of (1.1) as projection one can show that the combination technique is an exact projection into the underlying sparse grid space (and therefore of approximation order $O(h_n^2 \cdot \log(h_n^{-1})^{d-1})$) 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 [6].

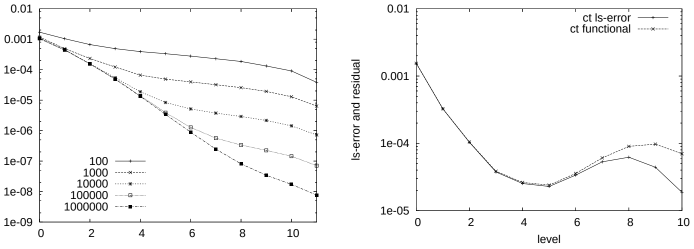
In the following we will show that both these assumptions do not hold for the regularised regression problem and that the combination technique can actually diverge.

1.3.1 Empirical convergence behaviour

We now 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 [3] we consider the function

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

Fig. 1.1. Left: Convergence of full grid solutions against highly refined solution measured using (1.1). Right: Value of the residual (1.1) and the least squares error for 5000 data using the combination technique with $\lambda = 10^{-6}$.



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 1.1, left we show the error of a full grid solution of level l measured against the one of level $n = 12$ using the functional (1.1) as a norm. We see that the error shows two different types of convergence behaviour, after some discretisation level the error decreases slower than with the usual h^2. Furthermore, the more data is used, the later this change in the error reduction rate takes place. These observations do not depend on the regularisation parameter λ.

A different picture arises if we employ the sparse grid combination technique. We measure the residual and the least squares error of the approximation using $m = 5000$ data and $\lambda = 10^{-6}$, the results are presented in Figure 1.1, right. One observes that after level $n = 3$ both error measurements increase on the training data, which cannot happen with a true variational discretisation ansatz. This effect is especially observed for small λ, already with $\lambda = 10^{-4}$ the now stronger influence of the smoothing term results in a (more) stable approximation method. Note that the instability is more common and significant in higher dimensions.

1.3.2 Asymptotics and errors of the full grid solution

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 $p(\underline{x}) = 1$, 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

$$J(f) = \lambda \int_\Omega |\nabla f(\underline{x})|^2 dx + \int_\Omega (f(\underline{x}) - f_*(\underline{x}))^2 dx \tag{1.4}$$

then $J(f) \approx R(f)$. Consider a finite element space $V_N \subset C(\Omega)$ with rectangular elements Q of side lengths $h_1, \ldots, h_d$ 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

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

holds for $k \to \infty$ 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:

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

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

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

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 a sparse grid situation 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

$$\|f - f_{\text{sg}}\|_1 \leq C \left(\|f - f_{\text{sg}}^{\text{best}}\|_1 + \frac{\|f\|_1 \sqrt{\sum_{\underline{l}} c_{\underline{l}}^2 2^{|\underline{l}|}}}{\sqrt{m}} \right) \qquad (1.5)$$

where, as usual, $c_{\underline{l}}$ are the combination coefficients.

To study this effect experimentally let us consider (1.4) with

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

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 (1.1). We measure the difference between the resulting discrete solutions for fixed number of data and the above continuous solution. In Figure 1.2, left, we show how this error, measured in the H^1-seminorm, behaves. At first we have the "usual" decrease in the error, but after about 1 sample point per element the error increases instead.

The bound (1.5) holds only asymptotically in k and thus for fixed k and very small mesh size it will break down. In the following we consider what happens asymptotically in this case for a regular grid ($h_i = h$). Recall that the full grid solution f_N satisfies the Galerkin equations

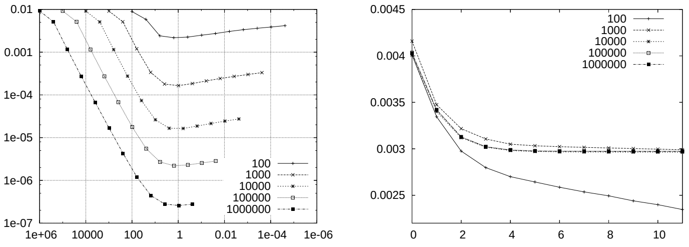
$$\lambda \int_\Omega \nabla f_N^T \nabla g dx + \frac{1}{m}\sum_{i-1}^{m}(f_N(\underline{x}_i) - y_i)g(\underline{x}_i) = 0, \quad \text{for all } g \in V_N. \tag{1.6}$$

Using an approximate Green's kernel $G_N(\underline{x}, \underline{x}_i)$ for the Laplacian one can write the solution as

$$f_N(\underline{x}) = \frac{1}{m\lambda}\sum_{i=1}^{m}(y_i - f_N(\underline{x}_i))G_N(\underline{x}, \underline{x}_i).$$

One can show that for $i \neq j$ the values $G_N(\underline{x}_i, \underline{x}_j)$ are bounded as $h \to 0$ and that $G_N(\underline{x}_i, \underline{x}_i) = O(|\log(h)|)$ for $d = 2$ and $G_N(\underline{x}_i, \underline{x}_i) = O(h^{2-d})$ for $d \neq 2$. One then gets:

Fig. 1.2. 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.



Proposition 1. *Let f_N be the solution of equation 1.6. Then $f_N(\underline{x}_i) \to y_i$ for $h \to 0$ and there exists a $C > 0$ such that*

$$|y_i - f_N(\underline{x}_i)| \leq C \frac{m\lambda}{|\log h|}, \quad \textit{if } d = 2$$

and

$$|y_i - f_N(\underline{x}_i)| \leq C m\lambda h^{d-2}, \quad \textit{if } d > 2.$$

Proof. Using the Green's kernel matrix G_N with components $G_N(\underline{x}_i, \underline{x}_j)$ one has for the vector f_N of function values $f_N(\underline{x}_i)$ the system

$$(G_N + m\lambda I) f_N = G_N y$$

where y is the data vector with components y_i.

It follows that $f_N - y = (G_N + m\lambda I)^{-1} G_N y - (G_N + m\lambda I)^{-1}(G_N + m\lambda I) = -m\lambda(G_N + m\lambda I)^{-1} y$. The bounds for the distance between the function values and the data then follow when the asymptotic behaviour of G_N mentioned above is taken into account. □

It follows that one gets an asymptotic overfitting in the data points and the data term in $R(f)$ satisfies the same bound

$$\sum_{i=1}^{m} (f_N(\underline{x}_i) - y_i)^2 \leq C \frac{m\lambda}{|\log h|}, \quad \text{if } d = 2$$

and

$$\sum_{i=1}^{m} (f_N(\underline{x}_i) - y_i)^2 \leq C m\lambda h^{d-2}, \quad \text{if } d \geq 3$$

and $h \to 0$. The case $d = 2$ is illustrated in Figure 1.2, right.

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 the fit picks up fast oscillations near the data points but only slow variations further away.

It is seen that the value of $R(f_N) \to 0$ for $h \to 0$. In the following we can give a bound for this for $d \geq 3$.

Proposition 2. *The value of functional J converges to zero on the estimator f_N and*

$$J(f_N) \leq C m\lambda h^{d-2}$$

for some $C > 0$. In particular, one has $\|\nabla f_N\| \leq C\sqrt{m\lambda h^{(d-2)}}$.

Proof. While we only consider regular partitioning with hyper-cubical elements Q, the proof can be generalised for 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 see that

$$\int_Q |\nabla b_Q|^2 dx \le C h^{d-2}.$$

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

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

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 $i \neq j$ and so

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

and, thus,

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

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

We conjecture that in the case of $d = 2$ one has $J(f_N) \le C m \lambda / |\log h|$. We would also conjecture, based on the observations, that f_N converges very slowly towards a constant function.

1.4 Projections and the combination technique

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 $V_n^S = \sum_{|\underline{l}| \le n} V_{\underline{l}}$. 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.

For illustration, consider the case of three spaces V_1, V_2 and $V_3 = V_1 \cap V_2$. The cosine of the angle $\alpha(V_1, V_2) \in [0, \pi/2]$ between the two spaces V_1 and V_2 is defined as

$$c(V_1, V_2) := \sup \left\{ (f_1, f_2) \mid f_i \in V_i \cap (V_1 \cap V_2)^\perp, \|f_i\| \leq 1, i = 1, 2 \right\}.$$

The angle can be characterised in terms of the orthogonal projections P_{V_i} into the closed subspaces V_i and the corresponding operator norm, it holds [2]

$$c(V_1, V_2) = \|P_1 P_2 P_{V_3^\perp}\|. \tag{1.7}$$

If the projections commute then one has $c(V_1, V_2) = 0$ and $\alpha(V_1, V_2) = \pi/2$ which in particular is the case for orthogonal V_i. However, one also gets $\alpha(V_1, V_2) = \pi/2$ for the case where $V_2 \subset V_1$ (which might contrary to the notion of an “angle”).

Numerically, we estimate the angle of two spaces using a Monte Carlo approach and the definition of the matrix norm as one has

$$c(V_1, V_2) = \|P_{V_1} P_{V_2} - P_{V_1 \cap V_2}\| = \sup_g \frac{\|P_{V_1} P_{V_2} g - P_{V_1 \cap V_2} g\|}{\|P_{V_2} g\|} \tag{1.8}$$

For the energy norm the angle between the spaces substantially depends on the positions of the data points $\underline{x}_i$. We consider in the following several different layouts of points and choose the function values y_i randomly. Then the ratio $\frac{\|P_{V_1} P_{V_2} g - P_{V_1 \cap V_2} g\|}{\|P_{V_2} g\|}$ is determined for these function values and data points and the experiment is repeated many times. The estimate chosen is then the maximal quotient.

In the experiments we choose $\Omega = (0, 1)^2$ and the subspaces V_1 and V_2 were chosen such that the functions were linear with respect to one variable while the h for the grid in the other variables was varied. In a first example, the data points are chosen to be the four corners of the square Ω. In this case, the angle turns out to be between 89.6 and 90 degrees. Lower angles corresponded here 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. Thus the regularise again does make 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. A very similar picture was seen if the points were chosen as the elements of the set $0.2i(1,1)$ for $i = 1, \ldots, 4$. In all cases mentioned above the angles decrease when smaller mesh sizes h are considered.

1.4.1 Optimised combination technique

In [7] 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 [6] this ansatz is presented in more detail and the name "opticom" for this optimised combination technique is suggested.

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 one gets

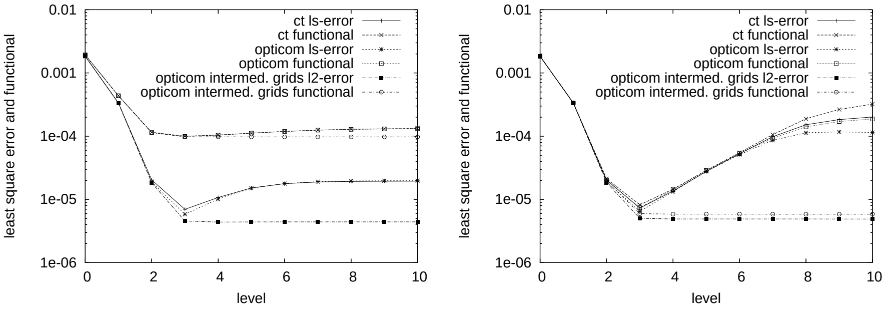
$$\begin{aligned} \theta(c_1, \ldots, c_s) = & \sum_{i,j=1}^{s} c_i c_j \langle P_i f, P_j f \rangle_{RLS} \\ & -2 \sum_{i=1}^{s} c_i \|P_i f\|^2_{RLS} + \|Pf\|^2_{RLS}. \end{aligned}$$

While this functional depends on the unknown quantity Pf, the location of the minimum of J does not. 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

$$\begin{bmatrix} \|P_1 f\|^2_{RLS} & \cdots & \langle P_1 f, P_s f \rangle_{RLS} \\ \langle P_2 f, P_1 f \rangle_{RLS} & \cdots & \langle P_2 f, P_s f \rangle_{RLS} \\ \vdots & \ddots & \vdots \\ \langle P_s f, P_1 f \rangle_{RLS} & \cdots & \|P_s f\|^2_{RLS} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \|P_1 f\|^2_{RLS} \\ \|P_2 f\|^2_{RLS} \\ \vdots \\ \|P_m f\|^2_{RLS} \end{bmatrix} \tag{1.9}$$

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

Fig. 1.3. Value of the functional (1.1) and the least squares error on the data, i.e. $\frac{1}{M}\sum_{i=1}^{M}(f(\underline{x}_i)-y_i)^2$, for the reconstruction of $e^{-x^2}+e^{-y^2}$ for the combination technique and the optimised combination technique for the grids $\Omega_{i,0}, \Omega_{0,i}, \Omega_{0,0}$ and the optimised combination technique for the grids $\Omega_{j,0}, \Omega_{0,j}, 0 \leq j \leq i$ with $\lambda = 10^{-4}$ (left) and 10^{-6} (right).



as it requires an embedding into a bigger discrete space which contains both V_i and V_j.

Using these optimal coefficients c_i the combination formula is now

$$f_n^c(\underline{x}) := \sum_{q=0}^{d-1} \sum_{|\underline{l}|_1=n-q} c_{\underline{l}} f_{\underline{l}}(\underline{x}). \tag{1.10}$$

Now let us consider one particular additive function $u = e^{-x^2} + e^{-y^2}$, which we want to reconstruct based on 5000 random data samples in the domain $[0,1]^2$. We use the combination technique and optimised combination technique for the grids $\Omega_{i,0}, \Omega_{0,i}, \Omega_{0,0}$. For $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$ we show in Figure 1.3 the value of the functional (1.1), in Table 1.1 the corresponding numbers for the residuals and the cosine of $\gamma = \angle(P_{U_1}u, P_{U_2}u)$ are given. We see that both methods diverge for higher levels of the employed grids, nevertheless as expected the optimised combination technique is always better than the normal one.

We also show in Figure 1.3 the results for an optimised combination technique which involves all intermediate grids, i.e. $\Omega_{j,0}, \Omega_{0,j}$ for $1 \leq j < i$, as well. Here we do not observe rising values of the functional for higher levels but a saturation, i.e. higher refinement levels do not substantially change the value of the functional.

1.5 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.

level	$cos(\gamma)$	e_c^2	e_o^2
1	-0.012924	$3.353704 \cdot 10^{-4}$	$3.351200 \cdot 10^{-4}$
2	-0.025850	$2.124744 \cdot 10^{-5}$	$2.003528 \cdot 10^{-5}$
3	-0.021397	$8.209228 \cdot 10^{-6}$	$7.372946 \cdot 10^{-6}$
4	-0.012931	$1.451818 \cdot 10^{-5}$	$1.421387 \cdot 10^{-5}$
5	0.003840	$2.873697 \cdot 10^{-5}$	$2.871036 \cdot 10^{-5}$
6	0.032299	$5.479755 \cdot 10^{-5}$	$5.293952 \cdot 10^{-5}$
7	0.086570	$1.058926 \cdot 10^{-4}$	$9.284347 \cdot 10^{-5}$
8	0.168148	$1.882191 \cdot 10^{-4}$	$1.403320 \cdot 10^{-4}$
9	0.237710	$2.646455 \cdot 10^{-4}$	$1.706549 \cdot 10^{-4}$
10	0.285065	$3.209026 \cdot 10^{-4}$	$1.870678 \cdot 10^{-4}$

Table 1.1. Residual for the normal combination technique e_c^2 and the optimised combination technique, as well as cosine of the angle $\gamma = \angle(P_{U_1}u, P_{U_2}u)$.

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 point 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.

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" [8] 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. Dietrich Braess. *Finite elements.* Cambridge University Press, Cambridge, second edition, 2001.
2. Frank Deutsch. Rate of convergence of the method of alternating projections. In *Parametric optimization and approximation (Oberwolfach, 1983)*, volume 72 of *Internat. Schriftenreihe Numer. Math.*, pages 96–107. Birkhäuser, Basel, 1985.
3. J. Garcke. *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern.* Doktorarbeit, Institut für Numerische Simulation, Universität Bonn, 2004.
4. J. Garcke, M. Griebel, and M. Thess. Data mining with sparse grids. *Computing*, 67(3):225–253, 2001.
5. 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.
6. M. Hegland, J. Garcke, and V. Challis. The combination technique and some generalisations. submitted, `http://wwwmaths.anu.edu.au/~garcke/paper/opticom.pdf`, 2005.
7. Markus 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.
8. B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.
9. A. N. Tikhonov and V. A. Arsenin. *Solutions of ill-posed problems.* W.H. Winston, Washington D.C., 1977.
10. G. Wahba. *Spline models for observational data*, volume 59 of *Series in Applied Mathematics.* SIAM, Philadelphia, 1990.