BOUNDS FOR APPROXIMATE DISCRETE TOMOGRAPHY SOLUTIONS

LAJOS HAJDU AND ROB TIJDEMAN

Abstract. In earlier papers we have developed an algebraic theory of discrete tomography. In those papers the structure of the functions $f : A \to \{0, 1\}$ and $f : A \to \mathbb{Z}$ having given line sums in certain directions have been analyzed. Here A was a block in $\mathbb{Z}^n$ with sides parallel to the axes. In the present paper we assume that there is noise in the measurements and (only) that A is an arbitrary or convex finite set in $\mathbb{Z}^n$. We derive generalizations of earlier results. Furthermore we apply a method of Beck and Fiala to obtain results of the following type: if the line sums in k directions of a function $h : A \to [0, 1]$ are known, then there exists a function $f : A \to \{0, 1\}$ such that its line sums differ by at most k from the corresponding line sums of h.

1. Introduction

Let n be a positive integer and let A be a finite subset of $\mathbb{Z}^n$. Further, let S be a finite set of directions. By a discrete tomography problem we mean asking for a function $f : A \to Z$ which satisfies prescribed line sums along the directions in S, where Z may be $\{0, 1\}, \mathbb{R}, \mathbb{Q}, \mathbb{Z}$ or some finite real set. The authors and others have developed an algebraic theory of the structure of the solutions of a discrete tomography problem, see [9], [10], [11], [12], [8], [14], [15], [6], [3]. It turns out that the real solutions of a discrete tomography problem form a linear manifold if there is at least one real solution, and that the integer solutions form a lattice in this linear manifold, provided that at least one integer solution exists.

2010 *Mathematics Subject Classification.* 94A08, 15A06.

Key words and phrases. Discrete tomography, approximate solutions.

Research supported in part by the OTKA grants K75566, K100339 and NK101680, and by the TÁMOP 4.2.1./B-09/1/KONV-2010-0007 and TÁMOP-4.2.2.C-11/1/KONV-2012-0001 projects. The projects are implemented through the New Hungary Development Plan, supported by the European Union, co-financed by the European Social Fund and the European Regional Development Fund.

Theorem 1 of [12] states that if A is a hyperblock with sides parallel to the axes, then any function $f : A \to Z$ with $Z = \mathbb{Q}$ or $\mathbb{Z}$ having zero line sums along the directions in S can be uniquely written as a linear combination of so-called switching components of S contained in A. In fact the statement and proof remain valid also for functions $f : A \to \mathbb{R}$. In Section 3 we prove that it suffices that $A \subseteq \mathbb{Z}^n$ is convex, but that the convexity requirement cannot be dropped.

If the line sums are measured with some noise, then it is not certain that some function satisfies the measured line sums along S. A natural question is then what the best approximate solution is. We shall show that there is some linear manifold which can be considered as the set of 'best real approximations' in the sense of least squares. An obvious choice is then to choose the shortest best real approximation, that is the orthogonal projection of the origin to that linear manifold. In Section 4 we present an algorithm to construct this shortest best real approximation and illustrate it by an example. In Section 5 we present an explicit system of linear equations which determines the shortest best real solution in case A is convex. As an application we generalize a result from [6] by giving an explicit expression for the shortest best real solution in case A is a rectangle with sides parallel to the axes and only row and column sums are given.

In the 80's Beck and Fiala [4] proved a 'balancing' theorem. In Section 6 we show that this theorem implies that if the line sums in k directions of a function $h : A \to [0, 1]$ are known, then there exists a function $f : A \to \{0, 1\}$ such that its line sums differ by at most k from the corresponding line sums of h.

We extend this result in Section 7 to the case that we are not searching for a binary image, but for an image f with a finite number of given real values. To do so we generalize the result of Beck and Fiala.

2. Notation

We use the following notation throughout the paper. Let n be a positive integer. For brevity, for $x_1, \dots, x_n \in \mathbb{R}$ and $u_1, \dots, u_n \in \mathbb{Z}$, we write $\underline{x} = (x_1, \dots, x_n)$, $\vec{x} = (x_1, \dots, x_n)^T$ and $\underline{x}^{\underline{u}} = \prod_{j=1}^{n} x_j^{u_j}$.

Let $\underline{d} \in \mathbb{Z}^n$ with $\gcd(d_1, \dots, d_n) = 1$ be such that $\underline{d} \neq \underline{0}$, and for the smallest j with $d_j \neq 0$ we have $d_j > 0$. We call $\underline{d}$ a *direction*. By *lines with direction* $\underline{d}$ we mean lines of the form $\underline{c} + t\underline{d}$ (with $\underline{c} \in \mathbb{Z}^n$ fixed, $t \in \mathbb{R}$ variable).

Let A be a finite subset of $\mathbb{Z}^n$. Write $A = \{\underline{a}_1, \dots, \underline{a}_s\}$ where $\underline{a}_1, \dots, \underline{a}_s$ are arranged in lexicographic increasing order. We call A *convex* if every $\underline{a} \in \mathbb{Z}^n$ which belongs to the closed convex hull of A

belongs to A itself. By the *minimal corner* of a set $B \subseteq \mathbb{Z}^n$ we mean the lexicographically smallest element $\phi(B)$ of B.

If $f : A \to \mathbb{R}$, then the *line sum* of f along the line $l = \underline{c}+t\underline{d}$ is defined as $\sum_{\underline{a}\in A\cap l} f(\underline{a})$. For any $f : A \to \mathbb{R}$, write $\vec{f} := (f(\underline{a}_1), \ldots, f(\underline{a}_s))^T$. We often identify f and $\vec{f}$. The *length* of $\vec{f}$ (or f) is defined as $|f| = |\vec{f}| = \sqrt{\sum_{\underline{a}\in A}(f(\underline{a}))^2}$.

Let k be a positive integer and $S = \{\underline{d}_1, \ldots, \underline{d}_k\}$ be a fixed set of directions. By the *line sums along* S we mean all the line sums along lines in a direction from S which pass through at least one point of A. For $\underline{d} = (d_1, \ldots, d_n) \in S$ put

$$f_{\underline{d}}(\underline{x}) = (\underline{x}^{\underline{d}} - 1) \prod_{d_j<0} x_j^{-d_j},$$

$F(\underline{x}) = \prod_{i=1}^k f_{\underline{d}_i}(\underline{x})$ and, for $\underline{u} \in \mathbb{Z}^n$, set $F_{\underline{u}}(\underline{x}) = \underline{x}^{\underline{u}}F(\underline{x})$. Obviously, the polynomial $F_{\underline{u}}$ has integer coefficients. We call the functions $F_{\underline{u}}$ the *switching polynomials* of S. Define the functions $m_{\underline{u}} : \mathbb{Z}^n \to \mathbb{Z}$ by

$$m_{\underline{u}}(\underline{v}) = \mathrm{coeff}(\underline{x}^{\underline{v}}) \text{ in } F_{\underline{u}}(\underline{x}) \text{ for } \underline{v} \in \mathbb{Z}^n.$$

We define $D_{\underline{u}}$ as the set of $\underline{v} \in \mathbb{Z}^n$ for which $m_{\underline{u}}(\underline{v}) \neq 0$ and call it a *switching component*. Let $\phi(\underline{u})$ denote the minimal corner of $D_{\underline{u}}$. It follows from the above definitions that

$$m_{\underline{u}}(\phi(\underline{u})) = \pm 1. \tag{1}$$

3. The structure of functions with zero line sums

We show that Theorem 1 of [12], extended to functions $f : A \to Z$ with $Z = \mathbb{R}$ or $\mathbb{Z}$, remains true under the weaker condition that A is convex.

Theorem 3.1. *Let A be a finite convex subset of $\mathbb{Z}^n$, and S a given set of directions. Then any function $f : A \to Z$ with $Z = \mathbb{R}$ or $\mathbb{Z}$, having zero line sums along S can be uniquely written in the form*

$$f = \sum_{D_{\underline{u}}\subseteq A} c_{\underline{u}} m_{\underline{u}}$$

with coefficients $c_{\underline{u}} \in Z$. Moreover, every such function f has zero line sums along S.

If there is no $\underline{u}$ for which $D_{\underline{u}} \subseteq A$, then the only function f with zero line sums along S is the trivial function $f = 0$. Otherwise, the functions with zero line sums along S form a proper linear subspace or a proper sublattice of the linear space or lattice of all functions $f : A \to Z$, respectively, according as $Z = \mathbb{R}$ or $Z = \mathbb{Z}$.

Proof. First we prove the statement in case $A \subseteq \mathbb{Z}^n$ is a hyperblock with sides parallel to the axes. Note that in this case the statement is just Theorem 1 of [12], up to that there the cases $Z = \mathbb{Q}$ and $\mathbb{Z}$ were considered. As we mentioned already, the statement and proof are similar if $Z = \mathbb{R}$. Therefore we only give the main steps of the proof, for details we refer to [12].

Clearly, without loss of generality we may assume that the elements of A have nonnegative entries. Let χ_f be the generating function of f on A, defined by

$$\chi_f(\underline{x}) = \sum_{\underline{a} \in A} f(\underline{a}) \underline{x}^{\underline{a}}.$$

Then one can easily check (see Lemma 1 in [12]) that the fact that f has zero line sums along a direction $\underline{d} \in S$ is equivalent to that $f_{\underline{d}}(\underline{x})$ divides $\chi_f(\underline{x})$ in $Z[\underline{x}]$. Since the functions $m_{\underline{u}}$ have zero line sums along the directions in S, this already proves the second statement of the theorem (in the case where A is an appropriate hyperblock).

On the other hand, from Lemma 3 and Corollary 2 of [12] it follows that the polynomials $f_{\underline{d}}(\underline{x})$ $(\underline{d} \in S)$ are pairwise non-associated irreducible elements of the unique factorization domain $Z[\underline{x}]$. Hence we immediately obtain that

$$F(\underline{x}) \mid \chi_f(\underline{x}) \text{ in } Z[\underline{x}].$$

Thus there exists a polynomial $h(\underline{x}) = \sum_{D_{\underline{u}} \subseteq A} c_{\underline{u}} \underline{x}^{\underline{u}}$ in $Z[\underline{x}]$ such that $\chi_f(\underline{x}) = h(\underline{x}) F(\underline{x})$. This equation can be rewritten as

$$\chi_f(\underline{x}) = \sum_{D_{\underline{u}} \subseteq A} c_{\underline{u}} F_{\underline{u}}(\underline{x}).$$

Now by the definitions of $\chi_f(\underline{x})$ and the switching components $m_{\underline{u}}$ we get

$$f = \sum_{D_{\underline{u}} \subseteq A} c_{\underline{u}} m_{\underline{u}},$$

which proves that the f can be written in the desired form.

We are left with proving the uniqueness of the representation. Suppose that for some coefficients $l_{\underline{u}} \in Z$ $(D_{\underline{u}} \subseteq A)$ we have

$$\sum_{D_{\underline{u}} \subseteq A} l_{\underline{u}} m_{\underline{u}}(\underline{a}) = 0 \text{ for all } \underline{a} \in A.$$

By the definitions of the switching components, at the minimal corner of $m_{\underline{0}}$ all the other switching components vanish. This immediately gives $l_{\underline{0}} = 0$. Running through the switching components $m_{\underline{u}}$ with $D_{\underline{u}} \subseteq A$ in increasing lexicographical order according to $\underline{u}$, we conclude that all the coefficients $l_{\underline{u}}$ are zero. This implies the uniqueness of the above representation, and the theorem follows for hyperblocks.

Let now A be an arbitrary convex set in $\mathbb{Z}^n$, and let A^* be a hyperblock with sides parallel to the axes such that $A \subseteq A^*$. Set $f(\underline{x}) = 0$ for $\underline{x} \in A^* \setminus A$. Then we know that

$$f = \sum_{D_{\underline{u}} \subseteq A^*} c_{\underline{u}} m_{\underline{u}} \tag{2}$$

with coefficients $c_{\underline{u}} \in \mathbb{R}$. It remains to prove that $c_{\underline{u}} = 0$ if $D_{\underline{u}}$ is not contained in A.

If $D_{\underline{u}}$ is not contained in A, then there exists some $\psi(\underline{u}) \in D_{\underline{u}}$ such that $\psi(\underline{u}) \notin A$. Since A is convex, there is a linear manifold L which extends a hyperface of the convex hull of A (which is a closed set) such that $\psi(u)$ and A are on different sides of L. Let H_L be the open halfspace generated by L which contains $\psi(u)$. Note that H_L does not contain any element of A. Consider the set U_L of all $\underline{u}$ such that $D_{\underline{u}} \subseteq A^*$ and $D_{\underline{u}}$ contains an element $\psi(\underline{u}) \in H_L$. Without loss of generality we assume that $\psi(\underline{u})$ has maximal Euclidean distance $d(\psi(\underline{u}), L)$ to L among the elements of $D_{\underline{u}} \cap H_L$ and, if there are more such elements with maximal distance to L, then $\psi(u)$ is the lexicographically smallest among them. Since the sets $D_{\underline{u}}$ for variable $\underline{u}$ are translates of each other, the vectors $\psi(\underline{u}) - \underline{u}$ are the same for all $\underline{u} \in U_L$. Now we arrange the elements of U_L according to the non-increasing distances $d(\psi(\underline{u}), L)$ of $\psi(\underline{u})$ to L. Thereafter we order the elements of U_L for which the distances $d(\psi(\underline{u}), L)$ are equal according to nondecreasing lexicographic order of $\underline{u}$. Consider the first element $\underline{u} \in U_L$ according to this ordering. By the above construction there is no other set $D_{\underline{u}}$ for $\underline{u} \in U_L$ which contains $\psi(\underline{u})$. Since $\psi(u) \notin A$ we infer $f(\psi(\underline{u})) = 0$, hence $c_{\underline{u}} = 0$. We proceed with the next element $\underline{u} \in U_L$ in the ordering and conclude by a similar reasoning that $c_{\underline{u}} = 0$ also for this $\underline{u}$. Continuing until we have had all elements of U_L, we conclude that $c_{\underline{u}} = 0$ for all $\underline{u} \in U_L$. Since $D_{\underline{u}}$ was an arbitrary set not contained in A, the first statement follows. The uniqueness and the second statement of the theorem follow immediately from the statement concerning the case where A is a hyperblock. $\square$

Remark 3.1. The above theorem can be interpreted in the following way: for functions $f : A \to Z$ having zero line sums along S, the switching components form a basis when $Z = \mathbb{R}$, and form a lattice basis when $Z = \mathbb{Z}$, respectively.

The following result is a consequence of Theorem 3.1.

Corollary 3.1. *In the notation of Theorem 3.1, for any* $h : A \to Z$ *and for any prescribed values from* Z *at the minimal corners of the switching components contained in* A *there exists a unique* $f : A \to Z$ *having the prescribed values at the minimal corners and having the same line sums along* S *as* h *has.*

Proof. According to Theorem 3.1, for any coefficients $c_{\underline{u}}$, the function

$$f := h + \sum_{D_{\underline{u}} \subseteq A} c_{\underline{u}} m_{\underline{u}}$$

has the same line sums along S as h. By (1) we obtain, following the ordering argument from the previous proof, that we have precisely one choice for each coefficient $c_{\underline{u}}$, determined by the value of h at $\phi(u)$ together with the coefficients $c_{\underline{v}}$ fixed already. □

Remark 3.2. The following example shows that in Theorem 3.1 we cannot drop the convexity requirement.
Let $A = \{(0,0),(0,1),(1,0),(1,2),(2,1),(2,2)\}$ and $S = \{(1,0),(0,1)\}$. Then for every $\underline{u}$ we have $D_{\underline{u}} - \underline{u} = \{(0,0),(0,1),(1,0),(1,1)\}$. Therefore A does not contain any switching component. However, there is a nontrivial function $f : A \to \mathbb{Z}$ with all line sums along S equal to 0:
$f(0,0) = 1, f(0,1) = -1, f(1,0) = -1, f(1,2) = 1, f(2,1) = 1,$
$f(2,2) = -1.$

4. The best approximating function for general domains

The aim of this section is to construct the function $f_0 : A \to \mathbb{R}$ such that f_0 fits optimally the measured line sums along S in the sense of least squares and, moreover, has minimal Euclidean length among such functions.

Note that the results of this chapter can be obtained by standard tools from linear algebra; see e.g. the book of Golub and Van Loan [7]. In particular, Theorem 4.1 below is a reformulation of Theorem 5.5.1 on page 257 of [7]. Note that here we use a different notation as in [7]. We do so because of two reasons: on the one hand, we want the notation to fit with the notation in the rest of the paper, and on the other hand, our intention is to give everything as explicitly as possible.

Let $A \subseteq \mathbb{Z}^n$ be a finite, non-empty set, and write $\underline{a}_1, \ldots, \underline{a}_s$ for its elements. For the rest of this section, fix the indexing of the elements.

Let s be as above, t be an arbitrary positive integer, and B a t by s matrix of real numbers. The range of the matrix B is denoted by

$$R(B) := \{B \cdot \vec{x} \ : \ \vec{x} \in \mathbb{R}^s\}.$$

Hence $R(B)$ is a subspace of $\mathbb{R}^t$, generated by the column vectors $\vec{b}_1, \ldots, \vec{b}_s$ of B. We have $0 \leq \dim(R(B)) \leq t$. Write B_1 for a matrix formed by a maximal linearly independent set of column vectors of B. Then $B_1 = B \cdot C_1$ where C_1 is a matrix of type $s \times (\mathrm{rank}(B))$ which has $\mathrm{rank}(B)$ entries 1 in distinct columns and all other entries equal to 0. Observe that $B_1^T \cdot B_1$ is invertible. Then, as it is well-known (see e.g. [7]), for any $\vec{b} \in \mathbb{R}^t$

$$\vec{b^*} = B_1 \cdot (B_1^T \cdot B_1)^{-1} \cdot B_1^T \cdot \vec{b} \tag{3}$$

is the vector from $R(B)$ which is closest to $\vec{b}$.

Define

$$\vec{l}_f = B \cdot \vec{f}.$$

Let B_2 be a matrix formed by a maximal linearly independent set of row vectors of B. Then $B_2 = C_2 \cdot B$ where C_2 is a matrix of type $(\mathrm{rank}(B)) \times t$ which has $\mathrm{rank}(B)$ entries 1 in distinct rows and all other entries equal to 0. Observe that $B_2 \cdot B_2^T$ is invertible. Then we have the following statement.

Theorem 4.1. *Let $A, B, B_2, C_2, \vec{b}, \vec{b^*}, f\ (= \vec{f})$ be as above. Put*

$$\vec{f_0} = B_2^T \cdot (B_2 \cdot B_2^T)^{-1} \cdot C_2 \cdot \vec{b^*}. \tag{4}$$

Then the corresponding $f_0 : A \to \mathbb{R}$ has the following properties:

(i) *for any $f : A \to \mathbb{R}$ we have $|\vec{l}_f - \vec{b}| \geq |\vec{l}_{f_0} - \vec{b}|$,*

(ii) *if $f : A \to \mathbb{R}$, $f \neq f_0$ such that $|\vec{l}_f - \vec{b}| = |\vec{l}_{f_0} - \vec{b}|$, then $|\vec{f}| > |\vec{f_0}|$.*

Proof. The theorem is a reformulation of Theorem 5.5.1 on page 257 of [7]. □

Remark 4.1. An alternative version of Theorem 4.1 can be obtained by using the Moore-Penrose pseudo inverse, cf. [7], or the proof of Theorem 1 in [3].

Remark 4.2. We apply Theorem 4.1 in the context of Discrete Tomography as follows. Let A be a finite subset of $\mathbb{Z}^n$ and S a set of directions. Let $l_1, \ldots, l_t$ be the measured line sums along S. Note that because of noise they need not be consistent. Then B is the s by t matrix whose entry B_{ij} equals 1 if the line corresponding to l_j passes through $\underline{a}_i$ and 0 otherwise. The vector $\vec{b}^*$ given in (3) represents the corresponding line sums along S which are consistent and provide the optimal choice in the sense that $\sum_{j=1}^t (l_j - b_j^*)^2$ is minimal among the consistent line sums b_j^* along S. Furthermore, the vector $\vec{f_0}$ constructed in Theorem 4.1 is the shortest best approximation in the sense that it is the shortest vector realizing the line sums given by $\vec{b^*}$. The corresponding function $f_0 : A \to \mathbb{R}$ may be considered as the optimal choice for the measured line sums $l_1, \dots, l_t$.

We illustrate the method by an example.

Example. We use the previous notation. Consider the following subset of $\mathbb{Z}^2$:

$$A := \{(1,0),(3,0),(0,1),(4,1),(0,2),(4,2),(1,3),(2,3),(3,3)\}.$$

As the set of directions, take

$$S := \{(1,0),(0,1),(1,-1),(1,1)\}.$$

The ordering of the points in A and directions in S are arbitrary, but fixed. As a (measured) line sum vector, take

$$\vec{b}^T := \left(1, \frac{23}{10}, \frac{7}{5}, 1, 1, 1, \frac{3}{2}, 1, \frac{6}{5}, 1, 1, 1, \frac{9}{10}, \frac{13}{10}, \frac{1}{2}, 1, \frac{6}{5}, \frac{3}{5}, \frac{1}{2}, \frac{17}{10}, \frac{7}{10}\right).$$

The entries of $\vec{b}$ belong to the lines

$$y = t \quad (t = 0,1,2,3), \qquad x = t \quad (t = 0,1,2,3,4)$$

$$y = x + t \quad (t = -3,-2,-1,0,1,2), \qquad y = -x + t \quad (t = 1,2,3,4,5,6)$$

which we keep in this order. Then the matrix B of line sums is given by

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1
\end{pmatrix}.$$

As one can easily check, $\operatorname{rank}(B) = 9$. So we can take the matrix C_1 as the 9×9 unit matrix. Thus $B_1 = B$. Then, by (3), the vector $\vec{b^*}^T$ is given by

$$\left(\tfrac{891}{800}, \tfrac{2457}{1600}, \tfrac{1019}{800}, \tfrac{4361}{3200}, \tfrac{167}{128}, \tfrac{103}{128}, \tfrac{111}{128}, \tfrac{103}{128}, \tfrac{963}{640}, \tfrac{4239}{3200}, \tfrac{859}{1600}, \tfrac{1211}{1600}, \tfrac{1433}{3200}, \tfrac{287}{200}, \tfrac{2511}{3200}, \tfrac{4239}{3200}, \tfrac{1179}{1600}, \tfrac{571}{1600}, \tfrac{153}{3200}, \tfrac{367}{200}, \tfrac{3151}{3200}\right).$$

As one can readily check, the indices of a maximal set of independent rows of B is given by

$$\{1, 2, 3, 4, 5, 6, 7, 10, 11\}.$$

That is, we may take

$$C_2 := \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

whence

$$B_2 = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}.$$

Finally, by (4) we obtain

$$\vec{f_0}^T = \left(\frac{1211}{1600}, \frac{571}{1600}, \frac{1817}{3200}, \frac{3097}{3200}, \frac{1179}{1600}, \frac{859}{1600}, \frac{153}{3200}, \frac{111}{128}, \frac{1433}{3200}\right).$$

5. The best approximating function for convex domains

The following theorem provides explicitly a system of linear equations which determines the best approximating function constructed in the previous section. We illustrate in the corollary the advantage of this explicit expression. The real number $l(Y_\tau)$ in the following theorem can be considered as the measured line sum of f along the line corresponding to Y_τ.

Theorem 5.1. *Let $A \subseteq \mathbb{R}^n$ be convex. Let S be a finite set of directions and $Y_1, \dots, Y_t$ the subsets of A which determine the lines along S. Suppose for $\tau = 1, \dots, t$ a real number $l(Y_\tau)$ is given. Let $U_A \subseteq A$*

be the set of minimal corners of the switching components contained in A*. Define* $f_0 : A \to \mathbb{R}$ *by the system of linear equations*

$$\sum_{\underline{v}\in D_{\underline{u}}} f_0(\underline{v}) m_{\underline{u}}(\underline{v}) = 0 \text{ for all } \underline{u} \text{ with } \phi(\underline{u}) \in U_A, \tag{5}$$

$$\sum_{\tau:\underline{u}\in Y_\tau} \sum_{\underline{v}\in Y_\tau} f_0(\underline{v}) = \sum_{\tau:\underline{u}\in Y_\tau} l(Y_\tau) \text{ for all } \underline{u} \text{ with } \phi(\underline{u}) \in A \setminus U_A. \tag{6}$$

Then f_0 *is a function such that*

$$\sum_{\tau=1}^{t} \left(\sum_{\underline{v}\in Y_\tau} f_0(\underline{v}) - l(Y_\tau) \right)^2 \tag{7}$$

is minimal and among such functions f_0 *is the one for which the value of* $|\vec{f_0}|$ *is minimal.*

Proof. By Theorem 4.1 the function $f_0 : A \to \mathbb{R}$ satisfying (7) for which $|\vec{f_0}|^2 = \sum_{\underline{v}\in A} (f_0(\underline{v}))^2$ is minimal is uniquely determined. We proceed with this function f_0 and consider it as a function for which each value $f_0(\underline{u})$ for $\underline{u} \in A$ is a variable. It follows by differentation of (7) to each $f_0(\underline{u})$ that

$$\sum_{\tau:\underline{u}\in Y_\tau} \sum_{\underline{v}\in Y_\tau} f_0(\underline{v}) = \sum_{\tau:\underline{u}\in Y_\tau} l(Y_\tau)$$

for all $\underline{u} \in A$. Hence f_0 satisfies (6).

We know that $\vec{f_0}$ is orthogonal to the linear subspace L of functions having zero line sums along S. According to Theorem 3.1 the functions $m_{\underline{u}}$ have zero line sums along S. Therefore they are in L for all $\underline{u} \in \mathbb{Z}^n$. Since the inner product of $\vec{f_0}$ and any vector from L is 0, f_0 satisfies (5) too.

The numbers of linear equations in (5) and (6) together equal the cardinality of A. Thus it suffices to show that they are linearly independent over $\mathbb{R}$ in order to prove that f_0 is completely determined by them. Because of the orthogonality of $\vec{f_0}$ and L, it is enough to prove that the equations in (5) are linearly independent as well as those in (6).

Since by Theorem 3.1 the functions $m_{\underline{u}}$ are linearly independent, the equations (5) are linearly independent as well.

Furthermore, in Theorem 3.1 it is shown that f_0 is uniquely determined by its values at U_A. This shows that the equations in (6) are linearly independent. We conclude that the linear equations in (5) and (6) are linearly independent indeed. $\square$

In the particular case that $A \subseteq \mathbb{Z}^2$ is a rectangular block, and we only have row and column sums, we give an explicit form of f_0. The result shows that the formula from [6] is also valid if there is noise in the measurements. We simplify our notation.

Corollary 5.1. *Let $A = \{(i,j) \in \mathbb{Z}^2 : 0 \le i < q, 0 \le j < p\}, S = \{(1,0),(0,1)\}$. Let c_i $(i = 0,\dots,q-1)$ and r_j $(j = 0,\dots,p-1)$ denote the measured column sums and row sums, respectively. Further, write $s_r = \sum\limits_{j=0}^{p-1} r_j, s_c = \sum\limits_{i=0}^{q-1} c_i$ and $T = \frac{ps_r+qs_c}{q+p}$. Then for any $(i,j) \in A$ we have*

$$f_0(i,j) = \frac{c_i}{p} + \frac{r_j}{q} - \frac{T}{qp}.$$

Observe that if $s_r = s_c$, then $T = s_r = s_c$.

Proof. Since

$$(\frac{r_j}{q}+\frac{c_i}{p}-\frac{T}{qp})-(\frac{r_j}{q}+\frac{c_{i+1}}{p}-\frac{T}{qp})-(\frac{r_{j+1}}{q}+\frac{c_i}{p}-\frac{T}{qp})+(\frac{r_{j+1}}{q}+\frac{c_{i+1}}{p}-\frac{T}{qp}) = 0$$

for all i and j, the equations (5) are satisfied. Furthermore

$$(\frac{r_1}{q}+\frac{c_i}{p}-\frac{T}{qp})+\dots+(\frac{r_p}{q}+\frac{c_i}{p}-\frac{T}{qp})+(\frac{r_j}{q}+\frac{c_1}{p}-\frac{T}{qp})+\dots+(\frac{r_j}{q}+\frac{c_n}{p}-\frac{T}{qp})$$

$$= \frac{s_r}{q} + c_i - \frac{T}{p} + r_j + \frac{s_c}{p} - \frac{T}{q} = c_i + r_j,$$

which shows that the equations (6) are also satisfied. □

6. Approximate solutions in the integral case

Let A be a finite subset of $\mathbb{Z}^n$. We assume that a function $h : A \to \mathbb{R}$ is given and provide information on the 'nearest' function $f : A \to \mathbb{Z}$ having approximately the same line sums along S as h.

If $n = 2$ and only row and column sums are given, we have the following result.

Theorem 6.1. *If $h : A \to \mathbb{R}$ is given, there exists a function $f : A \to \mathbb{Z}$ such that every two corresponding elements of f and h as well as every two corresponding row sums and column sums as well as the sums of all function values of f and h differ by less than 1.*

We apply the following result of Baranyai.

Lemma 6.1 ([2], Lemma 3)**.** *Let $[h_{ij}]$ be an l by m matrix of real elements. Then there exists an l by m integer matrix $[f_{ij}]$ such that*

$$|h_{ij} - f_{ij}| < 1 \text{ for all } i,j,$$

$$|\sum_i h_{ij} - \sum_i f_{ij}| < 1 \text{ for all } j,$$

$$|\sum_j h_{ij} - \sum_j f_{ij}| < 1 \text{ for all } i,$$

$$|\sum_i \sum_j h_{ij} - \sum_i \sum_j f_{ij}| < 1.$$

Proof of Theorem 6.1. Choose an l by m block A^* which covers A. For $(i,j) \in A^* \setminus A$ put $h(i,j) = 0$. This does not change the line sums. Applying Lemma 6.1, we get $f(i,j) = h(i,j) = 0$ for $(i,j) \in A^* \setminus A$ and the theorem follows. □

The following example shows that the bound 1 is best possible. Let $0 < \varepsilon < 1$, $l > 1/\varepsilon$, $m = 1$, $h(i,1) = \varepsilon$ for $i = 1, \dots, l$. Then $f(i,1) = 1$ for some i in order to avoid that the row sums of h and f differ more than 1. But then the i-th column sums of h and f have a difference $1 - \varepsilon$.

The crucial feature of the following general result is that the upper bound is independent of the size of A.

Theorem 6.2. *Let A be a finite set in $\mathbb{Z}^n$. Let $h : A \to \mathbb{R}$ and let k directions S be given. Then there exists a function $f : A \to \mathbb{Z}$ such that each difference between corresponding elements of h and f is less than* 1 *and each difference between corresponding line sums of h and f along S is at most $k - 1$.*

We introduce the following notation in order to apply a result of Beck and Fiala. Let $X = \{x_1, x_2, \dots\}$ be a finite set and $\mathcal{F}$ a family of subsets of X. Associate to every x_i a real number α_i. Let k be the degree of $\mathcal{F}$, that is the maximal number of elements of $\mathcal{F}$ to which some element of X belongs. Let $r(k)$ be the least value for which one can find integers $a_i,\ i = 1, 2, \dots$ so that $|a_i - \alpha_i| < 1$ and

$$|\sum_{x_i \in E} a_i - \sum_{x_i \in E} \alpha_i| \le r(k)$$

for all $E \in \mathcal{F}$. The following result is due to Beck and Fiala (see [4]). We shall prove a generalization of it in the next section.

Lemma 6.2. *In the above notation, we have*

$$r(k) \le k - 1 \text{ for } k \ge 2.$$

Beck and Fiala conjecture that $r(k) \le k/2$ is true even for small values of k. Bednarchak and Helm [5] and Helm [13] improved the Beck-Fiala bound to $r(k) \leq k - 3/2$ for $k \geq 3$ and $r(k) \leq k - 2$ for k sufficiently large, respectively.

Proof of Theorem 6.2. Let $Y_1, \ldots, Y_t$ denote the subsets of A which determine the line sums along S. Let $\mathcal{F} = \{Y_1, Y_2, \ldots, Y_t\}$. By Lemma 6.2 there exist integers $f(a)$ for all $a \in A$ with $f(a) \in \{\lfloor h(a) \rfloor, \lceil h(a) \rceil\}$ such that $\sum_{a \in Y_j} |f(a) - h(a)| \leq k - 1$ for $j = 1, \ldots, t$. □

Remark 6.1. Obviously, many variations of Theorem 6.2 are possible. E.g. adding the requirement that the sum of all values $f(a)$ differs little from the sum of all values $h(a)$ leads to an upper bound k in place of $k - 1$. The requirement that the difference between the sums of the values of f and h along any linear manifold parallel to the axes should be small leads to an upper bound $2^k - 2$.

Remark 6.2. By a probabilistic method a better dependence on k can be obtained at the cost of some dependence on A. A recent improvement by Banaszczyk [1] of a result of Beck implies that in Theorem 6.2 the upper bound $k-1$ can be replaced by $C\sqrt{k \log(\min(m, n))}$, where C is some constant.

7. Approximate solutions for grey values

Theorem 7.1. *Let $Z = \{z_1, \ldots, z_m\}$ be a set of m real numbers with $z_1 < \cdots < z_m$. Put $z = \max_i(z_{i+1} - z_i)$. Let $h : A \to [z_1, z_m]$ and let k directions S be given. Then there exists a function $f : A \to Z$ such that the difference between the values of f and h at any element of A is at most z and each difference between corresponding line sums of f and h along S is at most $(k - 1)z$.*

For the proof we derive the following extension of the lemma of Beck and Fiala.

Lemma 7.1. *Let $Z = \{z_1, \ldots, z_m\}$ be a set of m real numbers with $z_1 < \cdots < z_m$. Put $z = \max_i(z_{i+1} - z_i)$. Let $X = \{x_1, x_2, \ldots, x_s\}$ be a finite set and associate to every x_i a real number $\alpha_i \in [z_1, z_m]$. Then given any family $\mathcal{F}$ of subsets of X having maximum degree $k \geq 2$, there exist $a_i \in Z$ such that $a_i = z_j$ if $\alpha_i = z_j$, there is no element from Z in between α_i and a_i for all i and j, and*

$$\left| \sum_{x_i \in E} a_i - \sum_{x_i \in E} \alpha_i \right| \leq (k - 1)z$$

for all $E \in \mathcal{F}$.

Proof. We shall define a sequence $\alpha^0, \alpha^1, \dots, \alpha^p$ of s-dimensional vectors $\alpha^j = (\alpha_1^j, \dots, \alpha_s^j)$ and a sequence Y_j of subsets of X with the following properties:
(i) $\alpha_i^0 = \alpha_i$ for $i = 1, \dots, s$.
(ii) There is no element of Z in between α_i and α_i^j for $i = 1, \dots, s; j = 0, 1, \dots, p$.
(iii) $X \setminus Y_j$ is a set of points x for which $x \in Z$ for all j.
(iv) $Y_0 \supset Y_1 \supset \dots \supset Y_p$ and $|Y_j| = p - j$ for $0 \le j \le p$.
(v) $\alpha_i^j = \alpha_i^h$ for $j = h, \dots, p$ whenever $\alpha_i^h \in Z$.
(vi) If $|E \cap Y_j| > k$, then $\sum_{x_i \in E} \alpha_i^j = \sum_{x_i \in E} \alpha_i^{j+1}$ for all $E \in \mathcal{F}$.
(vii) For $j = 0, 1, \dots, p$ and all $E \in \mathcal{F}$ we have

$$\left| \sum_{x_i \in E} \alpha_i^j - \sum_{x_i \in E} \alpha_i \right| \le (k-1)z.$$

According to (iii) and (iv) the final vector α^p has all coordinates in Z.

We construct the sequence (α^j) by induction. Suppose α^j is defined satisfying the above conditions for j. Let

$$G_j = \{E \in \mathcal{F} : |E \cap Y_j| \ge k\}.$$

We distinguish between three cases. At every step j there is some i such that $x_i \in Y_j, \alpha_i^{j+1} \in Z$ and we set $Y_{j+1} = Y_j \setminus \{x_i\}$.
Case (a) $G_j = \emptyset$.
Case (b) $0 < |G_j| < |Y_j|$.
Case (c) $|G_j| \ge |Y_j|$.

Case (a). If G_j is empty, then choose α_i^{j+1} as the element from Z which is nearest to α_i for all i with $x_i \in Y_j$. It follows that

$$\left| \sum_{x_i \in E} \alpha_i - \sum_{x_i \in E} \alpha_i^{j+1} \right| \le (k-1)z \text{ for all } E \in \mathcal{F},$$

and the above conditions are satisfied for $j + 1$.
(It follows that $\alpha_i^{j+1} = \dots = \alpha_i^p = a_i$ for all i.)

In Case (b) associate a real variable β_i to every $i = 1, \dots, s$ and consider the system of equations

$$\sum_{x_i \in E \cap Y_j} \beta_i = 0 \text{ for } E \in G_j,$$

$$\beta_i = 0 \text{ for } x_i \notin Y_j.$$

A nontrivial solution $\{\beta_i\}_{i=1}^s$ exists, because in case (b) there are more variables than equations. Let t_0 be the smallest nonnegative value for which $\alpha_i^j + t\beta_i \in Z$ for some i with $x_i \in Y_j$. Put $\alpha_i^{j+1} = \alpha_i^j + t_0\beta_i$ for $i = 1, \dots, s$. It is easy to check that

$$\sum_{x_i \in E} \alpha_i^j = \sum_{x_i \in E} \alpha_i^{j+1} \text{ for all } E \in G_j.$$

Hence the above conditions are satisfied for $j+1$.

Case (c). Since each x_i has degree at most k in G_j, we may conclude that $|G_j| = |Y_j|$, each x_i has degree exactly k in G_j and $|E \cap Y_j| = k$ for every $E \in G_j$. Let α_i^{j+1} be the element from Z nearest to α_i for every $x_i \in Y_j$. Then $|\alpha_i^{j+1} - \alpha_i| \leq z/2$ for $x_i \in Y_j$. Since $k/2 \leq k-1$, we obtain

$$|\sum_{x_i \in E} \alpha_i^{j+1} - \sum_{x_i \in E} \alpha_i| \leq (k-1)z$$

for all $E \in \mathcal{F}$. Hence the above conditions are satisfied for $j+1$. (It follows that $\alpha_i^{j+1} = \cdots = \alpha_i^p = a_i$ for all i.)

Write $a_j = \alpha_i^p$ for $i = 1, \dots, s$. It is easy to check that in each case the relations (iii), (iv) and (vii) hold. This completes the proof. □

Proof of Theorem 7.1. Let $Y_1, \dots, Y_t$ denote the subsets of A which determine the line sums along S. By Lemma 7.1 there exists a function $f : A \to Z$ such that

$$\sum_{a \in A \cap Y_j} |f(a) - h(a)| \leq (k-1)z$$

for $j = 1, \dots, t$. □

Remark 7.1. A small adjustment must be made if the entries are not all in $[z_1, z_m]$. E.g. values of h smaller than z_1 are first replaced by z_1, values larger than z_m by z_m. Note that such an adjustment will also change the bounds on the line sums.

Remark 7.2. If we want to have relatively short vectors $\underline{f}, \underline{g}$, then we may apply Theorem 7.1 to the function f_0 from Theorem 4.1 which again will change the bounds on the line sums. A better, but more complicated approach is to follow the proof of Lemma 7.1 and to decide at every step j where Case (a) or (b) applies, what is the best way of rounding. In step (b) one can as well let t_0 be the smallest nonpositive value with the indicated property.

8. Acknowledgements

The authors are grateful to the anonymous referees for their thorough work and helpful and useful remarks.

References

[1] W. Banaszczyk, *On series of signed vectors and their rearrangements*, Random Structures and Algorithms, 40 (2012), pp. 301-316.

[2] Zs. Baranyai, *On the factorization of the complete uniform hypergraph*, in: Infinite and finite sets, A. Hajnal, R. Raod, V.T. Sós, eds., Coll. Math. Soc. János Bolyai, Vol. 10, North Holland, Amsterdam, 1975, pp. 91-107.

[3] K. J. Batenburg, W. Fortes, L. Hajdu, R. Tijdeman, *Bounds on the difference between reconstructions in binary tomography*, LNCS 6607 (2011), pp. 369-380.

[4] J. Beck, T. Fiala, *Integer-making theorems*, Discr. Appl. Math., 3 (1981), pp. 1-8.

[5] D. Bednarchak, M. Helm, *A note on the Beck-Fiala theorem*, Combinatorica, 17 (1997), pp. 147-149.

[6] B. E. van Dalen, L. Hajdu, R. Tijdeman, *Bounds for discrete tomography solutions*, arXiv:1104.5589, 29 Apr 2011, 16 pp.

[7] G. H. Golub, C. F. Van Loan, *Matrix Computations*, Third Edition, JHU Press, 1996, 728 pp.

[8] L. Hajdu, *Unique reconstruction of bounded sets in discrete tomography*, Electronic Notes in Discrete Mathematics, 20 (2005), pp. 15-25.

[9] L. Hajdu, R. Tijdeman, *Algebraic aspects of discrete tomography*, J. Reine Angew. Math., 534 (2001), 119-128.

[10] L. Hajdu, R. Tijdeman, *An algorithm for discrete tomography*, Linear Algebra Appl., 339 (2001), pp. 147-169.

[11] L. Hajdu, R. Tijdeman, *Algebraic aspects of emission tomography with absorption*, Theor. Comput. Sci., 290 (2003), pp. 2169-2181.

[12] L. Hajdu, R. Tijdeman, *Algebraic discrete tomography*, in: Advances in Discrete Tomography and its Applications, G.T. Herman, A. Kuba, eds., Birkhäuser, 2007, pp. 55-81.

[13] M. Helm, *On the Beck-Fiala theorem*, Discr. Math., 207 (1999), pp. 73-87.

[14] A. P. Stolk, *Discrete tomography for integer-valued functions*, PhD-thesis, Leiden University, 2011.

[15] A.P. Stolk, K. J. Batenburg, *An algebraic framework for discrete tomography*, SIAM Journal on Discrete Mathematics, 24, pp. 1056-1079.

Institute of Mathematics, University of Debrecen, H-4010 Debrecen, P.O. Box 12, Hungary

E-mail address: hajdul@science.unideb.hu

Mathematical Institute, Leiden University, 2300 RA Leiden, P.O. Box 9512, The Netherlands

E-mail address: tijdeman@math.leidenuniv.nl