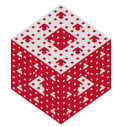
MATH 590: Meshfree Methods

Accuracy and Optimality of RKHS Methods

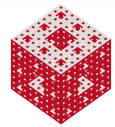
Greg Fasshauer

Department of Applied Mathematics
Illinois Institute of Technology

Fall 2014



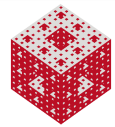
Outline



In Chapter 4 we saw that (simple) kriging yields the best linear unbiased predictor $\overset{\triangle}{Y}_{\boldsymbol{x}}$ based on given data $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, N\}$ since it minimizes the mean square error of any linear predictor based on the data.

In this chapter we will see that within the native Hilbert spaces associated with strictly positive definite kernels (or, more specifically, radial functions) the kernel (RBF) interpolant provides the best approximation to a given data function.

This optimality of interpolants in Hilbert space is the subject of the theory of optimal recovery described in the late 1950s by Michael Golomb and Hans Weinberger in [GW59].



In [GW59] the authors studied the following general problem:

Problem

Given the values $f_1 = \lambda_1(f), \ldots, f_N = \lambda_N(f) \in \mathbb{R}$, where $\{\lambda_1, \ldots, \lambda_N\}$ is a linearly independent set of linear functionals (called information functionals yielding the information about f), how does one "best" approximate the value $\lambda(f)$ (called a feature of f) where λ is a given linear functional and f is unknown?

Remark

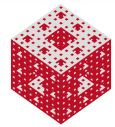
This is a very general problem formulation that includes

- *interpolation of data such as*
  - *function values,*
  - *values of derivatives of f,*
  - *integrals of f (such as averages or moments),*
- *as well as methods of approximation other than interpolation.*

Remark

This kind of problem is known in the literature as an optimal recovery problem.

Besides the seminal work by Golomb and Weinberger, optimal recovery was also studied in detail in [MRW76, MR77, MR80, MR85].



In a Hilbert space setting the solution to this optimal recovery problem is shown to be the minimum-norm interpolant.

More precisely, given a Hilbert space $\mathcal{H}$ and data $f_1 = \lambda_1(f), \ldots, f_N = \lambda_N(f) \in \mathbb{R}$ with $\{\lambda_1, \ldots, \lambda_N\} \subseteq \mathcal{H}^*$ (the dual of $\mathcal{H}$), the minimum-norm interpolant is that function $\overset{\star}{g} \in \mathcal{H}$ which satisfies

$$\lambda_j(\overset{\star}{g}) = f_j, \qquad j = 1, \ldots, N,$$

and

$$\overset{\star}{g} = \underset{\substack{g \in \mathcal{H} \\ \lambda_j(g) = f_j, j = 1, \ldots, N}}{\operatorname{argmin}} \|g\|_{\mathcal{H}}.$$

Remark

The kernel interpolant with kernel K satisfies these criteria if $\mathcal{H}$ is taken as the associated RKHS $\mathcal{H}_K(\Omega)$.

We will present three optimality results:

- The kernel interpolant for any strictly positive definite kernel K is the minimum norm interpolant from $\mathcal{H}_K(\Omega)$.
- The kernel interpolant provides the best approximation to f in the native space norm.
- The (cardinal form of the) kernel interpolant is more accurate (as measured by the pointwise error) than any other linear combination of the data (very similar to BLUP).

Remark

The proofs of the first two "optimality theorems" require the following two lemmas.

Lemma

Assume K is a symmetric strictly positive definite kernel on $\mathbb{R}^d$ and let s be the usual kernel interpolant to $f \in \mathcal{H}_K(\Omega)$ at the data sites $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \Omega$, i.e.,

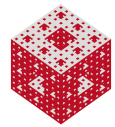
$$s(\boldsymbol{x}) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}, \boldsymbol{x}_j),$$

where the c_j are determined by $s(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$, $i = 1, \ldots, N$. Then

$$\langle s, s - g \rangle_{\mathcal{H}_K(\Omega)} = 0$$

for any other interpolant $g \in \mathcal{H}_K(\Omega)$, i.e., with $g(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$, $i = 1, \ldots, N$.

Note that $g \in \mathcal{H}_K(\Omega)$, a space that is larger than $\text{span}\{K(\cdot, \boldsymbol{x}_j) : \boldsymbol{x}_j \in \mathcal{X}\}$.



Proof.

The interpolant s is of the form $s = \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j)$, where the c_j are determined by $s(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$, $i = 1, \ldots, N$.

Using this, the symmetry of K and its reproducing property we have

$$\begin{aligned}
\langle s, s-g \rangle_{\mathcal{H}_K(\Omega)} &= \langle \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j), s-g \rangle_{\mathcal{H}_K(\Omega)} \\
&= \sum_{j=1}^{N} c_j \langle K(\cdot, \boldsymbol{x}_j), s-g \rangle_{\mathcal{H}_K(\Omega)} \\
&= \sum_{j=1}^{N} c_j \langle s-g, K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)} \\
&= \sum_{j=1}^{N} c_j (s-g)(\boldsymbol{x}_j) = 0
\end{aligned}$$

since both s and g interpolate f on $\mathcal{X}$. □

For the next result, we define

$$H_K(\mathcal{X}) = \text{span}\{K(\cdot, \boldsymbol{x}_j) : \ \boldsymbol{x}_j \in \mathcal{X}\}$$

as we did in Chapter 2, Part 3.
$H_K(\mathcal{X})$ is an N-dimensional subspace of the native space $\mathcal{H}_K(\Omega)$.

Lemma

Assume K is a strictly positive definite kernel on $\mathbb{R}^d$ and let $s \in H_K(\mathcal{X})$ be the interpolant to $f \in \mathcal{H}_K(\Omega)$ on $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \Omega$.
Then

$$\langle f - s, h \rangle_{\mathcal{H}_K(\Omega)} = 0$$

for all $h \in H_K(\mathcal{X})$.

Remark

In other words, the residual $f - s$ is orthogonal to the subspace $H_K(\mathcal{X})$ of $\mathcal{H}_K(\Omega)$. In particular, $\langle f - s, s \rangle_{\mathcal{H}_K(\Omega)} = 0$.

Proof.

Any $h \in H_K(\mathcal{X})$ can be written in the form $h = \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j)$ with appropriate coefficients c_j.
Using this as well as the reproducing property of K we have

$$\begin{aligned}
\langle f - s, h \rangle_{\mathcal{H}_K(\Omega)} &= \langle f - s, \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)} \\
&= \sum_{j=1}^{N} c_j \langle f - s, K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)} \\
&= \sum_{j=1}^{N} c_j (f - s)(\boldsymbol{x}_j) = 0
\end{aligned}$$

since s interpolates f on $\mathcal{X}$. □

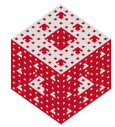
The following Pythagorean theorem (or "energy splitting" theorem) is an immediate consequence of the previous lemma.
It says that the native space "energy" of f can be split into

- the "energy" of the interpolant s and
- the energy of the residual $f - s$ (which — according to our lemma — is orthogonal to the interpolant).

Corollary

The orthogonality property of the preceding lemma implies the energy split

$$\|f\|^2_{\mathcal{H}_K(\Omega)} = \|f - s\|^2_{\mathcal{H}_K(\Omega)} + \|s\|^2_{\mathcal{H}_K(\Omega)}.$$



Proof.

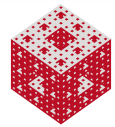
The statement follows from

$$\begin{aligned} \|f\|^2_{\mathcal{H}_K(\Omega)} &= \|f - s + s\|^2_{\mathcal{H}_K(\Omega)} \\ &= \langle (f-s) + s, (f-s) + s \rangle_{\mathcal{H}_K(\Omega)} \\ &= \|f-s\|^2_{\mathcal{H}_K(\Omega)} + 2\langle f - s, s \rangle_{\mathcal{H}_K(\Omega)} + \|s\|^2_{\mathcal{H}_K(\Omega)} \end{aligned}$$

and the fact that $\langle f - s, s \rangle_{\mathcal{H}_K(\Omega)} = 0$ by the lemma since the interpolant s itself is a special linear combination from $H_K(\mathcal{X})$. □

Remark

We will use the energy split in our sampling inequality error estimates below.

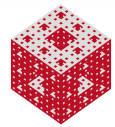


The following theorem presents the first optimality property formulated for strictly positive definite kernels. It is taken from [Wen05].

Theorem (Optimality I)

Suppose $K \in C(\Omega \times \Omega)$ is a strictly positive definite kernel. If the values $f_1, \ldots, f_N$ are given, then the interpolant s is the minimum-norm interpolant to $\{f_j\}_{j=1}^N$, i.e.,

$$s = \operatorname*{argmin}_{\substack{g \in \mathcal{H}_K(\Omega) \\ g(\boldsymbol{x}_j) = f_j, j=1,\ldots,N}} \|g\|_{\mathcal{H}_K(\Omega)}.$$



Proof.

Let $g \in \mathcal{H}_K(\Omega)$ be an arbitrary interpolant to $f_1, \ldots, f_N$.
Then the first orthogonality lemma tells us

$$\langle s, s - g \rangle_{\mathcal{H}_K(\Omega)} = 0.$$

Now

$$\begin{aligned} \|s\|^2_{\mathcal{H}_K(\Omega)} &= \langle s, s - g + g \rangle_{\mathcal{H}_K(\Omega)} \\ &= \langle s, s - g \rangle_{\mathcal{H}_K(\Omega)} + \langle s, g \rangle_{\mathcal{H}_K(\Omega)} \\ &= \langle s, g \rangle_{\mathcal{H}_K(\Omega)}, \end{aligned}$$

where the last step follows from the above orthogonality relation.
The Cauchy-Schwarz inequality yields

$$\|s\|^2_{\mathcal{H}_K(\Omega)} \leq \|s\|_{\mathcal{H}_K(\Omega)} \|g\|_{\mathcal{H}_K(\Omega)},$$

so that the statement follows. □

Example

The native space of (conditionally positive definite) thin plate splines

$$\kappa(r) = r^2 \log r,$$

$r = \|\boldsymbol{x}\|_2$ with $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$ is given by the Beppo-Levi space $\mathrm{BL}_2(\mathbb{R}^2)$.

The corresponding semi-norm in the Beppo-Levi space $\mathrm{BL}_2(\mathbb{R}^2)$ is (see [Fas07, Chapter 13])

$$|f|^2_{\mathrm{BL}_2(\mathbb{R}^2)} = \int_{\mathbb{R}^2} \left(\left| \frac{\partial^2 f}{\partial x^2}(\boldsymbol{x}) \right|^2 + 2 \left| \frac{\partial^2 f}{\partial x \partial y}(\boldsymbol{x}) \right|^2 + \left| \frac{\partial^2 f}{\partial y^2}(\boldsymbol{x}) \right|^2 \right) \mathrm{d}\boldsymbol{x},$$

which represents the bending energy of a thin plate.

By the optimality theorem the thin plate spline interpolant minimizes this energy. This explains the name of these functions.

Another nice property of the kernel interpolant is the fact that it is at the same time the best Hilbert-space approximation to the given data. Therefore the interpolant is not just any projection of f but the orthogonal projection (with respect to the native space inner product).

Theorem (Optimality II)

Let

$$H_K(\mathcal{X}) = \{h = \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j) : \ \boldsymbol{x}_j \in \mathcal{X}\},$$

where $K \in C(\Omega \times \Omega)$ is a strictly positive definite kernel.
If only the values $f_1 = f(\boldsymbol{x}_1), \ldots, f_N = f(\boldsymbol{x}_N)$ are given, then the interpolant s is the best approximation to f from $H_K(\mathcal{X})$ in $\mathcal{H}_K(\Omega)$, i.e.,

$$\|f - s\|_{\mathcal{H}_K(\Omega)} \leq \|f - h\|_{\mathcal{H}_K(\Omega)}$$

for all $h \in H_K(\mathcal{X})$.

Proof.

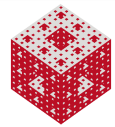
The native space $\mathcal{H}_K(\Omega)$ is the completion of $H_K(\Omega)$ with respect to the $\|\cdot\|_{H_K}$-norm (see Chapter 2) so that

$$\|f\|_{H_K(\Omega)} = \|f\|_{\mathcal{H}_K(\Omega)} \qquad \text{for all } f \in H_K(\Omega).$$

Therefore, we can characterize the best approximation $\overset{\star}{g}$ to f from $H_K(\mathcal{X})$ by

$$\langle f - \overset{\star}{g}, h \rangle_{\mathcal{H}_K(\Omega)} = 0 \qquad \text{for all } h \in H_K(\mathcal{X}).$$

However, the second orthogonality lemma shows that $\overset{\star}{g} = s$ satisfies this relation (and s qualifies since $\mathcal{X} \subseteq \Omega$). □



Remark

Optimality properties of kernel interpolants play an important role in applications such as

- *in the design of support vector machines in statistical learning theory*
- *or the numerical solution of partial differential equations.*

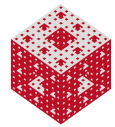
The optimality results imply that one could also start with some Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$ and ask to find the minimum norm interpolant (i.e., Hilbert space best approximation) to some given data.

In this way any given space defines a set of optimal basis functions, generated by the reproducing kernel of $\mathcal{H}$.

This is how Duchon approached the subject in his papers [Duc76, Duc77, Duc78, Duc80].

More recently, Kybic, Blu and Unser [KBU02a, KBU02b] take this point of view and explain from a sampling theory point of view how a thin plate spline can be interpreted as fundamental solution of the differential operator defining the semi-norm in the Beppo-Levi space $\mathrm{BL}_2(\mathbb{R}^2)$.

These contributions motivated our work on Green's kernels (see Chapter 6).



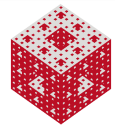
The third optimality result is in the context of quasi-interpolation, i.e.,

Theorem (Optimality III)

Suppose $K \in C(\Omega \times \Omega)$ is a strictly positive definite kernel and suppose that $\boldsymbol{x} \in \Omega$ is fixed.
Let $\overset{\star}{u}_j(\boldsymbol{x})$, $j = 1, \ldots, N$, be the values at $\boldsymbol{x}$ of the cardinal basis functions for interpolation with K.
Then

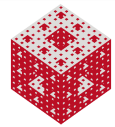
$$\left| f(\boldsymbol{x}) - \sum_{j=1}^{N} f(\boldsymbol{x}_j) \overset{\star}{u}_j(\boldsymbol{x}) \right| \leq \left| f(\boldsymbol{x}) - \sum_{j=1}^{N} f(\boldsymbol{x}_j) u_j \right|$$

for all choices of $u_1, \ldots, u_N \in \mathbb{R}$.



Remark

- *The theorem is proved in [Wen05].*
- *It says that the minimum norm interpolant s is also more accurate (in the pointwise sense) than any linear combination of the given data values.*
- *This is very similar to the MSE-optimality of kriging.*



Goal: to provide error estimates for scattered data interpolation with strictly positive definite functions.

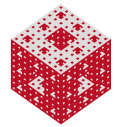
We will provide most of the details for the strictly positive definite case, but extension to conditionally positive definite kernels are possible.

In their final form we will want our estimates to depend on some kind of measure of the data distribution.

The measure that is usually used in approximation theory is the so-called fill distance

$$h = h_{\mathcal{X},\Omega} = \sup_{\boldsymbol{x}\in\Omega} \min_{\boldsymbol{x}_j\in\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}_j\|_2$$

already introduced in Chapter 1.



- The fill distance indicates how well the data fill out the domain Ω.
- It denotes the radius of the largest empty ball that can be placed among the data locations.

Convergence

We will be interested in whether the error

$$\|f - s^{(h)}\|_\infty$$

tends to zero as $h \to 0$, and if so, how fast.

Here $\{s^{(h)}\}_h$ presents a sequence of interpolation (or, more generally, projection) operators that vary with the fill distance h.

Remark

Most error bounds will focus on this worst-case setting. Some will be measured in the L_2-norm, i.e., for average case errors, or other L_p-norms.

Example

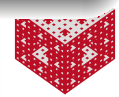
Let $s^{(h)}$ denote the interpolant to data given

- at $(2^n+1)^d$, $n = 1, 2, \ldots$, equally spaced points in the unit cube in $\mathbb{R}^d$
- so that $h = \frac{1}{\sqrt[d]{(2^n+1)^d - 1}} = 2^{-n}$.

The definition of the fill distance also covers scattered data such as sets of Halton points.
In fact, since Halton points are quasi-uniformly distributed we can assume $h \approx 2^{-n}$ for a set of $(2^n+1)^d$ Halton points in $\mathbb{R}^d$.

Remark

These relations explain the specific sizes of the point sets we used in earlier examples.



We measure the speed of convergence to zero in terms of approximation order.

We say that the interpolant $s^{(h)}$ has L_p-approximation order k if

$$\|f - s^{(h)}\|_p = \mathcal{O}(h^k) \quad \text{for } h \to 0.$$

Moreover, if we can also show that $\|f - s^{(h)}\|_p \neq o(h^k)$, then $s^{(h)}$ has exact L_p-approximation order k.

Remark

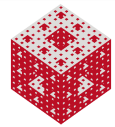
We will concentrate mostly on the case $p = \infty$ (i.e., pointwise estimates), but approximation order in other norms can also be studied.

Remark

Since we want to employ the machinery of reproducing kernel Hilbert spaces presented in the previous chapter we will concentrate on error estimates for functions $f \in \mathcal{H}_K$.

In order to keep the following discussion as transparent as possible we will restrict ourselves to strictly positive definite kernels.

With (considerably) more technical details the following can also be formulated for strictly conditionally positive definite kernels (see [Wen05] for details).



Lagrange Form of the Interpolant and Cardinal Basis Functions

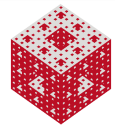
The key idea for the following discussion is to express the interpolant in Lagrange form, i.e., using so-called cardinal basis functions.

For radial basis function approximation this idea is due to [WS93].

In earlier chapters we established that, for any strictly positive definite kernel K, the linear system

$$\mathrm{K}\boldsymbol{c} = \boldsymbol{y}$$

with $\mathrm{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, N$, $\boldsymbol{c} = (c_1, \ldots, c_N)^T$, and $\boldsymbol{y} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N))^T$ has a unique solution.



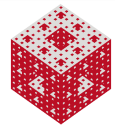
In order to obtain the cardinal basis functions $\overset{\star}{u}_j$, $j = 1, \ldots, N$, with the property $\overset{\star}{u}_j(\boldsymbol{x}_i) = \delta_{ij}$, i.e.,

$$\overset{\star}{u}_j(\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

we consider the linear system

$$\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x}), \tag{1}$$

where the matrix K is as above (and therefore invertible), $\overset{\star}{\boldsymbol{u}} = (\overset{\star}{u}_1, \ldots, \overset{\star}{u}_N)^T$, and $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_N))^T$.



Existence of Cardinal Functions

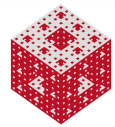
Theorem

Suppose K is a strictly positive definite kernel on $\mathbb{R}^d \times \mathbb{R}^d$. Then, for any distinct points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, there exist functions $\overset{\star}{u}_j \in \text{span}\{K(\cdot, \boldsymbol{x}_j), j = 1, \ldots, N\}$ such that $\overset{\star}{u}_j(\boldsymbol{x}_i) = \delta_{ij}$.
They are determined pointwise by solving the linear system (1), i.e.,

$$\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x}).$$

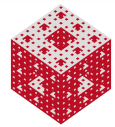
Therefore — if we know the cardinal functions — we can write the interpolant s to f at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ in the cardinal form

$$s(\boldsymbol{x}) = \sum_{j=1}^{N} f(\boldsymbol{x}_j)\overset{\star}{u}_j(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$



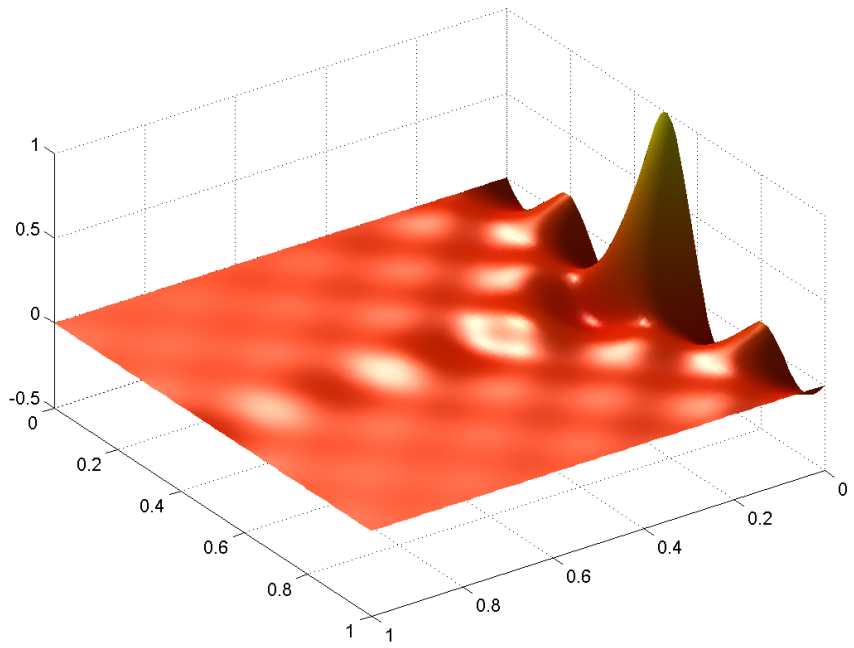
Remark

- *Cardinal functions do not depend on the data values of the interpolation problem.*
- *They do heavily depend on the data locations (see plots on following slides).*
- *Once the data sites are fixed and the basic function is chosen with an appropriate shape parameter (whose optimal value will depend on the data sites and values), then the cardinal functions are determined by the linear system (1).*
- *The cardinal functions are formally identical to the optimal kriging weights of Chapter 4.*



Example

Gaussian Cardinal Functions



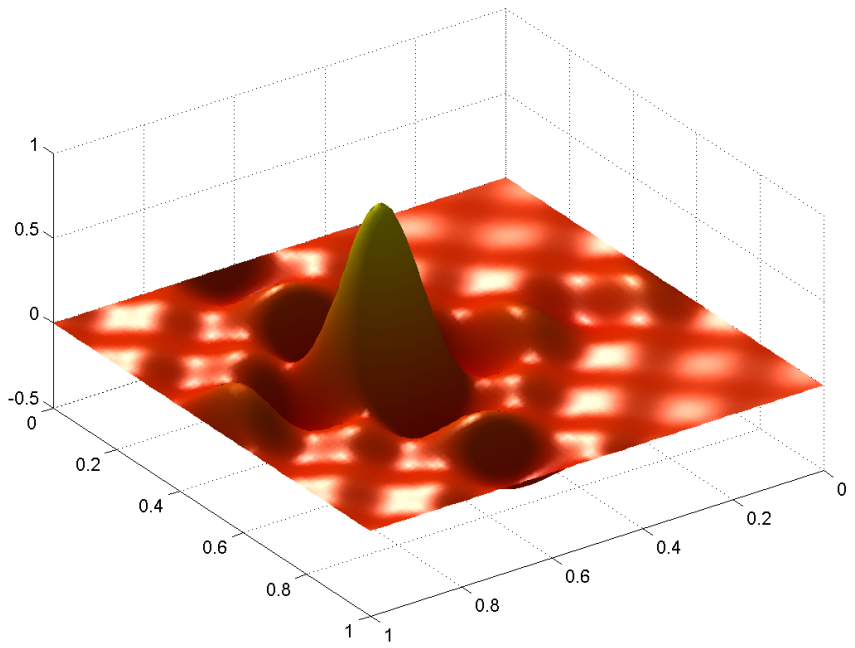
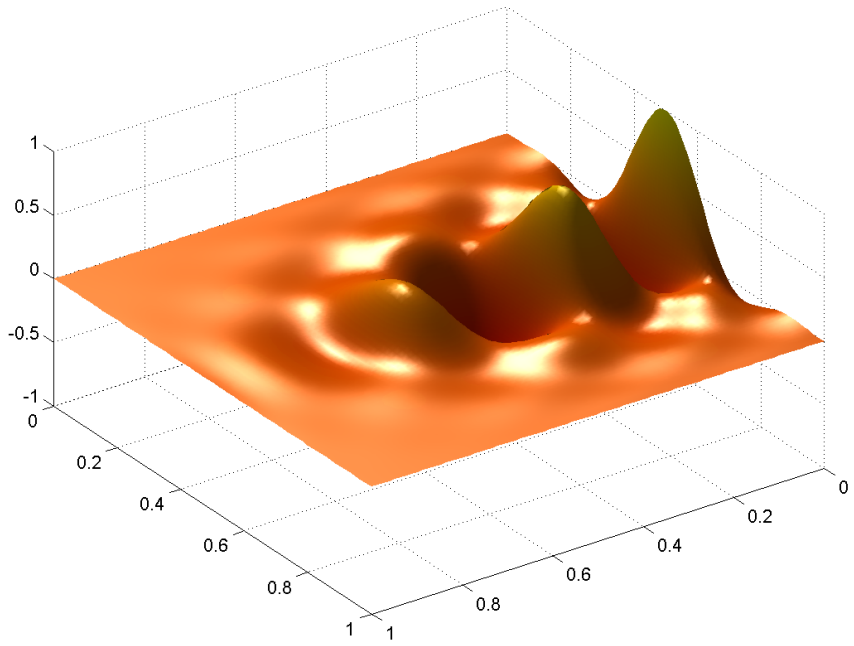


Figure: Cardinal functions for Gaussian interpolation (with $\varepsilon = 5$) on 81 uniformly gridded points in $[0, 1]^2$. Centered at an edge point (left) and at an interior point (right).

Example

More Gaussians



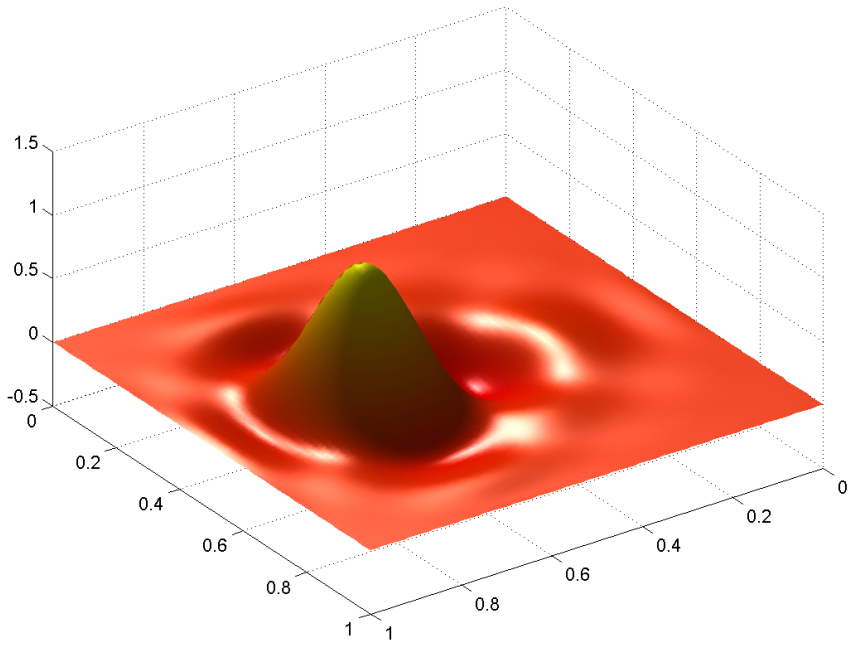
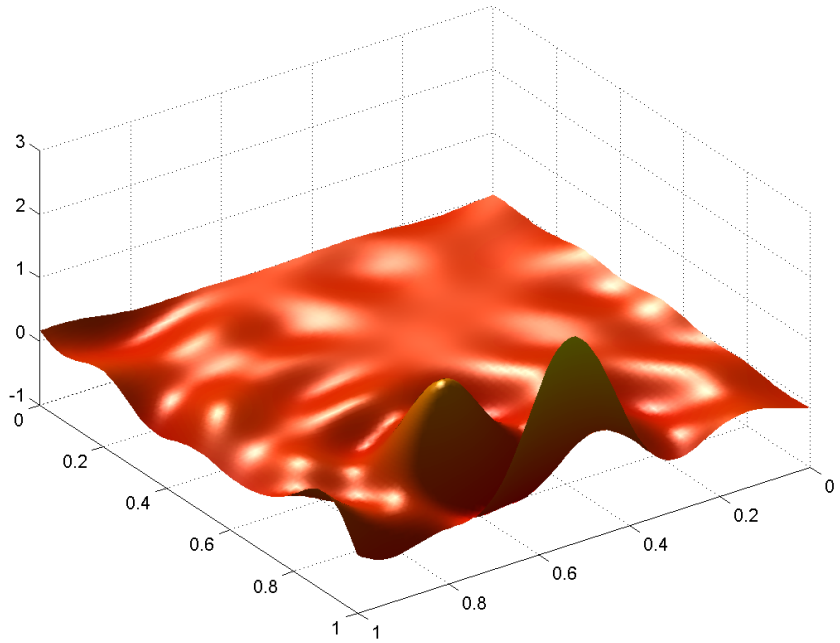


Figure: Cardinal functions for Gaussian interpolation (with $\varepsilon = 5$) on 81 tensor-product Chebyshev points in $[0, 1]^2$. Centered at an edge point (left) and at an interior point (right).

Example

More Gaussians



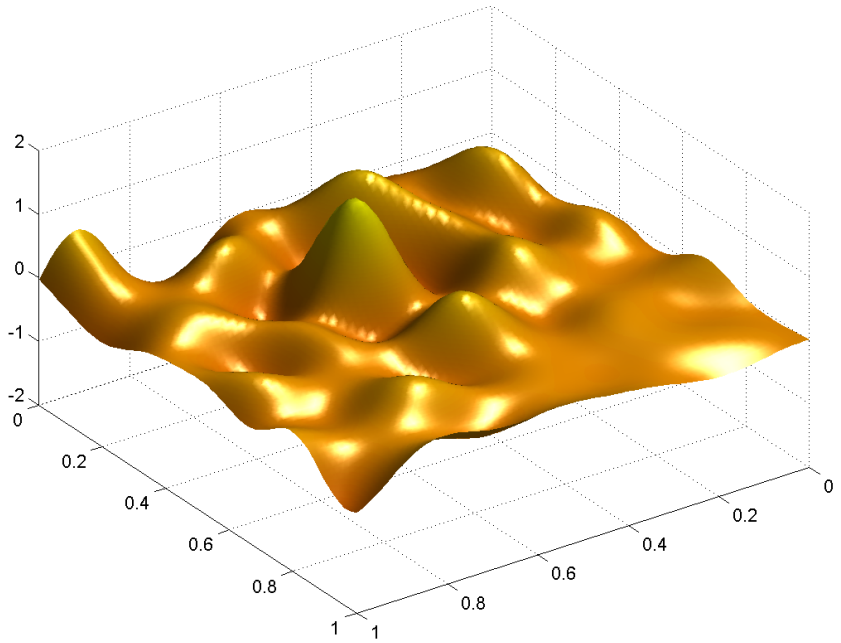
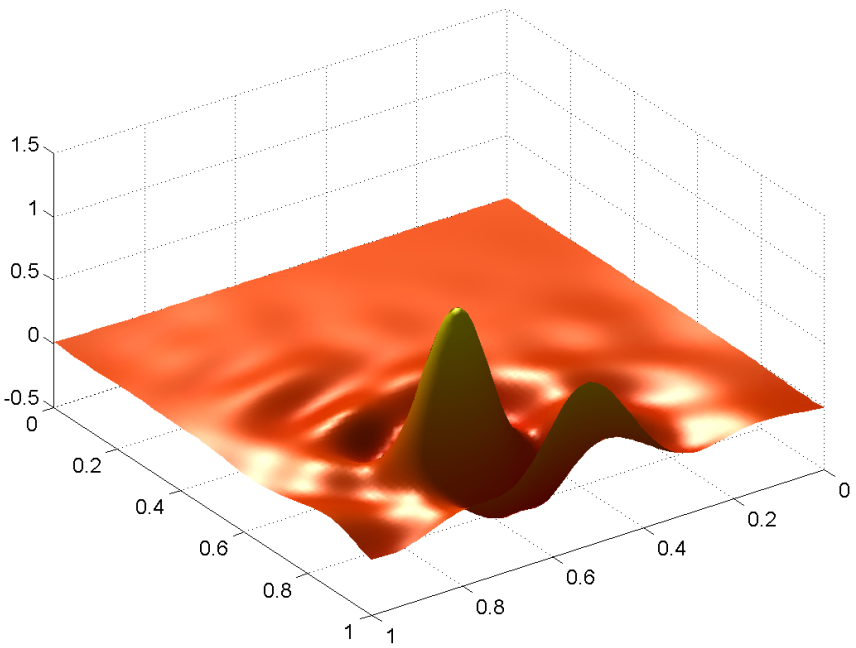


Figure: Cardinal functions for Gaussian interpolation (with $\varepsilon = 5$) on 81 Halton points in $[0, 1]^2$. Centered at an edge point (left) and at an interior point (right).

Example

Multiquadric Cardinal Functions



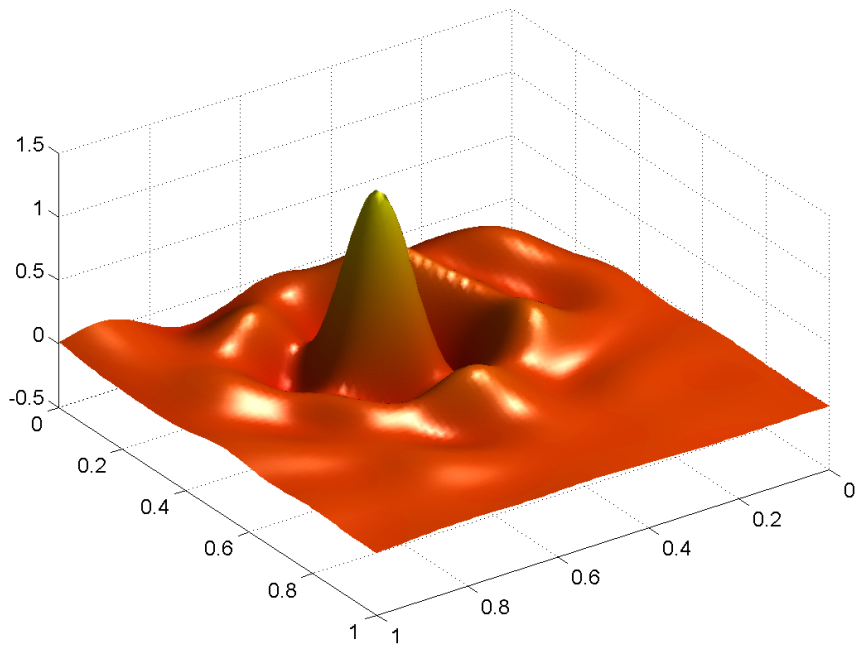
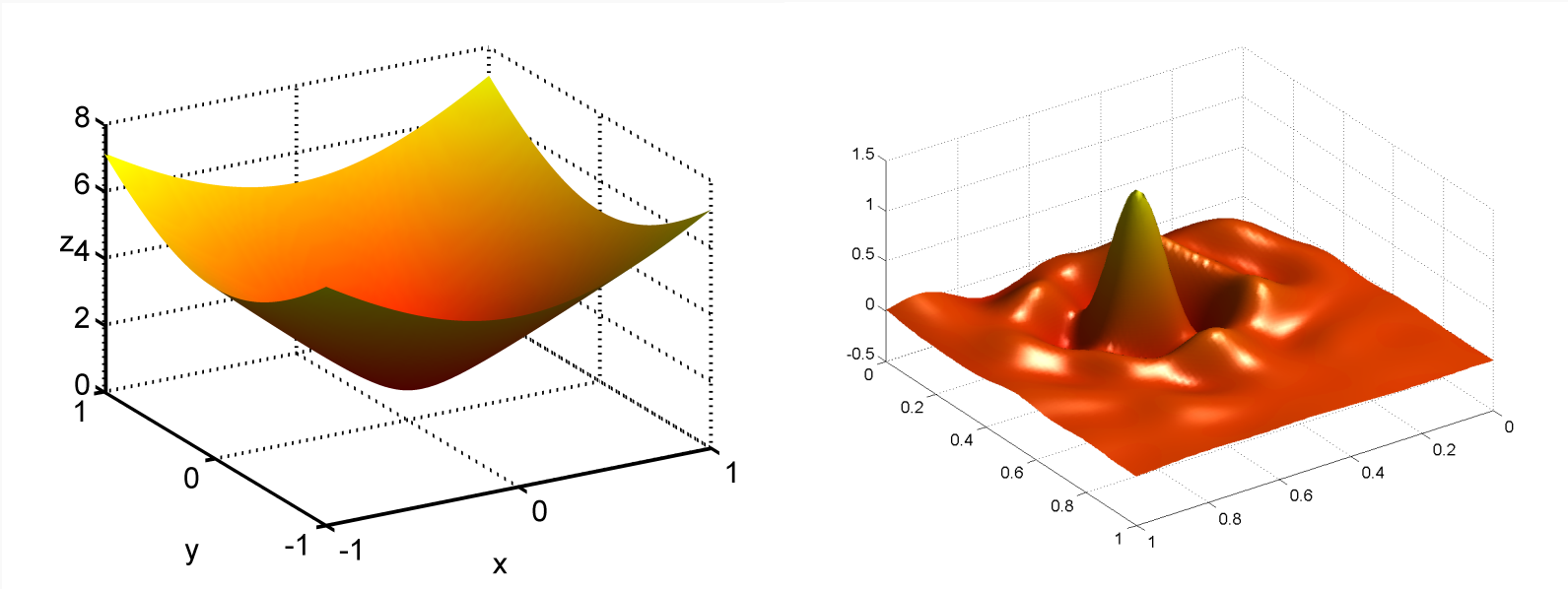


Figure: Cardinal functions for multiquadric interpolation (with $\varepsilon = 5$) on 81 Halton points in $[0, 1]^2$. Centered at an edge point (left) and at an interior point (right).

Remark

Basic functions that grow with increasing distance from the center point (such as multiquadrics) are sometimes criticized for being "counter-intuitive" for scattered data approximation.



The plot above shows that the associated cardinal functions are just as localized as those for the Gaussian basic functions, and thus the function space spanned by multiquadrics is a "good" local space.

The Power Function

Another important ingredient needed for our error estimates is the so-called power function.

For any strictly positive definite kernel $K \in C(\Omega \times \Omega)$, $\Omega \subseteq \mathbb{R}^d$, any set of distinct points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \Omega$, and an arbitrary vector $\boldsymbol{u} \in \mathbb{R}^N$, we define the quadratic form

$$Q(\boldsymbol{u}) = K(\boldsymbol{x}, \boldsymbol{x}) - 2\sum_{j=1}^{N} u_j K(\boldsymbol{x}, \boldsymbol{x}_j) + \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j K(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

Definition

Suppose $\Omega \subseteq \mathbb{R}^d$ and $K \in C(\Omega \times \Omega)$ is strictly positive definite. For any distinct points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \Omega$ the power function $P_{K,\mathcal{X}}$ is defined pointwise by

$$[P_{K,\mathcal{X}}(\boldsymbol{x})]^2 = Q(\overset{\star}{\boldsymbol{u}}(\boldsymbol{x})),$$

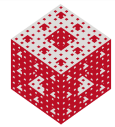
where $\overset{\star}{\boldsymbol{u}}$ is the vector of cardinal functions studied above.

Using the definition of the native space norm from Chapter 2 we can rewrite the quadratic form $Q(\boldsymbol{u})$ as

$$
\begin{aligned}
Q(\boldsymbol{u}) &= K(\boldsymbol{x},\boldsymbol{x}) - 2\sum_{j=1}^{N} u_j K(\boldsymbol{x},\boldsymbol{x}_j) + \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\
&= \langle K(\cdot,\boldsymbol{x}), K(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}_K(\Omega)} - 2\sum_{j=1}^{N} u_j \langle K(\cdot,\boldsymbol{x}), K(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}_K(\Omega)} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j \langle K(\cdot,\boldsymbol{x}_i), K(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}_K(\Omega)} \\
&= \langle K(\cdot,\boldsymbol{x}) - \sum_{j=1}^{N} u_j K(\cdot,\boldsymbol{x}_j), K(\cdot,\boldsymbol{x}) - \sum_{j=1}^{N} u_j K(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}_K(\Omega)} \\
&= \left\| K(\cdot,\boldsymbol{x}) - \sum_{j=1}^{N} u_j K(\cdot,\boldsymbol{x}_j) \right\|^2_{\mathcal{H}_K(\Omega)} = \left\| K(\cdot,\boldsymbol{x}) - \boldsymbol{k}(\cdot)^T \boldsymbol{u} \right\|^2_{\mathcal{H}_K(\Omega)} \qquad (2)
\end{aligned}
$$

Remark

- *The name power function was chosen by [Sch93] based on its connection to the power function of a statistical decision function (originally introduced in [NP36]).*
- *In the paper [WS93] the power function was referred to as kriging function. However, as we saw in Chapter 4, the power function corresponds to the square root of the kriging variance.*



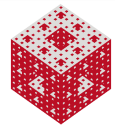
Using the linear system notation employed earlier, i.e., $\mathrm{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, N$, $\boldsymbol{u} = (u_1, \ldots, u_N)^T$, and $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_N))^T$, we note that we can also rewrite the quadratic form $Q(\boldsymbol{u})$ as

$$\begin{aligned} Q(\boldsymbol{u}) &= K(\boldsymbol{x}, \boldsymbol{x}) - 2\sum_{j=1}^{N} u_j K(\boldsymbol{x}, \boldsymbol{x}_j) + \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{k}(\boldsymbol{x})^T \boldsymbol{u} + \boldsymbol{u}^T \mathrm{K} \boldsymbol{u}. \end{aligned} \tag{3}$$

This suggests two alternative representations of the power function.

Using the matrix-vector notation for $Q(\boldsymbol{u})$, the power function is given as

$$P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{Q(\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}))} = \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{k}(\boldsymbol{x})^T \overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) + \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})^T \mathrm{K} \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})}.$$



However, by the definition of the cardinal functions $\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})$, and therefore we have the two new variants

$$\begin{aligned} P_{K,\mathcal{X}}(\boldsymbol{x}) &= \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - 2\boldsymbol{k}(\boldsymbol{x})^T\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) + \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})^T\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x})} \\ &= \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\overset{\star}{\boldsymbol{u}}(\boldsymbol{x})} \\ &= \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})^T\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x})}. \end{aligned}$$

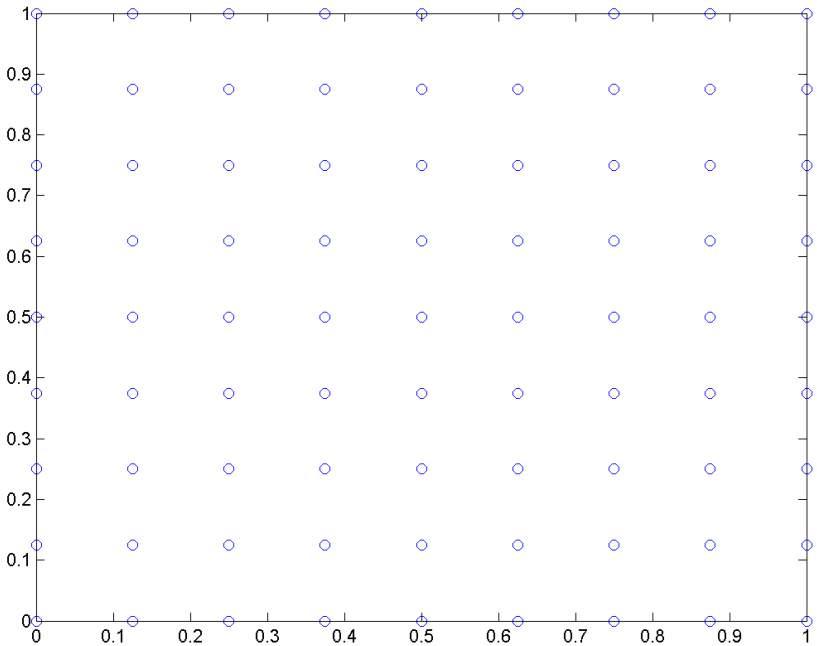
Remark

These formulas can be used for the numerical evaluation of the power function at $\boldsymbol{x}$.
To this end one has to first find the value of the cardinal functions $\overset{\star}{\boldsymbol{u}}(\boldsymbol{x})$ by solving the system $\mathrm{K}\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})$.
This results in

$$\boxed{P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})}.} \tag{4}$$

Example

Gaussian Power Function



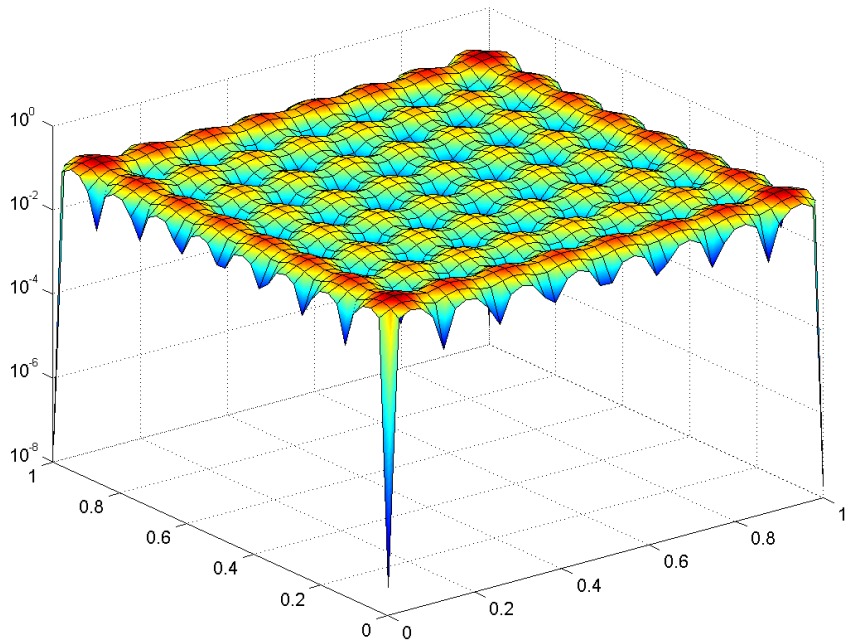
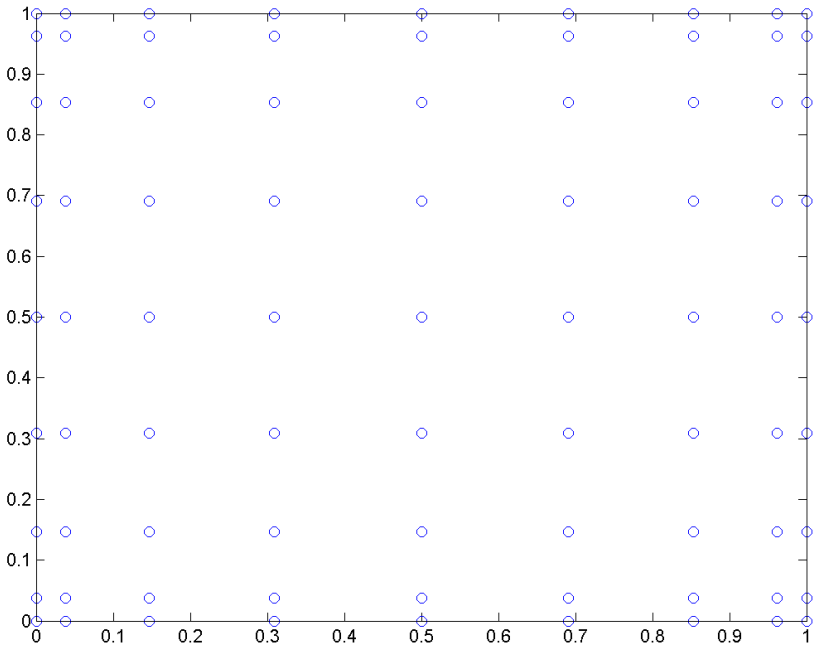


Figure: Data sites and power function for Gaussian interpolant with $\varepsilon = 6$ based on $N = 81$ uniformly gridded points in $[0, 1]^2$.

Example

More Gaussian Power Function



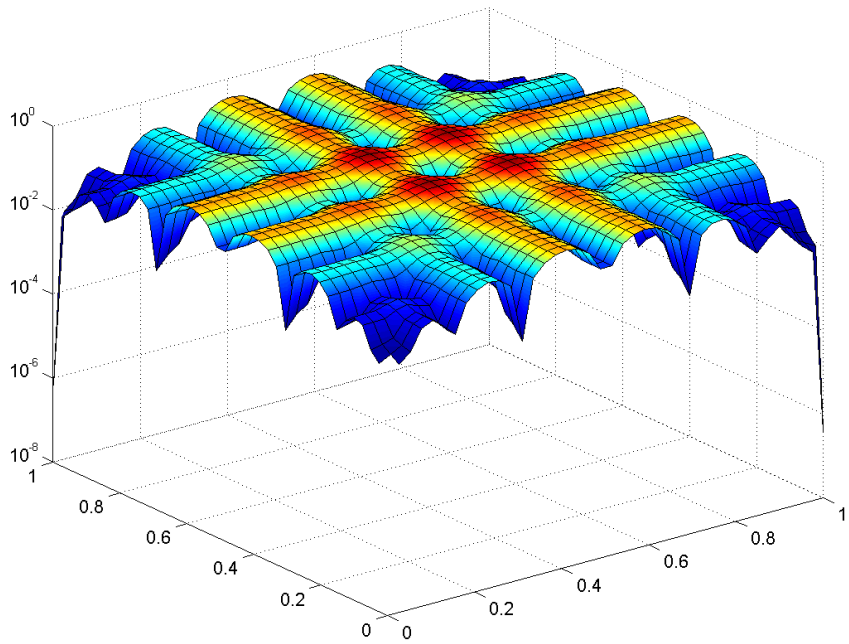
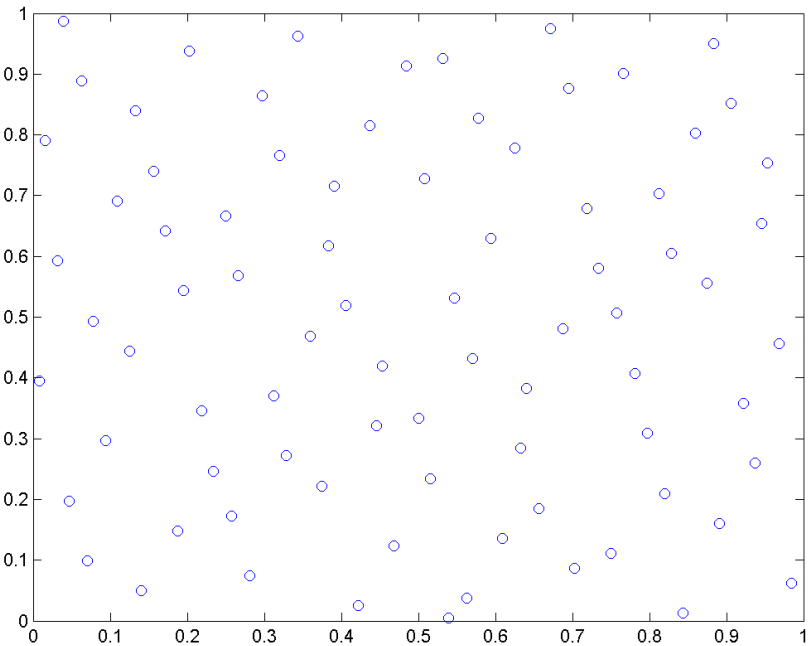


Figure: Data sites and power function for Gaussian interpolant with $\varepsilon = 6$ based on $N = 81$ tensor-product Chebyshev points in $[0, 1]^2$.

Example

More Gaussian Power Function



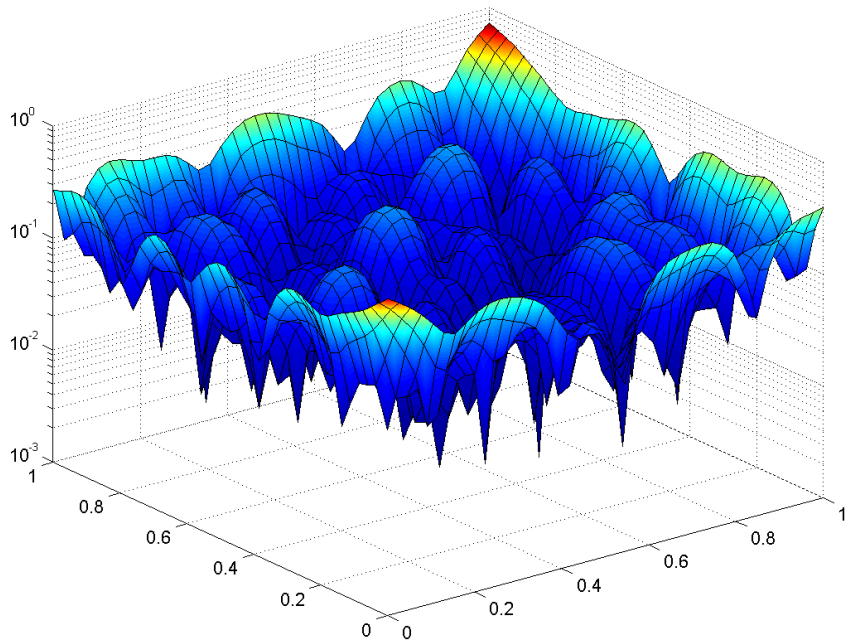


Figure: Data sites and power function for Gaussian interpolant with $\varepsilon = 6$ based on $N = 81$ Halton points in $[0, 1]^2$.

Remark

- *Dependence of the power function on the data locations is clearly visible.*
- *This connection was used in [DMSW05] to obtain an optimal set of data locations that are independent of the data values.*
- *Since* K *is a positive definite matrix whenever* K *is a strictly positive definite kernel we see that the power function satisfies the bounds*

$$0 \leq P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})^T \mathrm{K} \overset{\star}{\boldsymbol{u}}(\boldsymbol{x})} \leq \sqrt{K(\boldsymbol{x},\boldsymbol{x})}.$$

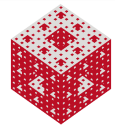
- *At this point — in the deterministic setting — the power function is mostly a theoretical tool that helps us better understand error estimates since we can decouple the effects due to the data function* f *from those due to the kernel* K *and the data locations* $\mathcal{X}$ *(see the following theorem).*
- *However, the kriging variance has a practical use in defining confidence intervals.*

Now we can give a first generic error estimate.

Theorem

Let $\Omega \subseteq \mathbb{R}^d$, $K \in C(\Omega \times \Omega)$ be strictly positive definite, and suppose that the points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are distinct. Denote the interpolant to $f \in \mathcal{H}_K(\Omega)$ on $\mathcal{X}$ by s. Then for every $\boldsymbol{x} \in \Omega$

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq P_{K,\mathcal{X}}(\boldsymbol{x}) \|f\|_{\mathcal{H}_K(\Omega)}.$$



Proof.

Since f is assumed to lie in the native space of K the reproducing property of K yields

$$f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_K(\Omega)}.$$

We express the interpolant in its cardinal form and apply the reproducing property of K. This gives us

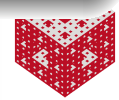
$$\begin{aligned} s(\boldsymbol{x}) &= \sum_{j=1}^{N} f(\boldsymbol{x}_j) \overset{\star}{u}_j(\boldsymbol{x}) \\ &= \sum_{j=1}^{N} \overset{\star}{u}_j(\boldsymbol{x}) \langle f, K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)} \\ &= \langle f, \sum_{j=1}^{N} \overset{\star}{u}_j(\boldsymbol{x}) K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)}. \end{aligned}$$

Proof (cont.)

Now all that remains to be done is to combine the two formulas just derived and apply the Cauchy-Schwarz inequality. Thus,

$$\begin{aligned} |f(\boldsymbol{x}) - s(\boldsymbol{x})| &= \left| \langle f, K(\cdot, \boldsymbol{x}) - \sum_{j=1}^{N} \overset{\star}{u}_j(\boldsymbol{x}) K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K(\Omega)} \right| \\ &\leq \|f\|_{\mathcal{H}_K(\Omega)} \left\| K(\cdot, \boldsymbol{x}) - \sum_{j=1}^{N} \overset{\star}{u}_j(\boldsymbol{x}) K(\cdot, \boldsymbol{x}_j) \right\|_{\mathcal{H}_K(\Omega)} \\ &= \|f\|_{\mathcal{H}_K(\Omega)} P_{K,\mathcal{X}}(\boldsymbol{x}), \end{aligned}$$

where we have used the representation (2) of the quadratic form $Q(\overset{\star}{\boldsymbol{u}}(\boldsymbol{x}))$ and the definition of the power function. □



One of the main benefits of the above theorem is that we are now able to estimate the interpolation error by considering two independent phenomena:

- the smoothness of the data (measured in terms of the native space norm of f — which is independent of the data locations, but does depend on K),
- and the distribution of the data (measured in terms of the power function — independent of the actual data values).

Remark

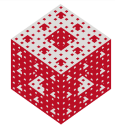
- *This is analogous to the standard error estimate for polynomial interpolation cited in most numerical analysis texts.*
- *Effects due to the use of any specific kernel K (or basic function in the translation invariant or radial case) are felt in both terms since the native space norm of f also varies with K.*
  - *In particular, changing a possible shape parameter ε will have an effect on both terms in the error bound.*

Error Bounds via Sampling Inequalities

We now use so-called sampling inequalities to obtain error bounds. Sometimes this technique is referred to as a zeros lemma.

The basic idea for sampling inequalities seems to have originated in the French school with the work of Duchon, Arcangéli and Atteia [Duc78, Arc74]. Another early paper is [MP85].

The paper [NWW05] started a long string of recent activity such as [ALdST07, ALdST09, ALdST12, GRZ13, HNW10, HNSW11, Mad06, NWW06, Rie08, RSZ10, RZ08, RZ14, WR05, ZS13].

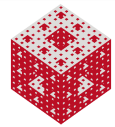


How sampling inequalities lead to error bounds

A sampling inequality formalizes in a rigorous way the following very general — and also intuitively convincing — idea:

It's impossible for a smooth function u to become uncontrollably large on its domain Ω provided

1. samples of u obtained on a sufficiently dense discrete subset $\mathcal{X}$ of Ω are small enough (or zero), and
2. the derivatives of u are bounded on Ω.



A typical sampling inequality looks like this:

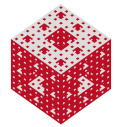
$$|u|_{W_q^m(\Omega)} \leq C \left(h^{k-m-d\left(\frac{1}{p}-\frac{1}{q}\right)_+} |u|_{W_p^k(\Omega)} + h^{-m} \|\boldsymbol{u}\|_{\ell_\infty} \right), \tag{5}$$

i.e., the weak semi-norm $|\cdot|_{W_q^m(\Omega)}$ is bounded by

- a strong semi-norm $|\cdot|_{W_p^k(\Omega)}$ (defined in terms of $m \geq 0$ and $k > m + \frac{d}{p}$), and
- a discrete norm $\|\cdot\|_{\ell_\infty}$ of the values $\boldsymbol{u} = (u(\boldsymbol{x}_1), \ldots, u(\boldsymbol{x}_N))^T$ of u on the set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ with fill distance h.

Note that

- $|u|_{W_p^k(\Omega)}$ is multiplied by a positive power of h (i.e., its contribution decreases as h gets small),
- while $\|\boldsymbol{u}\|_{\ell_\infty}$ is multiplied by a non-positive power of h (i.e., its contribution increases as h gets small).



In order to get a useful error bound we apply our sampling inequalities in the case where the function u is interpreted as the residual $f - s$ between the function f and its approximation s.

Then, for any successful approximation method, the samples of the residual will be small and will (hopefully) outweigh the h^{-m} factor.

In particular, if s is obtained by interpolation on $\mathcal{X}$, then the residual vector $\|\boldsymbol{u}\|_{\ell_\infty} = 0$, and we get

$$|f - s|_{W_q^m(\Omega)} \leq C h^{k-m-d\left(\frac{1}{p}-\frac{1}{q}\right)_+} |f - s|_{W_p^k(\Omega)}.$$

Remark

This bound does not quite look like an error bound yet since it involves the residual $f - s$ also on the right-hand side.

Therefore, an important assumption needs to be made:
We assume that our approximation satisfies the "energy split" (Pythagorean) lemma from above, i.e.,

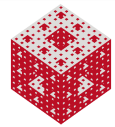
$$|f|^2_{W_p^k(\Omega)} = |f-s|^2_{W_p^k(\Omega)} + |s|^2_{W_p^k(\Omega)},$$

which implies, in particular, that

$$|f-s|_{W_p^k(\Omega)} \leq |f|_{W_p^k(\Omega)}.$$

This is certainly true if the native space $\mathcal{H}_K(\Omega)$ is equivalent to the Sobolev space $W_p^k(\Omega)$. Then we end up with the error bound

$$|f-s|_{W_q^m(\Omega)} \leq Ch^{k-m-d\left(\frac{1}{p}-\frac{1}{q}\right)_+}|f|_{W_p^k(\Omega)}.$$



To make this bound look a bit more familiar we consider

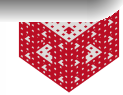
- $\Omega = [a, b]$,
- set $m = 0$ (no simultaneous approximation of derivatives),
- let $p = q = 2$ and
- convert the Sobolev (semi-)norm on the right-hand side to an L_2-norm of derivatives.

Then we get for any $f, s \in C^k([a, b])$

$$\|f - s\|_{L_2([a,b])} \leq Ch^k \|f^{(k)}\|_{L_2([a,b])}.$$

Remark

This is, e.g., the kind of error bound that is typical for piecewise polynomial spline interpolation.

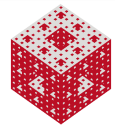


Sampling Inequalities for Kernel-based Approximation

We now look at error bounds for kernel-based interpolation methods obtained via sampling inequalities.

Sampling inequalities in 1D are relatively easy to understand and to obtain. So we first consider the univariate case in detail. This will cover, e.g., our iterated Brownian bridge kernels.

Then we outline the important steps for the much more complicated multivariate case.

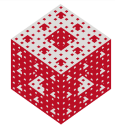


We consider a univariate kernel-based interpolation to data $\{x_i, f(x_i)\}_{i=1}^N$ at N distinct points $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \Omega = [a, b]$ with fill distance h in the domain Ω, i.e.,

$$h = \sup_{\boldsymbol{x} \in \Omega} \min_{\boldsymbol{x}_j \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}_j\|.$$

As discussed above, we must first obtain a sampling inequality for a generic function u of the same smoothness class as the kernel of interest.

Here we assume that $u \in C^k(\Omega)$ for some non-negative integer k.



We begin by constructing a *local* polynomial interpolant of order k (i.e., degree $k-1$) to the generic function $u \in C^k(\Omega)$ at the points $\mathcal{X}_x = \{t_1, \ldots, t_k\} \subset \mathcal{X}$.

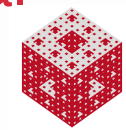
Remark

For convenience we denote the local points with t_j, $j = 1, \ldots, k$, instead of picking appropriate subsets of indices for the original points in $\mathcal{X}$.

We pick the local interpolation points in dependence on the point of evaluation x, i.e., $\mathcal{X}_x$ consists of the k nearest neighbors of x from the set of global interpolation points $\mathcal{X}$; this implies that we need $N \geq k$.

This ensure that the local interpolation points are chosen as "symmetrical" as possible around x.

Furthermore, we define the interval I_x as the smallest closed interval which contains the points $\mathcal{X}_x$.



The uniqueness of univariate polynomial interpolation ensures that there exists a constant C such that for any polynomial p of degree at most $k-1$ we have

$$\sup_{x \in I_x} |p(x)| \leq C \sup_{t_j \in \mathcal{X}_x} \left|p(t_j)\right|, \tag{6}$$

i.e., the polynomial p is *determined* by its values on the set $\mathcal{X}_x$.

Remark

- *A consequence of this property is that*

$$p(t_j) = 0 \text{ for all } t_j \in \mathcal{X}_x \quad \Longrightarrow \quad p \equiv 0 \text{ on } I_x,$$

  i.e., the points in $\mathcal{X}_x$ determine p.
- *In the univariate setting this is straightforward, but in the multivariate setting this causes a first difficulty, calling for the introduction of so-called norming sets.*

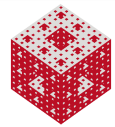
A useful exact error formula for polynomial interpolation (known as Kowalewski's exact remainder formula, see [Dav63, Eqn. (3.7.10)], [Mad06]) can be obtained starting from a Taylor expansion of u.

We give the derivation of this formula since it sheds some light on the procedure recommended for the multivariate setting.

Given $u \in C^k(\Omega)$ and an arbitrary point $t_j \in \mathcal{X}_x$, we begin with [Mad06, Eqn. (2.2.1)]

$$u(x) = u(t_j) - \sum_{i=1}^{k-1} \frac{u^{(i)}(x)}{i!}(t_j - x)^i + \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\,\mathrm{d}t, \quad (7)$$

which is obtained by swapping the roles of x and t_j in the usual version of Taylor's theorem with remainder.



Now we consider the k^{th} order Lagrange polynomials for interpolation on $\mathcal{X}_x$ which satisfy

- the cardinality conditions $L_j(t_i) = \delta_{ij}$, $i, j = 1, \ldots, k$ and
- reproduce polynomials up to order k, i.e.,

$$\sum_{j=1}^{k} p(t_j) L_j(x) = p(x), \quad x \in I_x, \tag{8}$$

In particular, we have

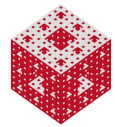
$$\sum_{j=1}^{k} L_j(x) \equiv 1.$$

This is known as a partition of unity.

Using this partition of unity property we can obtain a weighted average of the Taylor formula (7) for $j = 1, \ldots, k$, i.e., we multiply both sides of (7) by $L_j(x)$ and then sum over j from 1 to k resulting in

$$u(x) = \sum_{j=1}^{k} L_j(x) u(t_j) - \sum_{j=1}^{k} L_j(x) \sum_{i=1}^{k-1} \frac{u^{(i)}(x)}{i!} (t_j - x)^i + \sum_{j=1}^{k} L_j(x) \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\, \mathrm{d}t. \tag{9}$$

- The first sum is nothing but $p_k(x)$, the unique k^{th} order polynomial interpolating u on $\mathcal{X}_x$, since $u(t_j) = p_k(t_j)$, $j = 1, \ldots, k$, and p_k is defined by its Lagrange form $p_k(x) = \sum_{j=1}^{k} L_j(x) p_k(t_j)$.
- The reproduction of polynomials up to degree $k-1$ (8) implies that the double sum vanishes since $\sum_{j=1}^{k} L_j(x)(t_j - x)^i = 0$, $i = 1, \ldots, k-1$.



The result is Kowalewski's exact remainder formula as stated in [Mad06]:

$$u(x) = p_k(x) + \sum_{j=1}^{k} L_j(x) \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\,\mathrm{d}t. \tag{10}$$

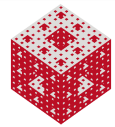
By replacing p_k by its Lagrange form as in (9), and then applying the triangle inequality to (10) and bounding the values of u in the first sum by their maximum we get

$$|u(x)| \le \Lambda_k \max_{t_j \in \mathcal{X}_x} |u(t_j)| + \sum_{j=1}^{k} |L_j(x)| \left| \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\,\mathrm{d}t \right|, \tag{11}$$

where we have introduced the abbreviation

$$\Lambda_k = \max_{x \in I_x} \sum_{j=1}^{k} |L_j(x)|$$

to denote the Lebesgue constant for polynomial interpolation at the points $t_1, \ldots, t_k$.

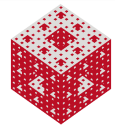


Remark

The Lebesgue constant depends on the distribution of interpolation points.

- *It grows logarithmically in k for Chebyshev points (which is the minimal rate of growth, but not with optimal constant),*
- *while it grows exponentially in k for equally spaced points.*

Note, however, that this growth is not important for us since k is fixed here.



In order to remove the Lagrange polynomials from the second term in (11) we make use of (6) so that

$$|L_j(x)| \leq C_j \max_{t_i \in \mathcal{X}_x} |L_j(t_i)| \leq C_j,$$

by the cardinality of the Lagrange polynomials.
This leaves us with

$$|u(x)| \leq \Lambda_k \max_{t_j \in \mathcal{X}_x} |u(t_j)| + C \sum_{j=1}^{k} \left| \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\, \mathrm{d}t \right|,$$

where $C = \max_{j=1,\ldots,k} C_j$.
Now we bound the integral using the Cauchy–Schwarz inequality, i.e.,

$$\left| \int_{t_j}^{x} \frac{(t_j - t)^{k-1}}{(k-1)!} u^{(k)}(t)\, \mathrm{d}t \right| \leq \left| \int_{t_j}^{x} \frac{(t_j - t)^{2k-2}}{((k-1)!)^2}\, \mathrm{d}t \right|^{1/2} \left| \int_{t_j}^{x} \left(u^{(k)}(t)\right)^2 \mathrm{d}t \right|^{1/2}$$

$$= \frac{|t_j - x|^{k-1/2}}{\sqrt{2k-1}(k-1)!} \left| \int_{t_j}^{x} \left(u^{(k)}(t)\right)^2 \mathrm{d}t \right|^{1/2}.$$

This results in

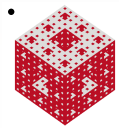
$$|u(x)| \leq \Lambda_k \max_{t_j \in \mathcal{X}_x} |u(t_j)| + C \sum_{j=1}^{k} \left\{ \frac{|t_j - x|^{k-1/2}}{\sqrt{2k-1}(k-1)!} \left| \int_{t_j}^{x} \left(u^{(k)}(t) \right)^2 \mathrm{d}t \right|^{1/2} \right\}.$$

We will now aim for a sampling inequality in the L_2-norm.
Therefore we next square both sides of this inequality and apply the Cauchy–Schwarz estimates

$$(A+B)^2 \leq 2A^2 + 2B^2 \quad \text{and} \quad \left(\sum_{j=1}^{k} A_j \right)^2 \leq k \sum_{j=1}^{k} (A_j)^2$$

to the right-hand side.
This yields

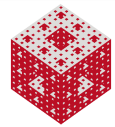
$$\begin{aligned} |u(x)|^2 \leq & 2\Lambda_k^2 \left(\max_{t_j \in \mathcal{X}_x} |u(t_j)| \right)^2 \\ & + \frac{2kC^2}{(2k-1)\left((k-1)!\right)^2} \sum_{j=1}^{k} |t_j - x|^{2k-1} \left| \int_{t_j}^{x} \left(u^{(k)}(t) \right)^2 \mathrm{d}t \right|. \end{aligned} \tag{12}$$



To obtain an L_2-norm estimate on the local interval I_x we will have to integrate both sides over $I_x = [\alpha, \beta]$ with appropriately chosen endpoints α and β.

We observe what happens if we integrate one of the summands on the right-hand side of (12) considering that $t_j \in [\alpha, \beta]$:

$$\begin{aligned}
&\int_\alpha^\beta |t_j - x|^{2k-1} \left| \int_{t_j}^x \left(u^{(k)}(t)\right)^2 \mathrm{d}t \right| \mathrm{d}x \\
&\quad = \int_\alpha^{t_j} \left(u^{(k)}(t)\right)^2 \int_t^\alpha |t_j - x|^{2k-1} \,\mathrm{d}x\,\mathrm{d}t + \int_{t_j}^\beta \left(u^{(k)}(t)\right)^2 \int_t^\beta |t_j - x|^{2k-1} \,\mathrm{d}x\,\mathrm{d}t \\
&\quad \leq \frac{|t_j - \alpha|^{2k}}{2k} \int_\alpha^{t_j} \left(u^{(k)}(t)\right)^2 \mathrm{d}t + \frac{|t_j - \beta|^{2k}}{2k} \int_{t_j}^\beta \left(u^{(k)}(t)\right)^2 \mathrm{d}t.
\end{aligned}$$

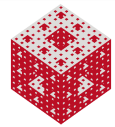


Since both $|t_j - \alpha|$ and $|t_j - \beta|$ are at most kh (because of the definition of fill distance) we get

$$\int_\alpha^\beta |t_j - x|^{2k-1} \left| \int_{t_j}^x \left(u^{(k)}(t)\right)^2 \mathrm{d}t \right| \mathrm{d}x \leq \frac{(kh)^{2k}}{2k} \int_\alpha^\beta \left(u^{(k)}(t)\right)^2 \mathrm{d}t.$$

We now plug this into the integrated version of (12) and use $\int_\alpha^\beta \left(u^{(k)}(t)\right)^2 \mathrm{d}t = \|u^{(k)}\|^2_{L_2(I_x)}$ to obtain

$$\begin{aligned}
\|u\|^2_{L_2(I_x)} &\leq 2kh\Lambda_k^2 \left(\max_{t_j \in \mathcal{X}_x} |u(t_j)| \right)^2 + \frac{2kC^2}{(2k-1)\left((k-1)!\right)^2} \sum_{j=1}^k \frac{(kh)^{2k}}{2k} \|u^{(k)}\|^2_{L_2(I_x)} \\
&= 2kh\Lambda_k^2 \left(\max_{t_j \in \mathcal{X}_x} |u(t_j)| \right)^2 + \frac{kC^2(kh)^{2k}}{(2k-1)\left((k-1)!\right)^2} \|u^{(k)}\|^2_{L_2(I_x)}.
\end{aligned}$$



Note that we can rewrite

$$\|u\|^2_{L_2(I_x)} \leq 2kh\Lambda_k^2 \left(\max_{t_j \in \mathcal{X}_x} |u(t_j)| \right)^2 + \frac{kC^2(kh)^{2k}}{(2k-1)\left((k-1)!\right)^2} \|u^{(k)}\|^2_{L_2(I_x)}.$$

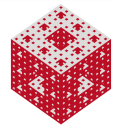
as

$$\|u\|^2_{L_2(I_x)} \leq C_1^2 h \left(\max_{t_j \in \mathcal{X}_x} |u(t_j)| \right)^2 + C_2^2 h^{2k} \|u^{(k)}\|^2_{L_2(I_x)}$$

provided the constants C_1 and C_2 (which depend on k and on the points in $\mathcal{X}_x$, but not on h or u) are defined accordingly.

Applying the inequality $\sqrt{A^2 + B^2} \leq A + B$ (for $A, B > 0$) we have the final form of the *local* sampling inequality

$$\|u\|_{L_2(I_x)} \leq C_1 \sqrt{h} \max_{t_j \in \mathcal{X}_x} |u(t_j)| + C_2 h^k \|u^{(k)}\|_{L_2(I_x)}.$$



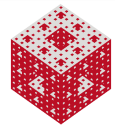
To finally obtain a *global* sampling inequality we use a technique introduced in [Duc78].
For this covering argument to apply we note that we can cover the interval $\Omega = [a, b]$ with a family of subintervals I_x as discussed above.
In fact,

- each of these subintervals has length at most kh and
- each point $x \in \Omega$ is covered by at most $k \leq N$ subintervals.

This allows us to sum up all the local estimates to arrive at

$$\|u\|_{L_2(\Omega)} \leq C_3 \sqrt{b-a} \max_{x_j \in \mathcal{X}} |u(x_j)| + C_4 h^k \|u^{(k)}\|_{L_2(\Omega)}, \tag{13}$$

where the constants C_3 and C_4 depend on k, N, and the distribution of the points in $\mathcal{X}$, but not on h and u.

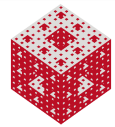


Note that

$$\|u\|_{L_2(\Omega)} \leq C_3\sqrt{b-a}\max_{x_j\in\mathcal{X}}|u(x_j)| + C_4 h^k \|u^{(k)}\|_{L_2(\Omega)},$$

looks like our typical sampling inequality (5) with

- $\Omega = [a, b]$,
- $C = \max\{C_3\sqrt{b-a}, C_4\}$,
- $m = 0$, and
- $p = q = 2$.



The sampling inequality (13) says that the continuous L_2-norm of u is bounded by

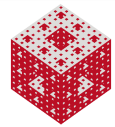
- the k^{th} power of the fill distance times the stronger H^k-norm of u and
- a discrete maximum-norm of the values of u on $\mathcal{X}$.

This estimate holds for any function $u \in C^k(\Omega)$.

In particular, it provides an error bound for any method that generates an interpolant $s \in C^k(\Omega)$ to a function $f \in C^k(\Omega)$ at the points in $\mathcal{X}$. In this case we have

$$\|f - s\|_{L_2(\Omega)} \leq Ch^k \|(f - s)^{(k)}\|_{L_2(\Omega)}$$

as outlined above.



In order to get the upper bound to depend only on the input data, i.e., f and $\mathcal{X}$ (or its fill distance h) we use the triangle inequality so that

$$\|f - s\|_{L_2(\Omega)} \leq Ch^k \left(\|f^{(k)}\|_{L_2(\Omega)} + \|s^{(k)}\|_{L_2(\Omega)} \right).$$

If we are working in a setting in which the interpolant is also a minimum norm interpolant (such as in a reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$) then

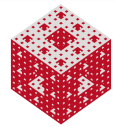
$$\|s^{(k)}\|_{L_2(\Omega)} \leq \|f^{(k)}\|_{L_2(\Omega)},$$

provided $\mathcal{H}_K(\Omega)$ is equivalent to the Sobolev space $H^k(\Omega)$, and we obtain the error bound

$$\|f - s\|_{L_2(\Omega)} \leq 2Ch^k \|f\|_{H^k(\Omega)}.$$

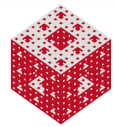
Remark

- *These orders in h, however, are only "half" of what is known as the optimal bounds in the literature.*
- *In order to "double" the orders one needs to take into account additional properties of the native Hilbert space.*
- *In particular, incorporating the boundary behavior of f and s is important to achieve those higher orders. We already saw this experimentally in Chapter 6.*



Remark

- *Estimates in other norms can be obtained similarly using Hölder's inequality instead of Cauchy–Schwarz.*
- *This approach can also be extended to other settings, such as (least squares) approximation or PDE error estimates — both in strong form and weak form.*
- *Details on more advanced versions of error bounds based on sampling inequalities are available in the original literature listed above.*



Sampling inequalities in higher dimensions

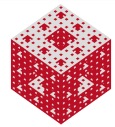
The main problem in higher dimensions is the misbehavior of polynomial interpolation:

- depending on the points in $\mathcal{X}$, the interpolating polynomial may not be unique, or it may not even exist (see, e.g., [GS00]).

Therefore, one usually resorts to

- oversampling and
- an approximation in terms of local (moving) least squares polynomials.

This approach then also calls for so-called norming sets as defined in [JSW99] (or as determining sets in [Mad06]).



A norming set Λ of a normed linear space $\mathcal{F}$ is a finite set of linear functionals (such as point evaluation functionals $\lambda_i(f) = f(\boldsymbol{x}_i),\ i = 1, \ldots, N$) defined on the dual of $\mathcal{F}$.

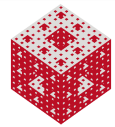
It provides a bound of the continuous $\mathcal{F}$-norm of f by only discrete samples, i.e.,

$$\|f\|_{\mathcal{F}} \leq C \sup_{\lambda_i \in \Lambda, \|\lambda_i\|=1} |\lambda_i(f)| \qquad \text{for all } f \in \mathcal{F} \tag{14}$$

with some positive constant C that depends on $\mathcal{F}$ and Λ.

We saw an example for the case of univariate polynomials and point evaluation functionals earlier:

$$\sup_{x \in I_x} |p(x)| \leq C \sup_{t_j \in \mathcal{X}_x} \left|p(t_j)\right|.$$



Following [RSZ10], we now sketch a standard way to prove a typical sampling inequality of the form

$$\|D^\alpha u\|_{L_q(\Omega)} \leq C \left(h^{k-|\alpha|-d\left(\frac{1}{p}-\frac{1}{q}\right)_+} |u|_{W_p^k(\Omega)} + h^{-|\alpha|} \|S_{\mathcal{X}} u\|_{\ell_\infty} \right).$$

Here $S_{\mathcal{X}}$ is a sampling operator on $\mathcal{X}$ and α is a multi-index so that D^α denotes a standard multivariate derivative operator of order $|\alpha|$.

For some domain $\mathcal{D}$, star-shaped with respect to a ball[1], let $\{a_j^{(\alpha)}, j = 1, \ldots, N\}$ be a local polynomial reproduction of degree k with respect to a discrete set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathcal{D}$, i.e.,

$$D^\alpha q(\boldsymbol{x}) = \sum_{j=1}^{N} a_j^{(\alpha)}(\boldsymbol{x}) q(\boldsymbol{x}_j)$$

holds for every multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq k$, all $\boldsymbol{x} \in \mathcal{D}$ and all $q \in \Pi_k^d(\mathcal{D})$, where Π_d^k denotes the space of all d-variate polynomials of degree at most k.

[1] $\mathcal{D}$ is called *star-shaped with respect to a ball* B if, for all $\boldsymbol{x} \in \mathcal{D}$, the closed convex hull of $\{\boldsymbol{x}\} \cup B$ is a subset of $\mathcal{D}$ (see, e.g., [BS94, (4.2.2) Definition]).

Then we have

$$|D^\alpha u(\boldsymbol{x})| \leq \|D^\alpha(u-p)\|_{L_\infty(\mathcal{D})} + \sum_{j=1}^{N} \left| a_j^{(\alpha)}(\boldsymbol{x}) \right| \left(\|u-p\|_{L_\infty(\mathcal{D})} + \|S_{\mathcal{X}} u\|_{\ell_\infty} \right)$$

for arbitrary $u \in W_p^k(\mathcal{D})$ and any polynomial $p \in \Pi_d^k(\mathcal{D})$.

Using a polynomial reproduction argument based on norming sets, the Lebesgue constant can be bounded by $\sum_{j=1}^{N} \left| a_j^{(\alpha)}(\boldsymbol{x}) \right| \leq 2$, if some moderate oversampling is allowed.

Remark

As a local polynomial approximation one usually chooses the averaged Taylor polynomials of degree k (see [BS94, Section 4.1].
In contrast to the averaging process we used in the one-dimensional setting, the multivariate average needs to be constructed with an appropriately defined measure which reproduces polynomials.

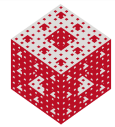
Doing all of this leads to a local sampling inequality of the form

$$\|D^\alpha u\|_{L_\infty(\mathcal{D})} \leq \frac{C}{(k-|\alpha|)!}\delta_{\mathcal{D}}^{k-d/p}\left(\delta_{\mathcal{D}}^{-|\alpha|}+h^{-|\alpha|}\right)|u|_{W_p^k(\mathcal{D})}+2h^{-|\alpha|}\|S_{\mathcal{X}}u\|_{\ell_\infty}$$

where $\delta_{\mathcal{D}}$ denotes the diameter of $\mathcal{D}$.

To derive sampling inequalities on a global Lipschitz domain Ω satisfying an interior cone condition, we cover Ω by domains $\mathcal{D}$ which are star-shaped with respect to a ball, satisfying $\delta_{\mathcal{D}} \approx h$ (see [Duc78] for details on such coverings).

Global estimates are obtained by summation or maximization over the local estimates (see, e.g., [Wen05, Theorem 11.32]) similar to what was done in the one-dimensional setting.



References I

[ALdST07] Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens, *An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing*, Numerische Mathematik **107** (2007), no. 2, 181–211.

[ALdST09] _____, *Estimates for functions in Sobolev spaces defined on unbounded domains*, Journal of Approximation Theory **161** (2009), no. 1, 198–212.

[ALdST12] _____, *Extension of sampling inequalities to Sobolev semi-norms of fractional order and derivative data*, Numerische Mathematik **121** (2012), no. 3, 587–608.

[Arc74] Rémi Arcangéli, *Études de problèmes de type elliptique ou parabolique avec conditions ponctuelles*, Ph.D. thesis, University of Toulouse, 1974.

[BS94] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer Verlag (New York), 1994.

[Dav63] Philip J Davis, *Interpolation and approximation.*, Blaisdell Pub. Co., New York, 1963.

References II

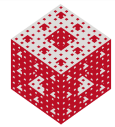
[DMSW05] S. De Marchi, R. Schaback, and H. Wendland, *Near-optimal data-independent point locations for radial basis function interpolation*, Advances in Computational Mathematics **23** (2005), no. 3, 317–330.

[Duc76] J. Duchon, *Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces*, Rev. Francaise Automat. Informat. Rech. Opér., Anal. Numer. **10** (1976), 5–12.

[Duc77] _____, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, Constructive Theory of Functions of Several Variables, Oberwolfach 1976 (W. Schempp and K. Zeller, eds.), Springer-Verlag (Berlin), 1977, Springer Lecture Notes in Math. 571, pp. 85–100.

[Duc78] _____, *Sur l'erreur d'interpolation des fonctions de plusieurs variables par les d^m-splines*, RAIRO Analyse Numerique **12** (1978), no. 4, 325–334.

[Duc80] _____, *Fonctions splines homogènes à plusiers variables*, Tech. report, Université de Grenoble, 1980.



References III

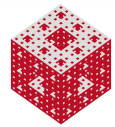
[Fas07] G. E. Fasshauer, *Meshfree Approximation Methods with* MATLAB, Interdisciplinary Mathematical Sciences, vol. 6, World Scientific Publishing Co., Singapore, 2007.

[GRZ13] M. Griebel, C. Rieger, and B. Zwicknagl, *Multiscale approximation and reproducing kernel Hilbert space methods*, 2013.

[GS00] Mariano Gasca and Thomas Sauer, *Polynomial interpolation in several variables*, Advances in Computational Mathematics **12** (2000), no. 4, 377–410.

[GW59] M. Golomb and H. F. Weinberger, *Optimal approximation and error bounds*, On Numerical Approximation (R. E. Langer, ed.), University of Wisconsin Press, 1959, pp. 117–190.

[HNSW11] T. Hangelbroek, F. J. Narcowich, X. Sun, and J. D. Ward, *Kernel approximation on manifolds II: the L_∞ norm of the L_2 projector*, SIAM Journal on Mathematical Analysis **43** (2011), no. 2, 662–684.



References IV

[HNW10] T. Hangelbroek, F. J. Narcowich, and J. D. Ward, *Kernel approximation on manifolds I: Bounding the Lebesgue constant*, SIAM Journal on Mathematical Analysis **42** (2010), no. 4, 1732–1760.

[JSW99] Kurt Jetter, Joachim Stöckler, and Joseph D. Ward, *Error estimates for scattered data interpolation on spheres*, Mathematics of Computation **68** (1999), no. 226, 733–747.

[KBU02a] J. Kybic, T. Blu, and M. Unser, *Generalized sampling: A variational approach — Part I: Theory*, IEEE Trans. Signal Proc. **50** (2002), 1965–1976.

[KBU02b] _____, *Generalized sampling: A variational approach — Part II: Applications*, IEEE Trans. Signal Proc. **50** (2002), 1977–1985.

[Mad06] W. R. Madych, *An estimate for multivariate interpolation II*, Journal of Approximation Theory **142** (2006), no. 2, 116–128.

[MP85] W. R. Madych and E. H. Potter, *An estimate for multivariate interpolation*, Journal of Approximation Theory **43** (1985), no. 2, 132–139.

References V

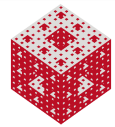
[MR77] C. A. Micchelli and T. J. Rivlin, *A survey of optimal recovery*, Optimal Estimation in Approximation Theory (Proc. Internat. Sympos., Freudenstadt, 1976) (C. A. Micchelli and T. J. Rivlin, eds.), Plenum Press (New York), 1977, pp. 1–54.

[MR80] _____, *Optimal recovery of best approximations*, Resultate Math. **3** (1980), no. 1, 25–32.

[MR85] _____, *Lectures on optimal recovery*, Numerical analysis, Lancaster 1984 (P. R. Turner, ed.), Springer (Berlin), 1985, Lecture Notes in Math., 1129, pp. 21–93.

[MRW76] C. A Micchelli, T. J. Rivlin, and S. Winograd, *The optimal recovery of smooth functions*, Numer. Math. **26** (1976), no. 2, 191–200.

[NP36] J. Neyman and E. S. Pearson, *Contributions to the theory of testing statistical hypotheses*, Statistical Research Memoirs **1** (1936), 1–37.



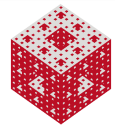
References VI

[NWW05] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comp. **74** (2005), 743–763.

[NWW06] _____, *Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions*, Constr. Approx. **24** (2006), no. 2, 175–186.

[Rie08] Christian Rieger, *Sampling Inequalities and Applications*, Ph.D. thesis, Universität Göttingen, 2008.

[RSZ10] Christian Rieger, Robert Schaback, and Barbara Zwicknagl, *Sampling and stability*, Mathematical Methods for Curves and Surfaces (Morten Dæhlen, Michael Floater, Tom Lyche, Jean-Louis Merrien, Knut Mørken, and Larry L. Schumaker, eds.), Lecture Notes in Computer Science, vol. 5862, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 347–369.



References VII

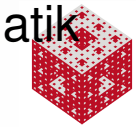
[RZ08] Christian Rieger and Barbara Zwicknagl, *Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning*, Advances in Computational Mathematics **32** (2008), no. 1, 103–129.

[RZ14] _____, *Improved exponential convergence rates by oversampling near the boundary*, Constructive Approximation **39** (2014), no. 2, 323–341.

[Sch93] R. Schaback, *Comparison of radial basis function interpolants*, Multivariate Approximation: From CAGD to Wavelets (Kurt Jetter and Florencio Utreras, eds.), World Scientific Publishing (Singapore), 1993, pp. 293–305.

[Wen05] H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, vol. 17, Cambridge University Press, 2005.

[WR05] Holger Wendland and Christian Rieger, *Approximate interpolation with applications to selecting smoothing parameters*, Numerische Mathematik **101** (2005), no. 4, 729–748.



References VIII

[WS93] Z.-M. Wu and R. Schaback, *Local error estimates for radial basis function interpolation of scattered data*, IMA Journal of Numerical Analysis **13** (1993), no. 1, 13 –27.

[ZS13] Barbara Zwicknagl and Robert Schaback, *Interpolation and approximation in Taylor spaces*, Journal of Approximation Theory **171** (2013), 65–83.