Diophantine approximations and integer points of cones

Martin Henk[*] Robert Weismantel[†]

Abstract

The purpose of this note is to present a relation between directed best approximations of a rational vector and the elements of the minimal Hilbert basis of certain rational pointed cones. Furthermore, we show that for a special class of these cones the integer Carathéodory property holds true.

Keywords: Simultaneous Diophantine approximation, Hilbert basis of rational cones, Carathéodory property.

1 Introduction

Throughout this paper we resort to the following notation. For integral points $z^1, \dots, z^m \in \mathbb{Z}^n$, the set

$$C := \text{cone}\left\{z^1, \dots, z^m\right\} = \left\{\sum_{i=1}^{m} \lambda_i z^i : \lambda \in \mathbb{R}^m_{\geq 0}\right\}$$

is called a *rational polyhedral cone.* It is called *pointed* if there exists a hyperplane $\{x \in \mathbb{R}^n : a^T x = 0\}$ such that $\{0\} = \{x \in C : a^T x \leq 0\}$. Here we are interested in generating systems of the integral points contained in such a cone.

Definition 1.1. *Let $C \subseteq \mathbb{R}^n$ be a rational polyhedral cone. A finite subset $H = \{h^1, \dots, h^t\} \subseteq C \cap \mathbb{Z}^n$ is called a* Hilbert basis *of C if every $z \in C \cap \mathbb{Z}^n$ has a representation of the form*

$$z = \sum_{i=1}^{t} \lambda_i h^i,$$

with non-negative integral multipliers $\lambda_1, \dots, \lambda_t$. A minimal Hilbert basis with respect to inclusion is also called an integral basis *of the cone C and it is denoted by $\mathcal{H}(C)$.*

The name Hilbert basis was introduced by Giles and Pulleyblank [GP79] in the context of totally dual integral systems. It was shown by Gordan [G1873]

[*] Supported by a “Leibniz Preis” of the German Science Foundation (DFG) awarded to M. Grötschel.

[†] Supported by a “Gerhard-Hess-Forschungsförderpreis” of the German Science Foundation (DFG).

that every rational polyhedral cone has an integral basis and for pointed cones we have the following result due to van der Corput [Cor31]: *The integral basis $\mathcal{H}(C)$ of a rational, pointed cone $C \subseteq \mathbb{R}^n$ is uniquely determined by*

$$\mathcal{H}(C) = \Big\{ z \in C \cap \mathbb{Z}^n \backslash \{0\} : z \text{ can not be written as the sum of two other elements of } C \cap \mathbb{Z}^n \backslash \{0\} \Big\}. \tag{1.1}$$

Although Hilbert bases play a role in various fields of mathematics, like combinatorial convexity, geometry of numbers, special desingularizations of toric varieties, or in integer programming, their geometrical structure is not very well understood yet. Besides the property to be "irreducible" (cf. (1.1)) no other geometrical property or characterization of the elements of a Hilbert basis is known. This paper tries to give a bit more insight by studying Hilbert bases (integral bases) of cones associated with the problem of simultaneous Diophantine approximation.

To this end let $A \in \mathbb{Z}^{m \times n}$ be an integral matrix of rank n and let $f_A : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be the norm on $\mathbb{R}^n$ defined by

$$f_A(x) = |Ax|_{\mathbf{1}} = \sum_{j=1}^{m} |a^j x|,$$

where $|\cdot|_{\mathbf{1}}$ denotes the $l_{\mathbf{1}}$-norm and a^j denotes the j-th row of A . With respect to that norm we study the

Directed Simultaneous Dioph. Approximation Problem (DSDAP) Let $p_1, \ldots, p_n, p_{n+1} \in \mathbb{Z}$, $p_{n+1} > 0$, $N \in \mathbb{N} > 0$. Find integers $q_1, \ldots, q_n \in \mathbb{Z}$ and an integer $q_{n+1} \in \mathbb{N} > 0$ such that $q_{n+1} \leq N$ and

1. $a^j \left(q_1/q_{n+1}, \ldots, q_n/q_{n+1}\right)^T \geq a^j \left(p_1/p_{n+1}, \ldots, p_n/p_{n+1}\right)^T, \quad j = 1, \ldots, m,$
2. $f_A \left(q_1/q_{n+1} - p_1/p_{n+1}, \ldots, q_n/q_{n+1} - p_n/p_{n+1}\right)$ is minimal.

Observe, that by neglecting the restrictions of the form 1., the problem reduces to the "standard" simultaneous Diophantine approximation problem of n rationals with respect to the norm f_A.

It has been known for a long time that the two-dimensional simultaneous Diophantine approximation problem (n=1) can be solved in polynomial time by the method of *continued fractions* as described in Khintchine [Khi56], Perron [Per13] and Grötschel, Lovász and Schrijver [GLS88]. Moreover, in the two-dimensional case best approximations have a nice geometric structure. More precisely, for a given $p = (p_1, p_2)^T \in \mathbb{Z}^2$ and $N \in \mathbb{N}$ let $C(I_1, p) = \text{cone}\{(1,0)^T, p\}$ and $C(-I_1, p) = \text{cone}\{(-1,0)^T, p\}$. Then it was shown by Klein [K1895] that a point $(q_1, q_2)^T$ on the lower convex hull of one of the two *Klein polyhedra* $K^+ = \text{conv}\{C(I_1, p) \cap \mathbb{Z}^2 \backslash \{0\}\}$, $K^- = \text{conv}\{C(-I_1, p) \cap \mathbb{Z}^2 \backslash \{0\}\}$ yields a best approximation of p_1/p_2, that is, $|q_1/q_2 - p_1/p_2|$ is minimal among all rationals whose denominator is bounded by N. In particular, an appropriate point $(q_1^+, q_2^+)^T$ $((q_1^-, q_2^-)^T)$ lying on the lower convex hull of K^+ (K^-) gives a best approximation of the directed problem: $\min |q_1/q_2 - p_1/p_2|$, $q_2 \leq N$ and $q_1/q_2 \geq p_1/p_2$ $(q_1/q_2 \leq p_1/p_2)$ (see also [BP94], [Fin93], [DS82]).

Whereas the two dimensional case ($n = 1$) of the problem has been very well understood, this much insight could not be gained for higher dimensions. We show in this paper that for every dimension and every $N \in \mathbb{N}$, a solution to the directed simultaneous approximation problem with smallest denominator belongs to the minimal Hilbert basis of the cone

$$C(A,p) = \Big\{ x \in \mathbb{R}^{n+1} : p_{n+1} a^j (x_1, \ldots, x_n)^T - x_{n+1} a^j (p_1, \ldots, p_n)^T \geq 0, \\ j = 1, \ldots, m, \quad x_{n+1} \geq 0 \Big\}. \tag{1.2}$$

Since the integral basis of any two dimensional rational pointed cone C consists of the integral points lying on the lower convex hull of conv $\{C \cap \mathbb{Z}^2 \backslash \{0\}\}$ (see e.g. [Oda88] and the references within), Klein's result follows indeed.

In the particular case when A coincides with the $(n \times n)$ identity matrix I_n, we have $C(I_n, p) = \text{cone}\,\{e^1, \ldots, e^n, p\}$, where $e^i \in \mathbb{R}^{n+1}$ denotes the i-th unit vector and $p = (p_1, \ldots, p_{n+1})^T$. In section 3 we show that if all the integers p_i are successively divisible then the cone $C(I_n, p)$ has the integer Carathéodory property, that is, each $z \in C(I_n, p) \cap \mathbb{Z}^n$ can be written as a non-negative integral combination of at most n elements of the minimal Hilbert basis. It has been conjectured by Cook, Fonlupt&Schrijver [CFS86] that this integer analogue of Carathéodory's theorem holds true for any rational pointed cone. This conjecture was proven by Sebö [Seb90] in dimensions ≤ 3 and has recently been disproved by Bruns, Gubeladze, Henk, Martin and Weismantel [BGHMW] in dimensions $n \geq 6$.

2 Simultaneous Diophantine Approximation

In this section we consider the directed simultaneous Diophantine approximation problem DSDAP introduced in the previous section.

Theorem 2.1. *Among all solutions of DSDAP let $q_1, \ldots, q_{n+1}$ be one with denominator q_{n+1} as small as possible. Then the vector $q = (q_1, \ldots, q_{n+1})^T$ is an element of the integral basis of $C(A,p)$, that is, $q \in \mathcal{H}(C(A,p))$.*

Proof. For abbreviation we write $\overline{x}$ for the first n components of a vector $x \in \mathbb{R}^{n+1}$ and $\tilde{x}$ for the "rational" vector $(x_1/x_{n+1}, \ldots, x_n/x_{n+1})$ if $x_{n+1} \neq 0$. On account of the restrictions 1., the vector q is an element of the cone $C(A,p)$ and conversely, for each vector $x \in C(A,p)$ with $x_{n+1} > 0$ the associated rational vector $\tilde{x}$ satisfies the restrictions 1. By the choice of the norm f_A this implies that for $x, y \in C(A,p)$, $x_{n+1} > 0, y_{n+1} > 0$ the norm $f_A(\cdot)$ is linear on conv $\{\tilde{x} - \tilde{p}, \tilde{y} - \tilde{p}\}$, that is,

$$\begin{aligned} f_A(\lambda \tilde{x} + (1-\lambda)\tilde{y} - \tilde{p}) &= f_A\left(\lambda(\tilde{x} - \tilde{p}) + (1-\lambda)(\tilde{y} - \tilde{p})\right) \\ &= \lambda f_A(\tilde{x} - \tilde{p}) + (1-\lambda) f_A(\tilde{y} - \tilde{p}), \quad 0 \leq \lambda \leq 1. \end{aligned} \tag{2.1}$$

Observe, that $C(A,p)$ is a pointed cone since $\text{rank}(A) = n$.

Now, suppose that q is not an element of the integral basis of $C(A,p)$. Then we can find two vectors $v, w \in C(A,p) \cap \mathbb{Z}^n \backslash \{0\}$ such that $q = v + w$ (cf. (1.1)) and therefore

$$\tilde{q} = \frac{1}{v_{n+1} + w_{n+1}} \overline{v} + \frac{1}{v_{n+1} + w_{n+1}} \overline{w} \tag{2.2}$$

Let $v_{n+1} \geq w_{n+1} \geq 0$. We have to distinguish two cases.
I. Let $w_{n+1} = 0$. By (2.2) and (1.2) we get for each $j \in \{1, \dots, m\}$

$$a^j\left(\tilde{q} - \tilde{p}\right) = a^j \tilde{v} + \frac{1}{v_{n+1}} a^j \overline{w} - a^j \tilde{p} \geq a^j\left(\tilde{v} - \tilde{p}\right).$$

Since $v, w \in C(A, p) \backslash \{0\}$ and $\mathrm{rang}(A) = n$ the equality cannot hold everywhere, and summation over all indices $j = 1, \dots, m$ yields $f_A(\tilde{q} - \tilde{p}) > f_A(\tilde{v} - \tilde{p})$. This contradicts the fact that $(q_1, \dots, q_{n+1})$ is a solution of DSDAP.
II. Let $v_{n+1} \geq w_{n+1} > 0$. Then we may write (cf. (2.2))

$$\tilde{q} = \frac{v_{n+1}}{v_{n+1} + w_{n+1}} \tilde{v} + \frac{w_{n+1}}{v_{n+1} + w_{n+1}} \tilde{w}$$

and (2.1) implies

$$f_A(\tilde{q} - \tilde{p}) = \frac{v_{n+1}}{v_{n+1} + w_{n+1}} f_A(\tilde{v} - \tilde{p}) + \frac{w_{n+1}}{v_{n+1} + w_{n+1}} f_A(\tilde{w} - \tilde{p}).$$

Thus, by the minimality of q we conclude that $f_A(\tilde{v} - \tilde{p}) = f_A(\tilde{w} - \tilde{p}) = f_A(\tilde{q} - \tilde{p})$. This contradicts the minimality of the denominator q_{n+1} and completes the proof. □

We remark that for $A = I_n$, $f_A(\cdot)$ is the l_1-norm. Then DSDAP is the problem to find a best approximation "from above" of the given rationals by other rationals whose common denominator is bounded. As pointed out in the introduction, the solutions of this problem for $n = 1$ can be interpreted as the lattice points lying on the lower convex hull of the conv $\left\{C(I_1, p) \cap \mathbb{Z}^2 \backslash \{0\}\right\}$. However, in general it is not sufficient to consider only the lattice points on the lower convex hull of $C(I_n, p)$. To see this, let $p = (1, \dots, 1, r)^T \in \mathbb{Z}^{n+1}$, $n \geq 2$, $r > 1$ and let $N = r - 1$. Then $C(I_n, p) = \mathrm{cone}\,\{e^1, \dots, e^n, p\}$ and the lower convex hull of conv $\left\{C(I_n, p) \cap \mathbb{Z}^{n+1} \backslash \{0\}\right\}$ is given by conv $\{e^1, \dots, e^n, p\}$. Obviously, $1/(r-1), \dots, 1/(r-1)$ is a solution of DSDAP, but the vector $(1, \dots, 1, r-1)^T$ is not contained in conv $\{e^1, \dots, e^n, p\}$.

3 Carathéodory property for special cones

The Carathéodry property also holds for a special family of cones that we investigated in Section 2.

Theorem 3.1. *Let $C = \mathrm{cone}\{e^1, \dots, e^{n-1}, p\}$ with $p = (p_1, \dots, p_n)^T \in \mathbb{N}^n$ satisfying $p_1 = 1$ and $p_i | p_{i+1}$, $1 \leq i \leq n-1$, that is, the numbers p_i are successively divisible. Let $\mathcal{H}(C)$ be the integral basis of C. Then for each $z \in C \cap \mathbb{Z}^n$ there exist at most n elements $b^1, \dots, b^n$ of $\mathcal{H}(C)$ such that $z = \sum_{i=1}^n v_i b^i$ with $v_i \in \mathbb{N}$.*

Proof. We use double induction with respect to the dimension n and the last coordinate p_n of the vector p. For $n = 2$ the theorem follows from the more general result of Sebö [Seb90]. Let $n \geq 3$. If $p_n = 1$ then the generators of C constitute a basis of $\mathbb{Z}^n$ and the result follows. Hence, let $p_n \geq 2$ and let

$$\begin{aligned} P(C) &= \left\{ z \in \mathbb{Z}^n : z = \sum_{i=1}^{n-1} \lambda_i e^i + \lambda_n p, \quad 0 \leq \lambda_i < 1 \right\} \\ &= \left\{ \left(1, \lceil j \cdot p_2/p_n \rceil, \dots, \lceil j \cdot p_{n-1}/p_n \rceil, j\right)^T : 1 \leq j \leq p_n - 1 \right\}. \end{aligned}$$

We prove that

$$\mathcal{H}(C) = \{e^1, \dots, e^{n-1}, p\} \cup P(C).$$

For simplification we distinguish two cases.

I. $p_2 = 1$. Let $z \in C \cap \mathbb{Z}^n$. We first analyze the case when $z_1 \geq z_2$. For a vector $x \in \mathbb{R}^n$ let $\tilde{x} = (x_2, \dots, x_n)^T$ be its orthogonal projection onto the plane $\{x \in \mathbb{R}^n : x_1 = 0\}$ (identified with $\mathbb{R}^{n-1}$). Then $\widetilde{z}$ is an integral vector of the $(n-1)$-dimensional cone $\widetilde{C} = \operatorname{cone}\{\widetilde{e}^2, \dots, \widetilde{e}^{n-1}, \widetilde{p}\}$ which is of the same type as the cone C. Hence, by induction hypothesis with respect to the dimension we can find $\widehat{b}^1, \dots, \widehat{b}^{n-1} \in \mathcal{H}(\widetilde{C})$ such that $\widetilde{z} = \sum_{i=1}^{n-1} v_i \widehat{b}^i$, $v_i \in \mathbb{N}$. Now, it easy to see that $b^i = (1, \widehat{b}^i)^T \in \mathcal{H}(C)$ and since $\widehat{b}^i_2 = 1$ we get

$$z = \sum_{i=1}^{n-1} v_i b^i + (z_1 - z_2) e^1.$$

Of course, the case $z_2 \leq z_1$ can be treated in the same way with respect to the orthogonal projection onto the plane $\{x \in \mathbb{R}^n : x_2 = 0\}$.

II. $p_2 > 1$. Let $v = (1, 1, p_3/p_2, \dots, p_n/p_2)^T \in P(C)$. Then we may write $v = (1/p_2)p + ((p_2 - 1)/p_2)e^1$ and thus v is contained in the relative interior of a 2-face of the cone C. Hence, $C = C_1 \cup C_2$ with

$$C_1 = \operatorname{cone}\{e^1, \dots, e^{n-1}, v\} \quad \text{and} \quad C_2 = \operatorname{cone}\{e^2, \dots, e^{n-1}, v, p\}.$$

Next we claim

$$C_1 \textit{ and } C_2 \textit{ satisfy the Carathéodory property.} \tag{3.1}$$

Obviously, C_1 is of the same type as C but with $v_n < p_n$ and therefore, we may assume that the Carathéodory property holds for this cone. Now, let U be the unimodular matrix determined by $Uv = e^1$, $Ue^i = e^i$, $i = 2, \dots, n$ and let

$$\overline{p} = Up = \left(1, p_2 - 1, \frac{p_3}{p_2}(p_2 - 1), \dots, \frac{p_n}{p_2}(p_2 - 1)\right)^T.$$

Then $UC_2 = \operatorname{cone}\{e^1, \dots, e^{n-1}, \overline{p}\}$ and this cone is of the same type as C but with $\overline{p}_n < p_n$. Therefore, we can assume that the Carathéodory property holds for the cone UC_2 and hence, also for C_2.

Finally, we claim

$$\mathcal{H}(C_1) \cup \mathcal{H}(C_2) = \mathcal{H}(C). \tag{3.2}$$

Obviously,

$$\begin{aligned}\mathcal{H}(C_1) &= \{e^1, \dots, e^{n-1}, v\} \cup \left\{ \left(1, \left\lceil j\frac{v_2}{v_n} \right\rceil, \dots, \left\lceil j\frac{v_{n-1}}{v_n} \right\rceil, j\right) : 1 \leq j \leq v_n \right\} \\ &= \{e^1, \dots, e^{n-1}, v\} \cup \left\{ \left(1, \left\lceil j\frac{p_2}{p_n} \right\rceil, \dots, \left\lceil j\frac{p_{n-1}}{p_n}, \right\rceil, j\right)^T : 1 \leq j \leq \frac{p_n}{p_2} \right\}.\end{aligned}$$

Thus $\mathcal{H}(C_1) \subset \mathcal{H}(C)$. For the cone UC_2 we get

$$\mathcal{H}(UC_2) = \{e^1, \dots, e^{n-1}, \overline{p}\} \cup \left\{ w^j : j = 1, \dots, \frac{p_n}{p_2}(p_2 - 1) \right\},$$

where $w^j = (1, \lceil jp_2/p_n \rceil, \ldots, \lceil jp_{n-1}/p_n \rceil, j)^T$, $1 \le j \le p_n/p_2(p_2-1)$. Now,

$$U^{-1}w^j = \left(1, \left\lceil \left(j + \frac{p_n}{p_2}\right) \frac{p_2}{p_n} \right\rceil, \ldots, \left\lceil \left(j + \frac{p_n}{p_2}\right) \frac{p_{n-1}}{p_n} \right\rceil, j + \frac{p_n}{p_2} \right)^T$$

and this shows $\mathcal{H}(C_2) \subset \mathcal{H}(C)$. Now, (3.2) follows from the trivial observation $\mathcal{H}(C) \subset \mathcal{H}(C_1) \cup \mathcal{H}(C_2)$.

On account of $C = C_1 \cup C_2$, (3.1) and (3.2) we get the desired result. □

We remark that with small modifications of the above proof one can show that a cone C as in Theorem 3 admits a unimodular partition, that is, one can find subcones C_i generated by the elements of $\mathcal{H}(C)$, such that i) the generators of C_i form a basis of $\mathbb{Z}^n$, ii) the union of the subcones C_i covers C and iii) the intersection of two distinct subcones is a face of both. Of course, this property implies the Carathéodory property.

Acknowledgment. We would like to thank the referee for helpful comments and for suggesting simplifications of the proofs.

References

[BP94] A.D. Bryuno and V.I. Parusnikov, *Klein polyhedrals for two cubic Davenport forms*, Mathematical Notes, vol. **46**, no. 3-4, 994–1007, (1994).

[BGHMW] W. Bruns, J. Gubeladze, M. Henk, A. Martin, and R. Weismantel, *A counterexample to an integer analogue of Carathеéodory's theorem*, J. reine angew. Math. **510**, 179–185, (1999).

[CFS86] W. Cook, J. Fonlupt, and A. Schrijver, *An integer analogue of Carathéodory's theorem*, J. Comb. Theory (B) **40**, 1986, 63–70.

[Cor31] J.G. van der Corput, *Über Systeme von linear-homogenen Gleichungen und Ungleichungen*, Proceedings Koninklijke Akademie van Wetenschappen te Amsterdam **34**, 368–371 (1931).

[DS82] A. Dress and R. Scharlau, *Indecomposable totally positive numbers in real quadratic orders*, Journal of Number Theory, vol. **14** no. 3, 292–306 (1982).

[Fin93] Yu.Yu. Finkel'shtein, *Klein polygons and reduced regular continued fractions*, Russ. Math. Surveys **48** (3), 198–200 (1993).

[GP79] F.R. Giles and W.R. Pulleyblank, *Total dual integrality and integer polyhedra*, Lineare Algebra Appl. **25**, 191–196 (1979).

[G1873] P. Gordan, *Über die Auflösung linearer Gleichungen mit reellen Coefficienten*, Math. Ann. **6**, 23–28 (1873).

[GLS88] M. Grötschel, L. Lovász and A. Schrijver, *Geometric algorithms and combinatorial optimization*, Springer Verlag Berlin (1988).

[Khi56] A. Khintchine, *Kettenbrüche*, Teubner Verlag, Leipzig (1956).

[K1895] F. Klein, *Über eine geometrische Auffassung der gewöhnlichen Kettenbruchentwicklung*, Nachr. Ges. Wiss. Göttingen, Math.-Phys. **3**, 357–359 (1895).

[Oda88] T. Oda, *Convex bodies and algebraic geometry*, Springer-Verlag, (1988).

[Per13] O. Perron, *Die Lehre von den Kettenbrüchen*, Teubner Verlag, Leipzig (1913).

[Seb90] A. Sebö, *Hilbert bases, Carathéodory's Theorem and combinatorial optimization*, in Proc. of the IPCO conference, Waterloo, Canada, 431–455 (1990).

[Sch80] A. Schrijver, *On cutting planes*, Annals of Discrete Mathematics **9**, 291–296 (1980).

Martin Henk
Technische Universität Wien
Institut für Analysis
Wiedner Hauptstr. 8-10/1142
A-1040 Wien, Austria
henk@osiris.tuwien.ac.at

Robert Weismantel
Universität Magdeburg
Institut für Mathematische Optimierung
Universitätsplatz 2
39106 Magdeburg, Germany
weismantel@imo.math.uni-magdeburg.de