Stable multiquadric approximation by local thinning: schemes of the algorithms

Licia Lenarduzzi

Istituto di Matematica Applicata e Tecnologie Informatiche del C. N. R. Sez. Milano, via Bassini 15, 20133 Milano (Italy). licia@mi.imati.cnr.it

Summary. We consider the case that the functional evaluations are relevant to a very regular function and that the configuration of the data determines bad conditioning of the interpolation matrix relevant to the multiquadric basis.

By shifts of the multiquadric basis we construct a global least squares approximant with matrix of maximal rank. The choice of the centers is done by determining good local conditions everywhere on the domain such that there is a good global condition too.

The method is robust. In fact, with different modalities of sampling, it provides accurate approximating functions; we refer to the examples shown in [2].

Here we give the details of the algorithms for three different situations that occur in the applications:

• data sites uniformly scattered
• data sites from a non-uniform density depending on the features of the function
• cluster sampling.

Keywords: scattered data, thinning, arbitrary distribution, non-stationary multiquadric approximant

AMS classification: 65D10, 41A05

1 Introduction

It is well known that shifts of the multiquadric (MQ): \( \phi_\delta(r) = \sqrt{r^2 + \delta^2} \), with \( r \) the Euclidean distance, can provide very accurate interpolants of very smooth functions, see for example [9]. In particular the approximation can feature spectral convergence as the points get more dense, when the unknown function is very regular (it was proved in [4] and in [14]).

Moreover the parameter \( \delta \) plays a significant role, because, as already stressed in [10] and in [1], it allows to have a basis more adherent to the unknown function; in addition, it has influence both on the accuracy and on the condition of the interpolation matrix.

For this reason an efficient technique has been studied in order to provide a method with good accuracy and stability for a sample of standard size (we mean large enough not to miss any salient feature of the function).
This technique consists in solving a global least squares problem of collocation matrix $A_{X,T}$ where $T$ denotes the center set.

The centers are obtained by a procedure that involves the computation of the conditions of some local matrices of small size.

In fact, while it is known that the index $K_2(A_{X,X})$ gets worse the smaller is the minimal distance $q_X$ among the data sites, experimentally one observes that, for a given value of $q(X)$, the condition of $A_{X,X}$ depends on the configuration of the data points too. Furthermore a bad local configuration implies a bad global condition.

In [2] we have described the technique to calculate the approximating function and we have provided some examples to show the validity of the procedure and we have compared them with the results by some known methods of the literature; in the current report the details of the algorithm to determine the set $T$ are given.

For the case of scattered data, an initial check on the size of the center set is described in §3; then a local thinning is sketched in §4 and the selection of thinning thresholds is described.

The details of the thinning process are in §5.

In §6 the case of clustered data is treated, because the algorithm is more simple in this case.

In §7 an example of scattered data with varying density is presented.

2 Preliminaries

2.1 Notations

Given a function $f \in C^\alpha(D), \ \alpha > 1$ and $D \subset \mathbb{R}^2$, let $(X, f_X)$ a sample of size $N$. The points of $X$ are pairwise distinct and we denote by $q(X)$ the minimal distance among the $X$ locations.

We consider the linear space spanned by the multiquadric functions

$$\{ \phi(\| \cdot - x_i \|, \delta) := \sqrt{\| \cdot - x_i \|^2 + \delta^2}; \ x_i \in X \},$$

with the parameter $\delta$ fixed. We denote $A_{X,X} := \{ \phi(\| x_i - x_j \|, \delta) \}_{x_i, x_j \in X}$ the matrix of interpolation and $K_2(A_{X,X})$ its spectral condition number. Our hypothesis is that the rank of $A_{X,X}$ as calculated by MatLab is not maximal (in the following, to be short, we shall say that there is a warning about $A_{X,X}$).

Let $T$ a center set on $D$ of size $M$ less or equal than $N$. The points of $T$ are pairwise distinct. We denote by $s_{X,T}$ the least squares approximant of $(X, f_X)$ with centers at $T$.

In order to obtain centers better separated, according to the theory [18], we adjust the covering radius of $T$ on $X$ in the $l_2$ sense; we mean that: for each point $P_i \subset T$ we consider the set

$$X(P_i) := V_T(P_i) \cap X,$$

being

$$V_T(P_i) = \{ x \in \mathbb{R}^2 : \min_{P_i \in T} \| x - P_i \| = \| x - P_i \| \}$$
the Voronoi tile of $P_i$ with respect to $T$ and we compute the barycenter of $X(P_i)$. This barycenter updates $P_i$ in $T$.

In the following we take $D = [0, 1]^2$; the examples of [2] were run with $\delta = 0.35$ in a non-stationary way. We choose $\delta = 0.35$ because, in our experience, this value is suitable to give a good smoothness and a small error.

2.2 Thinning

Thinning algorithms refer to different strategies for point removal from a set $T$ of locations. In the book by Iske, [11], there is a comprehensive description. We recall Algorithm 16 (Thinning) from §4.2 of [11]:

**INPUT:** $T$ with $|T| = M$, and $\{n = 1, \ldots, M - 1\}$;

1. Let $T_M = T$;
2. For $k = 1, \ldots, n$
   1. Locate a removable point $x \in T_{M-k+1}$;
   2. Let $T_{M-k} = T_{M-k+1} \setminus x$;

**OUTPUT:** $T_{M-n} \subset T$, of size $|T_{M-n}| = M - n$.

To select a specific strategy, it is necessary to give a definition for a removable point in (2a) of the Algorithm 16. The different classical thinning strategies tend to favour well distributed planar point sets and lead to construct a multiresolution representation of the given $T$:

$$T_{M-n_L-1} \subset T_{M-n_L-2} \subset \ldots \subset T_{M-n_1} \subset T$$

to be used for a multilevel approximation scheme.

The strategy of thinning that we refer to is the old strategy of [7]: a removable point is that one that minimizes the distance to its nearest neighbour in the current subset $T_{M-k} \subset T$. Then, by taking $n = M$ in Algorithm 16, we determine a sorting of the $T$ points on the basis of the order according to which they are removed; as $k$ increases, the subset $T_{M-k} \subset T$ of points not yet removed has increasing uniformity.

We denote by

$$Y := \{y^{(1)}, \ldots, y^{(M)}\}$$

the points of $T$ rearranged by such a sorting; we denote thinning$^T$ this strategy of thinning to stress the set on which it operates.

We shall make use also of another sorting of the points of $T$. The distances of a point from its contiguous points within the Delaunay triangulation $D(T)$ are considered (see [15] about the Delaunay triangulation; two points are contiguous in the sense that they are connected by an edge of the triangulation); these distances are averaged. We denote $Y^{D(T)}$ the set $T$ rearranged for increasing average distance.

Points in regions with higher density of location come first in the order: this order allows a selection more robust than the order of $Y$, because of which a point can be first because it belongs to an area with low density but, by chance, with just one neighbour at distance $q(T)$.

We have introduced the notation thinning$^T$ to stress the difference from a new strategy of thinning that we are going to present and according to which the points
of $T$ turn out to be ordered depending on the condition of the local matrix of interpolation with $\phi$; we shall denote the new strategy $\text{thinning}^{T,\phi}$ to stress the set on which it operates but also the dependence on the basis function; by running Algorithm 16 according to $\text{thinning}^{T,\phi}$, as $n$ increases, the subset $T_{M-n} \subset T$ of points not yet removed will be characterized by the fact that the value of the condition is not large on each local subset $S_r \subset T_{M-n}$ of small size.

3 Check on the size of the center set for the least squares approximation of scattered data

It is well known that centers on an hexagonal grid determine the maximal $q_T$ for assigned fill distance $h_T$. Anyway, having fixed $\delta$, there is a cardinality of centers $M$ such that $A_{X,T}$ gives a warning. For such a reason, having fixed $\delta$, we have calculated the value of $M_\delta$ in the following way:

Being assigned $M = 1600$ data points $X$ scattered on $D$, as center set $T$ we have taken the restriction to $D$ of a grid of equilateral triangles with edge $q(T)$.

In the case $\delta = 1$ we have experimented a warning for the matrix $A_{X,T}$ of $M = 72$ regular centers of the hexagonal grid with $q(T) = 0.13$, while the matrix $A_{X,T}$ for $q(T) = 0.1325$ that still leads to $M = 72$ centers has no warning. We put $M_1 = 72$.

In the case $\delta = 0.35$ we have experimented a warning for the matrix $A_{X,T}$ relevant to centers on the hexagonal grid with $q(T) = 0.062$ in number $M = 314$ (with value of condition equal to 3.27 $\epsilon(12)$), while the one relative to centers with $q(T) = 0.063$ and in number $M = 304$ has no warning. So we choose $M_{0.35} = 314$ to maintain as many centers as possible.

When there is no warning for the regular grid, we have noticed that for small perturbations of the same regular grid, still there is no warning while for larger perturbations (that make the value of $q(T)$ much lower) we enter in the region of warning. Therefore it is natural, when the size $M$ of the center set is greater than some suitable $M_\delta$ that indicates the fuzzy transition from the region of warning to the region of no warning, to look for a set $T$ of size $M_\delta$ of locations fairly regularly distributed on $D$, because it might be that it gives no warning for $A_{X,T}$.

3.1 Determination of $T$ of size $M_\delta$ from data sites uniformly distributed

We make use of the simplest scheme $\text{thinning}^X$ of [7] to obtain the set $T = X_{M_\delta}$ of size $M_\delta$. As we do not have to satisfy the constraint that $T$ is a subset of $X = X_{M_\ell}$, we improve its covering radius, measured in a $l_2$ sense, on $X$.

If there is a warning for $A_{X,T}$, the set $T$ is the input $T_0$ to $\text{thinning}^{T,\phi}$ (see §4 and §5).

3.2 Determination of $T$ of size $M_\delta$ for $X$ distributed according to the functional features

From $T = X$ we discard locations at random till the size $M = M_\delta$ is reached.
Stable multiquadric approximation

This set \( T \) matches the density of \( X \) and very likely \( A_{X,T} \) is not stable (that is why we performed the quite rough and fast thinning at random, without taking much care in placing the centers apart).

However we perform the step of improving the covering radius of \( T \), measured in a \( l_2 \) sense, on \( X \) and, very likely, we must perform thinning \( T^{\phi} \) of \( \xi \).

So if there is a warning for \( A_{X,T} \), the set \( T \) is the input \( T_0 \) to thinning \( T^{\phi} \) (see \( \xi \) and \( \xi \)).

For both the \( X \) configurations: if there is no warning for \( A_{X,T} \), \( s_{X,T} \) provides our recovering of \( f \). On the contrary, if there is warning for \( A_{X,T} \), we apply to \( T \) the procedure thinning \( T^{\phi} \) of \( \xi \), and we calculate \( s_{X,T} \) on the updated set \( T \).

4 Thinning \( T^{\phi} \)

Let the input be \( T = T_0 \) of size \( M_0 \) in the case \( N \geq M_0 \); while in the case \( N < M_0 \) we take \( T_0 = X \). We call \( M_0 \) the size of \( T_0 \). We assume that there is a warning for \( A_{X,T_0} \). We thin \( T_0 \) to a new \( T \) by looking at the local conditions to remove those bad configurations that determine the bad global condition; this is an effort to be efficient and at the same time to maintain the size of \( T \) close to the size of \( T_0 \).

Before presenting the algorithm, we must specify that, for an assigned threshold, and for a \( S \subset T \), with “thinning \( S^{\phi}(\text{threshold}) \)”, we shall mean that: within \( S \) we select points of \( T \) according to the sorting of the vector \( Y \), defined in (1), to be removed such that the condition of the local interpolation matrix is less than or equal to \( 5 \cdot \text{threshold} \).

With “thinning \( T^{\phi}(\text{threshold}) \) within \( V_k = \cup_{i=1}^{j(k)} S_i(k) \)”, being \( V_k \subset T \), we shall mean that: within each local \( S_i(k) \) we select points of \( T \) according to the sorting of the vector \( Y \) to be removed such that the condition of each local interpolation matrix is less than or equal to \( 5 \cdot \text{threshold} \).

4.1 Determination of the thresholds

Let \( S_1 \) and \( S_2 \) be two disjoints local sets of data sites of size \( m(1) \) and \( m(2) \) respectively.

• For \( S_1 \) compute the condition number of the submatrices \( \{ A^{(\mu)}_{S_1,S_1} \} \) excluding a number \( \mu \) of points of \( S_1 \), chosen according to the ordering of \( Y \).

Let \( A \) be the set

\[
A := \{ K_2(A^{(\mu)}_{S_1,S_1}) \}_{\mu=0}^{m(1)-1} := \{ a_{\mu} \}_{\mu=0}^{m(1)-1},
\]

\( a_{\mu} \geq a_{\mu+1} \).

• In analogous way, for \( S_2 \), compute

\[
B := \{ K_2(A^{(\mu)}_{S_2,S_2}) \}_{\mu=0}^{m(2)-1} := \{ b_{\mu} \}_{\mu=0}^{m(2)-1},
\]

\( b_{\mu} \geq b_{\mu+1} \) condition of the local matrix of interpolation excluding a number \( \mu \) of points of \( S_2 \), chosen according to the ordering of \( Y \).
- For an element \( a_i \) of the set \( A \), and an element \( b_j \) of the set \( B \), we establish a relationship of similarity: the similarity is larger the smaller is the difference \( | a_i - b_j | \).

  On the basis of such a relationship the set \( A \) is split into as many subsets as possible: \( A = \bigcup_{i=1}^{t} A_i \) (in an analogous way the set \( B \) is split: \( B = \bigcup_{i=1}^{t} B_i \)) such that:

  each element of \( A_i \) has a corresponding element in \( B_i \) to which is the most similar; vice versa each element of \( B_i \) has a corresponding element in \( A_i \) to which is the most similar.

  For each \( i = 1, \ldots, l \):

  Let \( a_r > a_{r+1} > \ldots > a_{r+n_i} > \ldots > a_s \) be the elements of \( A_i \) and \( b_1 > b_{1+i} > \ldots > b_{1+n_i} > \ldots > b_1 \) the elements of \( B_i \).

  If in \( A_i = \{a_r, \ldots, a_{r+n_i}, \ldots, a_s\} \) a value \( a_{r+n_i}/a_{r+n_i+1} \) is too large (or in \( B_i \), a value \( b_{1+n_i}/b_{1+n_i+1} \) is too large) split \( A_i \) (respectively \( B_i \)) in two subsets where the ratio is too large; then split also \( B_i \) (respectively \( A_i \)) in two subsets where the maximal ratio occurs, such that \( A \) and \( B \) are partitioned in the same number of subsets.

  To fix the ideas, we refer to the values \( A \) for a set \( S_1 \) of \( m(1) = 5 \) data:

  \[
  A = \{3.24 \ e(10), 4.69 \ e(9), 1.38 \ e(5), 0.61 \ e(5), 1\}
  \]

  and to the values \( B \) for a set \( S_2 \) of \( m(2) = 11 \) data:

  \[
  B = \{9.77 \ e(11), 5.91 \ e(9), 1.44 \ e(9), 4.14 \ e(8), 2.58 \ e(8), 1.30 \ e(8), 1.8 \ e(7), 2 \ e(6), 4 \ e(5), 46 \ e(0), 1\}
  \]

  The partition on the basis of the relation of similarity is

  \[
  A_1 = \{3.24 \ e(10), 4.69 \ e(9)\};
  \]

  \[
  A_2 = \{1.38 \ e(5), 0.61 \ e(5)\};
  \]

  \[
  A_3 = \{1\}
  \]

  \[
  B_1 = \{9.77 \ e(11), 5.91 \ e(9)\};
  \]

  \[
  B_2 = \{1.44 \ e(9), 4.14 \ e(8), 2.58 \ e(8), 1.30 \ e(8), 1.8 \ e(7), 2 \ e(6), 4 \ e(5)\};
  \]

  \[
  B_3 = \{46 \ e(0), 1\}.
  \]

  For example \( A_1 \) and \( B_1 \) cannot be split further on the basis of the relation of similarity, because the element most similar to \( 3.24 \ e(10) \) is \( 5.91 \ e(9) \).

  On the other hand \( 3.24 \ e(10)/4.69 \ e(9) \) and \( 9.77 \ e(11)/5.91 \ e(9) \) are not so large.

  Now we define a set of thresholds more and more selective: we put \( i = 1 \) and we consider \( \text{threshold} := \min \{a_r, b_1\} \), with \( a_r \in A_1 \) and \( b_1 \in B_1 \);

  for increasing \( i \), we put \( \text{threshold} := \max \{a_r, b_1\} \), with \( a_r \in A_i \) and \( b_1 \in B_i \) the first time and then we put \( \text{threshold} := \min \{a_r, b_1\} \).

  Relevant to the current example, the first choice of threshold (the least selective) is for \( i = 1 \) and it is \( \text{threshold} = 3.24 \ e(10) \) that implies to throw one point of \( S_2 \); then for \( i = 2 \) it is \( \text{threshold} = 1.44 \ e(9) \) that implies to throw one point of \( S_1 \) and one point of \( S_2 \); then, for \( i = 2 \), it is \( \text{threshold} = 1.38 \ e(5) \) that implies to throw two points of \( S_1 \) and eight points of \( S_2 \), and so on.

  \textbf{Remark:} there can be some computational saving in this way: compute only \( a_r \) of \( A_1 \) and \( b_1 \) of \( B_1 \); then, if there is a warning for \( A_{X,T} \), we compute all the subsets of \( A \) and of \( B \).
4.2 Computation of the sets $S_1$ and $S_2$

• In the following, for a set $I$ of indices of points relevant to a set $P$, we shall indicate the subset of the points of $P$ with index in $I_1 \subset I$ with $P(I_1)$.

• $J$ indicates the set $\{1, \ldots, M_0\}$.

For assigned $T$, we consider $Y^D(T)$ the rearrangement of its points according to the increasing average distance.

⋄ Construction of the two sets $S_1$ and $S_2$ of size $m = 20$ (the closest $m$ points) with centers respectively at the first element of $Y^D(T)$, let us say $z_1$, and at the first element $\in Y^D(T) \setminus S_1$, let us say $z_2$, in zones dense of points.

• We individuate $z_1$ and we indicate $S_1$ the set of the $m$ points $t_i \in T$ closest to $z_1$ ($z_1$ included), with respect to dist$_\infty$. Let $R_1$ be the distance between $z_1$ and its $(m - 1)$-th closest point.

• We indicate $z_2$ the first element in $Y^D(T)(J^1)$; let $S_2$ the set of the $m$ points $t_i \in T$ closest to $z_2$ ($z_2$ included), with respect to dist$_\infty$. Let $R_2$ be the distance between $z_2$ and its $(m - 1)$-th closest point.

4.3 Details about the construction of a next subset $S_i(k)$

• Every other set $S_i$ (that we shall indicate $S_i(k)$ if considered at step $k$ of the algorithm of §5), that we shall consider on the current subset $U \subset T_0$, is of size $m = 7$ and it is taken far from those already taken at the same step $k$.

In detail: let $J$ the cardinality of $U$; if we have already taken $\{S_j(k)\}_{j=1}^{r-1}$ of centers $\{z_j(k)\}_{j=1}^{r-1}$, we define $J^{r-1}$ as the subset of $J$ of the points $u_i(k)$ of $U = U(k)$ such that the following inequalities hold:

$$\text{dist}_\infty(u_i(k), z_j(k)) \geq 2R_j \quad j = 1, \ldots, r - 1.$$ 

We consider $S_r$ centered at $z_r$, the first element in $Y(J^{r-1})$.

• We repeat by increasing $r$ till a value $l$ is reached that is relevant to the set $J^l$ empty.

5 Algorithm for scattered data

⋄ Inizialization:

$T = T_0$;

let $D(T)$ its Delaunay triangulation.

• Calculate the sets $Y$ and $Y^D(T)$ derived from sortings of $T$.

• Locate $S_1 \subset T$ and $S_2 \subset T$ in dense regions of $T$ (see §4.2); sort the points of $S_1$ according to thinning$^{S_1}$ and sort the points of $S_2$ according to thinning$^{S_2}$;

• Computation of the vectors $A$ and $B$ (see §4.1)

• Computation of the set of the thresholds to be applied to the local conditions, see §4.1;
Main Loop:

While there is a warning about \( \text{rank}(A_X,T) \) and for a choice more and more selective of the threshold, perform:

- step \( k = 1 \)  
  \( T := T_0 \)  
  \( U := T_0 \)
  - on \( U \setminus (S_1 \cup S_2) \) we consider other sets \( S_i(k) \), \( i = 3, \ldots, l(k) \) centered at points removable before according to \( Y \); at the same time the \( \{S_i(k)\}_{i=1}^{l(k)} \) are scattered on \( T \) and almost not overlapping, see §4.3.

  We define \( V_k = \bigcup_{i=1}^{l(k)} S_i(k) \).

  - we run thinning \( T,\phi \) (threshold) within \( V_k \) to thin \( T \).

  - \( U = U \setminus V_k; \ k = 2 \).

- step \( k \)
  - if \( U \) turns out to be empty, then →  
    adjust the covering of \( T \) on \( X \),
    compute \( \text{rank}(A_X,T) \)
    break the loop on \( k \); go to the Main Loop

  - Otherwise we consider sets \( S_i(k) \), \( i = 1, \ldots, l(k) \), centered at points of \( U \) which are removable sooner according to \( Y \); at the same time the \( S_i(k) \) are scattered on \( U \) and almost not overlapping, see §4.3.

  We define \( V_k = \bigcup_{i=1}^{l(k)} S_i(k) \).

  - We run thinning \( T,\phi \) (threshold) within \( V_k \) to thin \( T \).

  - If, with the current step, no points of \( T \) were removed, then →  
    adjust the covering of \( T \) on \( X \),
    compute \( \text{rank}(A_X,T) \)
    break the loop on \( k \); go to the Main Loop

  - Otherwise put \( U = U \setminus V_k; \ k = k + 1; \) go to step \( k \).

5.1 Illustration of the scheme

We show how the algorithm works for data scattered but with varying density on \( D \) by two figures.

In Fig. 1 you see the locations of the data and the center \( z_1 \) of \( S_1 \) as a square; as expected it is placed where the density is high. One must know that there are two data locations extremely close between themselves, with locations \((0.0384, 0.5955)\) and \((0.0384 + 1 \ e(-9), 0.5955)\) respectively; the other locations are well separated.

About the determination of the threshold: \( A \) turns out to be partitioned in \( A_1 \) with elements of the order \( 1 \ e(10) \), \( A_2 \) with elements whose order is between \( 1 \ e(9) \) and \( 1 \ e(10) \) and \( A_3 \) with elements of the order \( 1 \ e(7) \); there is no need to split within any \( A_i \). Similarly for \( B \).

The values of the threshold more and more selective that the algorithm has worked out are: \( 1.53 \ e(10) \), \( 0.9 \ e(9) \), \( 0.4 \ e(9) \), \( 0.153 \ e(9) \).

In Fig. 2 you see the sets \( S_i(1), i = 1, \ldots, l(1) \) framed. The data locations circled are those removed within \( V_1 = \bigcup_{i=1}^{l(1)} S_i(1) \). At step \( k = 2 \) no further points will be removed.

5.2 Computational cost

An upper bound for the cost of thinning \( X \) the set \( X \) down to \( T_0 = X_{M_X(\phi)} \) is \( O(N\log(N)) \), see [11].
Fig. 1. Data sites and point \( z_1 \) squared.

For a center set of size \( M_0 \leq M_{\delta(\phi)} \), we have a preprocessing phase:

- the calculation of the Delaunay triangulation that can be performed with a cost \( O(M_0 \log M_0) \) (see §2.2 in [11]).
- the construction of a heap (standard data structure) on \( Y^{D(T_0)} \): at the node of the heap relevant to \( x_i \), we associate the indices of the neighbours in the Delaunay triangulation and the significance of the node is provided by the average distance from the neighbours.
- the construction of a heap on \( Y(T_0) \): at the node of the heap relevant to \( t_i \), we associate the distance from its nearest neighbour in \( T_0 \) as significance.

Each update of a heap of \( M_0 \) nodes costs \( O(\log M_0) \).

- we individuate about \( N_s := M/m \) sets \( \{S_r\} \). On \( S_1 \) and \( S_2 \) we must perform \( O(m^4/12) \) operations with \( m = 20 \) for the linear algebra; on each \( S_r \), \( r > 3 \) usually we must calculate just few spectral condition numbers and so the computational cost is \( O(m^3/3) \) with \( m = 7 \).
We must calculate the rank of the global matrix $K$ times with cost $O(N \cdot M^2 \cdot K/3)$ for $K$ less or equal than 4 in our experience.

Therefore the total cost, to individuate $T$ and then to determine the coefficients for $sX,T$, is of the order $O(M_0 \log M_0) + O((m = 20)^4/12) + O((N_s - 2) \cdot (m = 7)^3/3) + O(N \cdot M^2 \cdot K/3)$, plus eventual cost $O(N \log N)$.

6 Clustered data

6.1 Clusters: check on the size of the center set

We do the following assumptions:

- clusters of about the same size of the order of ten
• clusters well separated: the distance between any two clusters definitely larger than the distances within a cluster.

To calculate $M_3$ we restrict the grid of equilateral triangles to the regions of the clouds.

6.2 Thinning $T, \phi$ for clusters of points $X = \bigcup_{r=1}^{\alpha} S_r$.

• In the case $N \geq M_3$ we take $X_{M_3}$ by thinning $X$ the set $X$ and by improving the covering radius of $X_{M_3}$ on $X$, we obtain the set $T_0$ to be provided in input. In the case $N < M_3(\phi)$ we take $T_0 = X$. We assume that there is a warning about $A_{X,T_0}$.

• Let the input be $T = T_0$; calculate the ordering $Y^{D(T)}$ and the ordering $Y$.

• individuate the cluster $S_j$ to which the first element of $Y^{D(T)}$ belongs; denote $n(j) := \#S_j$; sort the points of $S_j$ according to thinning $S_j$.

• individuate the cluster $S_k$ in $T \setminus S_j$ of size $n(k)$ to which the first element of $Y^{D(T)} \setminus S_j$ belongs; calculate the set of the thresholds to be applied to the local conditions, as specified in §4.1.

• For a choice more and more selective of the threshold and while rank$(A_{X,T})$ is not maximum
  • Run thinning $(T, \phi)(threshold)$ within $\bigcup_{r=1}^{\alpha} S_r$ to thin $T$.
  • adjust the covering of $T$ on $X$.
  • compute rank$(A_{X,T})$

6.3 Computational cost

We have a preprocessing phase:

• the upper bound for the cost of the eventual thinning $X$ down to $T_0$ is $O(N \log N)$, see [11].

• the calculation of the Delaunay triangulation that can be run in $O(M_0 \log(M_0))$ operations, if $M_0$ is the size of $T_0$

• for the $i$–th cluster the construction of a heap structure of $n(i)$ nodes with associated distance from the nearest neighbour as significance.

Therefore the total cost to individuate $T$ and then to calculate the coefficients of $s_{X,T}$ is of the order

$$O(M_0 \log M_0) + O\left(\sum_{i=1}^{n} n(i)^{3/3}\right) + O(N \cdot M^2 \cdot K/3),$$

plus eventual cost $O(N \log N)$.

7 Example

Let us consider a sample of size $N = 809$ from a distribution with variable density depending on the behavior of the Franke’s function.

It is $M_3 = 314$. 
The set $T$ selected, by four iterations, is the one shown in Fig. 3 of size 234 and with $\mathcal{K}_2(A_{X,T}) = 6.6 \epsilon(12)$. The behaviour is shown in Fig. 4 and the error is

$$e_{\infty}(X,T) = 4.11 \epsilon(-4);$$

it is of the same order as that of the interpolant of $(X, f_X)$. Remark: the graphic of the interpolant turns out to be without undue oscillations in this case, despite the warning about the rank of $A_{X,X} (\mathcal{K}_2(A_{X,X}) = 1.5 \epsilon(20))$, a known situation described in [6] too, and so we take the value of the error of the interpolant as benchmark.

References

Fig. 3. $X$ dotted, $T$ circled.