THINNING ALGORITHMS IN SCATTERED DATA MODELLING

ARMIN ISKE*

Abstract. This short note intends to show the utility of thinning algorithms for multilevel scattered data approximation. To this end, we briefly review earlier work [2, 3], before we report on recent developments at Munich University of Technology.

1. Introduction. Scattered data modelling requires reconstructing an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ from a given data vector $D_Z(f) = (f(z_1), \ldots, f(z_N))^T \in \mathbb{R}^N$ of function values sampled from f at a finite set $Z = \{z_1, \ldots, z_N\} \subset \mathbb{R}^d$ of locations. Especially in situations where the number N is extremely large, and the points in Z are unevenly distributed, multilevel interpolation and approximation schemes are commonly used. These modern techniques usually require decomposing the data into a hierarchy

$$X_L \subset X_{L-1} \subset \cdots \subset X_1 \subset X_0 = Z \tag{1.1}$$

of nested subsets beforehand. Each of the subsets $X_j \subset Z$ will then represent a *level of detail* of the given data.

We shall use the term *thinning algorithm* for algorithms which serve to generate hierarchies of the form (1.1) by recursive point removals from Z. Likewise, the expression *thinning* is associated with the class of recursive point removal schemes.

2. Multilevel Interpolation. Let us briefly discuss how to use the hierarchy (1.1) for multilevel interpolation. According to the proposed scheme in [2], a sequence $s_L, \ldots, s_0$ of model functions is recursively computed as follows. We let $s_{L+1} \equiv 0$. For $j = L, \ldots, 0$ an interpolant Δs_j to the residual $D_{X_j}(f - s_{j+1})$ from the previous step is first computed, before we let $s_j = s_{j+1} + \Delta s_j$. Thus, at each step $L \geq j \geq 0$, by the update $s_j = s_{j+1} + \Delta s_j$ we obtain an interpolant to f on the set X_j, i.e.

$$D_{X_j}(s_j) = D_{X_j}(f), \quad 0 \leq j \leq L.$$

In particular, the function $s \equiv s_0$ interpolates the unknown function f on Z. The basic idea behind this multilevel scheme is to capture a global trend of the function f by the coarse approximation s_L, before finer details of f are gradually added by using the sequence of functions Δs_j. In other words, the sequence $s_L, \ldots, s_0$ of interpolants can be understood as a sequence of representations for f at $L+1$ different scales.

3. Thinning. It should be quite obvious that the performance of the multilevel interpolation scheme from the previous section heavily relies on the *quality* of the hierarchy (1.1). In [2], compactly supported radial basis functions were used for the modelling. Due to available error bounds for radial basis function interpolation it is desirable that the *difference*

$$d(X_{j+1}, X_j) = \max_{y \in X_j} \min_{x \in X_{j+1}} \|y - x\|, \quad 0 \leq j < L, \tag{3.1}$$

$\|\cdot\|$ being the Euclidean norm on $\mathbb{R}^d$, between two consecutive subsets is reasonably small. Loosely speaking, each subset $X_{j+1} \subset X_j$, $0 \leq j < L$, is supposed to represent its superset X_j reasonably well in the above sense.

*Technische Universität München, Zentrum Mathematik, Arcisstraße 21, D-80290 München, Germany (iske@mathematik.tu-muenchen.de).

For the purposes of this presentation we shall solely use the criterion (3.1) for the construction of the hierarchy (1.1). Note that (3.1) depends not on the function values $D_Z(f)$. Therefore, we shall here merely study a *data-independent* thinning strategy. A variant of *data-dependent* thinning is subject of [1].

From a rather abstract viewpoint on thinning, we associate each point removal scheme of such kind with a *thinning operator* T, which returns for any nonempty subset $X \subset Z$ one unique subset $T(X)$ of size $|T(X)| = |X| - 1$, i.e. T operates on X by removing one unique point from X. Accordingly, the n-fold composition $T^n = T \circ \ldots \circ T$, $1 \leq n \leq |X|$, returns a unique subset $T^n(X) \subset X$ of size $|X| - n$. Thus, by

$$T^n(Z) \subset T^{n-1}(Z) \subset \ldots \subset T^2(Z) \subset T(Z) \subset T^0(Z) = Z \tag{3.2}$$

we obtain a *dense* hierarchy of the form (1.1). We shall use nested sequences of the above form (3.2) for the construction of a hierarchy (1.1). In fact, having generated (3.2), it remains to select a sequence $n_L > n_{L-1} > \ldots > n_1 > 0$ of reasonable *breakpoints*, which gives us by $X_j = T^{n_j}(Z)$, $1 \leq j \leq L$, a *coarser* hierarchy (1.1). But the a priori selection of the breakpoints is, especially for the application from the previous section, a nontrivial task.

For controlling the relative error between the optimal value

$$d_n^*(Z) = \min_{\substack{X \subset Z \\ |X| = |Z| - n}} d(X, Z)$$

and the *current* difference $d(T^n(Z), Z)$ at run time, we shall make use of the bound

$$\sigma_n \leq d_n^*(Z) \leq d(T^n(Z), Z) \leq \alpha_n \cdot d_n^*(Z) \tag{3.3}$$

available from [5], where $\alpha_n = d(T^n(Z), Z)/\sigma_n$, with specific predetermined values $\sigma_n = \sigma_n(Z)$ (for details see [5]). To this end, we record the numbers α_n during the thinning. Whenever α_n is small, the difference $d(T^n(Z), Z)$ is by (3.3) close to the optimum $d_n^*(Z)$ in which case the number n is a good candidate for the breakpoint n_1. Having chosen n_1, we then store the subset $X_1 = T^{n_1}(Z)$ and proceed recursively by letting $Z = X_1$, i.e. the next subset X_2 in (1.1) will be generated by using thinning on X_1, and so on.

Now let us turn to specific choices for T. To this end, it will be sufficient to discuss the special case of one sublevel $L = 1$ in (1.1), i.e. we focus on the selection of merely one subset $X \equiv X_1 \subset Z$ from Z. In view of minimizing (3.1) for $j = 0$, we remark that the problem of finding an algorithm which outputs for any possible input pair (Z, n), $1 \leq n < |Z|$, an *optimal* subset $X^* \subset Z$ of size $|Z| - n$ satisfying $d(X^*, Z) = d_n^*(Z)$ is referred to as the *k-center problem* [4]. This covering problem is well-known to be NP-hard. Moreover, the problem of finding an *α-approximation algorithm* which outputs for any (Z, n) a subset $X \subset Z$ of size $|Z| - n$ satisfying

$$d(X, Z) \leq \alpha \cdot d_n^*(Z) \tag{3.4}$$

is for any $\alpha < 2$ NP-complete. Note that the bound (3.3) is similar to (3.4), but α_n in (3.3) depends on both Z and $T^n(Z)$, whereas the universal constant α in (3.4) does not even depend on Z. We expect to obtain tight bounds by using (3.3), but we defer details to later.

An appropriate answer to NP-complete optimization problems, such as the *k*-center problem, are *greedy* approximation algorithms. In our particular context, a thinning operator T_* is said to be greedy, if for any possible argument X, the operator T_* minimizes the difference $d(T_*(X), X)$. In this sense, T_* achieves to make the best possible choice for the removal of *one* point. Note that a thinning operator is greedy, if and only if by its action on any possible X, one point $x \in X$ which minimizes the Euclidean distance to its nearest neighbour in X is removed.

4. Exchange. In order to balance the short-sightedness of greedy thinning, we have in [5] used an additional ingredient for the construction of (1.1). The basic idea is, for any given $X \subset Z$, to repeatedly exchange point pairs between X and $Y = Z \backslash X$, whenever the exchange pays off in terms of the reduction of $d(X, Z)$. We say that a point pair $(x, y) \in X \times Y$ is *exchangeable*, iff

$$d(X, Z) > d((X \setminus x) \cup y, Z).$$

Moreover, the subset $X \subset Z$ is said to be *locally optimal* in Z, iff there is no exchangeable pair $(x, y) \in X \times Y$.

We shall interpret the exchange of one point pair as an operation on X by an *exchange operator* E, defined as

$$E(X) = \begin{cases} X & \text{if } X \text{ is locally optimal;} \\ (X \setminus x) \cup y & \text{for one specific exchangeable pair } (x, y) \in X \times Y. \end{cases}$$

We fix the operator E and use the notation E^n for the n-fold composition of E. Observe that for any subset X, there is a unique index $n = n(X) < \infty$, such that

$$E^{n+\ell}(X) = E^n(X) \neq E^{n-1}(X) \text{ for all } \ell \geq 0,$$

in which case the set $E^n(X)$ is locally optimal in Z. Indeed, this is because the sequence $\{d(E^k(X), Z)\}_{k=1}^n$ of nonnegative differences is strictly decreasing and Z was assumed to be finite. Therefore, the operator $E^\infty = E^{n(\cdot)}$ is a *projector* onto the locally optimal subsets of Z.

In [5], we have used exchange in combination with greedy thinning for the construction of *filters*, operators which achieve to select locally optimal subsets of Z. Among others, we have been working with filters of the form $F_n = E^\infty \circ T_*^n$ in order to construct point hierarchies (1.1) by letting $X_{j+1} = F_{n_j}(X_j)$, $0 \leq j < L$.

5. Model Problem. For the special case of two dimensions, i.e. $d = 2$, we have implemented the recursive filtering scheme given by F_n. We considered using the data set *Hurrungane* displayed in Figure 5.1 (top left) as a model problem. The following Table 5.1 reflects our results. In this table, T-CPU stands for the computational costs (in CPU units) consumed for computing the set $T_*^{n_j}(X_j)$ by using thinning, and X-CPU denotes the additional costs required for computing X_{j+1} by the subsequent exchanges.

REFERENCES

[1] N. Dyn, M. S. Floater, and A. Iske, Data-dependent Thinning, in preparation.
[2] M. S. Floater and A. Iske, *Multistep Scattered Data Interpolation using Compactly Supported Radial Basis Functions*, J. Comp. Appl. Math. **73** (1996), 65-78.
[3] M. S. Floater and A. Iske, Thinning algorithms for scattered data interpolation, BIT **38:4** (1998), 705-720.
[4] *Approximation algorithms for NP-hard problems*, D. S. Hochbaum (ed.), PWS Publishing Company, Boston, 1997.
[5] A. Iske and St. Pöhn, Adaptive Scattered Data Filtering, in preparation.

TABLE 5.1
Hierarchical filtering using F_n on Hurrungane.

j	$\|X_j\|$	n_j	$\alpha(T_*^{n_j}(X_j), X_j)$	$\alpha(X_{j+1}, X_j)$	T-CPU	X-CPU	$n(X_j)$
0	23,092	3899	4.53724	1.25576	24.31	98.18	1729
1	19,193	4932	4.59619	1.29533	26.98	130.03	2294
2	14,261	4904	4.11624	1.30673	23.54	132.88	2166
3	9,357	4222	3.96326	1.37028	15.74	74.44	1681
4	5,135	3523	4.61398	1.86060	10.51	17.68	721

FIG. 5.1. **The point hierarchy $X_{j+1} = F_{n_j}(X_j)$, $0 \le j < L = 5$.**