APPROXIMATION OF LINEAR FORMS BY LATTICE POINTS

WITH APPLICATIONS TO SIGNAL PROCESSING

BY
I. VAUGHAN L. CLARKSON

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

January 1997

DECLARATION

The contents of this thesis are the results of original research, and have not been submitted for a higher degree at any other university or institution.

Some parts of this thesis have been published or submitted for publication in refereed journals or presented at conferences. Listed below is the bibliographic information pertaining to these preliminary presentations of results.

Clarkson, I. V. L., Howard, S. D. & Mareels, I. M. Y. (1996*a*). Estimating the period of a pulse train from a set of sparse, noisy measurements. In *Proceedings of the International Symposium on Signal Processing and its Applications (ISSPA '96)*, vol. 2, 885–888.

Clarkson, I. V. L. & Mareels, I. M. Y. (1996). Finding best simultaneous Diophantine approximations using sequences of minimal sets of lattice points. Submitted to *Math. Comp.*

Clarkson, I. V. L., Perkins, J. E. & Mareels, I. M. Y. (1993). On the novel application of number theoretic methods to radar detection. In *Proceedings of the International Conference on Signal Processing Applications and Technology (ICSPAT '93)*, vol. 1, 1202–1211.

Clarkson, I. V. L., Perkins, J. E. & Mareels, I. M. Y. (1995). An algorithm for best approximation of a line by lattice points in three dimensions. In *CANT '95 Conference Abstracts and Other Information*, 77–91.

Clarkson, I. V. L., Perkins, J. E. & Mareels, I. M. Y. (1996*b*). Number theoretic solutions to intercept time problems. *IEEE Trans. Inform. Theory*, **42** (3), 959–971.

The material contained herein is original with me, under the supervision of Professor Iven M. Y. Mareels.

.....

I. Vaughan L. Clarkson

ACKNOWLEDGEMENTS

The well-known danger with an acknowledgements section is that almost always someone is left off who is deserving of thanks. Nevertheless, I shall now do my best.

The initial thanks goes to Professor Doug Gray, whose direction, advice and powers of persuasion caused many things to happen that were beneficial for me. Not least of these was his rôle in the formation of the Cooperative Research Centre for Robust and Adaptive Systems and his instigation of proceedings both at DSTO and at ANU that resulted in the waiving of a number of rules which allowed me to pursue the course of my studies under somewhat unusual conditions.

I would like to thank my collaborators, including but not limited to Jane Perkins, Stephen Howard, Iven Mareels and Andy Pollington, for their helpful inputs, encouragement and for listening to many wild and impossible ideas without laughing.

My superiors at DSTO, including Des Lamb (now retired), Mal Brown and John Curtin, deserve thanks for their unwavering support of my work and for maintaining faith in my ability to complete it.

The Cooperative Research Centre for Robust and Adaptive Systems has been the instrument which facilitated my study at the ANU and the source of a considerable amount of funding throughout my studies. The Centre and its staff deserve my thanks.

My supervisor, Professor Iven Mareels, is perhaps the one most officially responsible for my safe passage through to graduation and he has discharged his duty of supervision with diligence and enthusiasm. I am very grateful to him for his constant interest and involvement in my work, for providing critical comment at all stages throughout the course of my studies and for his friendly and helpful disposition.

Finally, I would like to thank my family. I would like to thank my mother, to whom I owe my persistence, and my father, to whom I owe my love of learning. I would like to thank my wife, Emma, for the love and support she ceaselessly gave, for tolerating the long hours and the short weekends, and for always "being there." Thanks too to my two sons, Sam and Tom, who reminded me that there are more important and enjoyable things in life than study.

ABSTRACT

This thesis is concerned with the theory of Diophantine approximation, simultaneous Diophantine approximation and the geometry numbers and its application to problems in signal processing involving periodic pulse trains. The contributions of the thesis are in computational mathematics and algorithms for signal processing.

We begin with a review of Diophantine approximation. The notions of best Diophantine approximation are developed. The relationship of Euclid's algorithm with best homogeneous approximation and Cassels' algorithm with best inhomogeneous approximation is explored. We also examine the relationship of best approximations with other mathematical objects. We discover a direct relationship between the successive maxima of certain periodograms and the best Diophantine approximations of a real number.

After reviewing some of the theory of the geometry of numbers, such as point lattices, Minkowski's theorems, lattice reduction and the LLL algorithm, we study simultaneous Diophantine approximation. We develop algorithms which are able to find successive best simultaneous Diophantine approximations for lattices of ranks two and three under quite general conditions. For lattices of rank three, the algorithms are able to find successive best approximations by lattice points of both lines and planes. We also review Brun's algorithm and more modern algorithms such as the HJLS algorithm of Hastad *et al.* (1989) for simultaneous Diophantine approximation for lattices of arbitrary rank.

We then discuss the problem of calculating the intercept time, or simultaneous coincidence, of a number of periodic pulse trains. Where the phase is unknown, we calculate the probability of intercept. We show that the problems can be interpreted as problems of Diophantine approximation and simultaneous Diophantine approximation. We give expressions for the intercept time and probability of intercept under a number of different conditions.

We also discuss the problem of estimation of the period and phase of a periodic pulse train of which only a short, sparse and noisy record exists of the times-of-arrival of the pulses. We apply a simultaneous Diophantine approximation algorithm derived from the LLL algorithm to the estimation problem. In numerical simulations, we find that it frequently obtains very good estimates even for records from which 99.9% of pulses are missing, of which only nine remain and for which errors in the measurement of time-of-arrival are as much as 1% of the period.

CONTENTS

PREAMBLE

This thesis has turned out rather differently than I had imagined at its beginning and it has been written under slightly unusual conditions. The opportunity to study at the ANU arose through the creation of the Cooperative Research Centre for Robust and Adaptive Systems — $(CR)^2ASys$ — in 1991. At that time, I was not long finished my undergraduate studies at the University of Queensland, and I had recently taken up a full-time position at the Electronic Warfare Division (EWD) at the Defence Science and Technology Organisation (DSTO) in Salisbury, South Australia. Professor Doug Gray, who had been one of the initial "conspirators" in the establishment of the Centre and who was at that time the Research Leader of the Signal and Information Processing Branch at EWD, offered me the possibility of study at the ANU. I eagerly accepted. It was proposed that I should receive my day-to-day supervision at DSTO, retaining my position as a Professional Officer in the Commonwealth Public Service, with only short visits to Canberra. However, this posed difficulties with enrolment at the ANU, which at that time did not allow such extended absences from Canberra. Special permission was required and a year passed before it was obtained. My official enrolment began in February, 1993.

Prior to the commencement of these studies, I had been involved with the signal processing discipline of frequency estimation and, through the Pulse Train Project of $(CR)^2ASys$, with pulse train signal processing. A special interest had been in the application of parallel processing to signal processing algorithms. Also during that time, I was responsible for writing an initial version of software that has come to be known as *IDEA*, an Interactive Deinterleaver for ELINT[1] Applications, as well as being involved in writing the specification of the current version. The software was designed to be a powerful tool for the deinterleaving of radar pulse trains and its development is continuing. It was envisaged at the outset that my thesis, as a part of the activities of the Pulse Train Project, would focus on the problem of deinterleaving pulse trains and especially *time-of-arrival* deinterleaving, where the periods of the pulse trains are the chief or sole source of information used for their separation. As it turned out, the thesis barely mentions deinterleaving. However, the original aspiration has not been entirely extinguished, for the thesis, I believe, has succeeded in strengthening the theoretical foundation for the study of pulse trains, from which algorithms for deinterleaving can be derived and analysed.

[1]Electronic Intelligence.

It was certainly not envisaged that my studies would lead me into the field of approximation of linear forms by lattice points, a phrase which I use to encompass the fields of Diophantine approximation, simultaneous Diophantine approximation and that part of the geometry of numbers that deals with lattice reduction and short lattice vectors (although in the last case, the linear form to be approximated degenerates to a point). These fields are not usually associated with signal processing. Indeed, my knowledge of these subjects was extremely limited at the commencement of my studies.

My attention was initially diverted by the problem of calculating intercept times of periodic pulse trains. This problem is related to deinterleaving, but the relationship is not direct. I became interested in the observation, which had been made initially by Greg Noone and communicated to me by Jane Perkins, that the probability of intercept of two pulse trains can be calculated by considering the positions of pulses on a "folded interval" (or a circle) having the length (or circumference) of one of the periods and that the pulses exhibit a "clustering" effect on that interval as the amount of observation is increased. It seemed to me that the structure of the clustering could be exploited. I soon discovered, and the results are presented in Chapter 5, that an efficient method of calculating the probability of intercept could be obtained using Euclid's algorithm. This led me to study Diophantine approximation and thence the other number theoretic subjects explored in this thesis. My efforts were chiefly directed towards finding an expression and an algorithm to allow efficient calculation of the intercept time or probability of intercept in problems involving more than two pulse trains. It was only rather late in my studies that I realised the application of the theory to the estimation of the period of pulse trains, a topic which is discussed in Chapter 6, and to a number of other areas in signal processing which time (and space!) has not allowed me to examine in this thesis. However, the period estimation problem is indeed closely related to the deinterleaving problem.

Out of this haphazard excursion into number theory, it is my fervent hope that something coherent, interesting and useful has emerged.

NOTATION

1. We use $\mathbb{R}$, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ to denote the real numbers, natural numbers $(1, 2, 3, \dots)$, integers and rational numbers, respectively.

2. We generally use bold face and a roman style to denote a vector, *e.g.* $\mathbf{x} \in \mathbb{R}^n$, and we use bold face, roman style and upper case to denote a matrix, *e.g.* $\mathbf{A} \in \mathbb{R}^{n \times n}$, unless otherwise noted, in contrast to an italic style for scalars, *e.g.* $x \in \mathbb{R}$.

3. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we adopt the convention that $\mathbf{a}_j$ refers to the j^{th} column vector of $\mathbf{A}$ and a_{jk} is the element in the j^{th} row and k^{th} column of $\mathbf{A}$.

4. The superscript T, in reference to a matrix, denotes its *transpose*.

5. We reserve $\mathbf{I}$ for the identity matrix, $\mathbf{0}$ and $\mathbf{1}$ for the vectors consisting entirely of zeros and ones, respectively.

6. As a slight abuse of notation, we will occasionally and without notice use a vector which has been defined in $\mathbb{R}^n$ as if it were a matrix in $\mathbb{R}^{n \times 1}$. That is, all vectors are assumed to be *column vectors* unless otherwise specified. For instance, for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we use $\mathbf{x} \cdot \mathbf{y}$ and $\mathbf{x}^T \mathbf{y}$ interchangeably.

7. If $x \in \mathbb{C}$ then $\mathfrak{R}\{x\}$ denotes its real part and $\mathfrak{I}\{x\}$ denotes its imaginary part.

8. We define the SIGNUM function of a real number as

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

9. We use $\lfloor x \rfloor$ to denote the greatest integer less than or equal to the real number x. That is, $\lfloor x \rfloor \leqslant x < \lfloor x \rfloor + 1$. Similarly, we use $\lceil x \rceil$ to denote the least integer greater than or equal to x. We use $\lfloor x \rceil$ to denote a nearest integer to x. Where x is a half-integer, its value is unspecified. Where it is necessary to specify which of the nearest integers is required, we shall make this clear in the text.

10. We use $\operatorname{Int} \mathcal{S}$ and $\operatorname{vol} \mathcal{S}$ to denote the *interior* and *volume* of a set $\mathcal{S}$ (see Definition 3.4 and Definition 3.8 of Chapter 3, respectively).

11. The use of vertical bars, $|\cdot|$, when applied to a discrete set, indicates the number of elements in the set.

12. We will frequently make use of Bachmann's O-NOTATION for expressing asymptotic quantities. Recall that $g(x) = O(f(x))$ for real-valued functions f and g of a real number implies that there exists some positive constants x_0 and M such that $x \geqslant x_0$ implies that $|g(x)| \leqslant M|f(x)|$.

CHAPTER 1

INTRODUCTION

1. Problem Statement

Many physical phenomena exhibit some form of periodicity. From the ticking of a clock to the quantisation of energy, they pervade the physical world. Their study in mathematics has been ongoing throughout its history. This thesis has been motivated by the need to understand the interactions between periodic processes with differing periods and to estimate the periods of infrequently observed periodic processes. That this should lead to the study of integers is not surprising, for the purest representation of a periodic process is the embedding of the integers in the continuum. The study of integers is variously known as arithmetic or the theory of numbers. In this thesis, we explore the closely related branches of number theory — Diophantine approximation, simultaneous Diophantine approximation and the geometry of numbers — and their application to two problems in signal processing and system analysis.

The signal processing applications we will examine are motivated by ELECTRONIC SUPPORT MEASURES (ESM). This is a discipline of electronic warfare. It is the theory and practice of monitoring the electromagnetic spectrum in order to provide information on the location and nature of both friendly and hostile sources of radiation. In an operational environment, the information must be timely and accurate, enabling the marshalling of offensive and defensive resources, such as weapon systems, electronic countermeasures and electronic counter-countermeasures. One of the most important sources of radiation that must be detected and analysed are radars. Radars typically emit a sequence of radio frequency pulses in a periodic sequence. Moreover, the search patterns of receiving equipment often have a periodic nature. The periodicity in the radar emitter arises because of the ease of processing radar returns from periodic sequences of pulses. The periodicity in ESM receivers arises from the need to search a large range of parameter space with rather sensitive equipment that can only examine a small range of parameters at once. For example, it may be necessary to scan over bearing and carrier frequency. Therefore, the receiver may have an antenna that rotates at a fixed rotational speed and tunes to each frequency of interest in a fixed and repeating sequence. This sort of periodic behaviour can also be found in radars, which also seek to acquire information about the environment, although through active rather than passive means.

Therefore, many problems in ESM reduce to problems which involve periodic pulse trains. Understanding how they interact is vital to being able to analyse and design ESM equipment, and to process the signals which they receive.

The ESM problems which we treat are intercept time problems and the problem of estimating the parameters of a periodic pulse train of which only a short, incomplete and noisy record exists of the times-of-arrival (TOAs).

The intercept time problem is one of predicting the time of intercept between two of more periodic pulse trains, or calculating the probability of an intercept over a certain time interval if there is incomplete information. Our interest in the intercept time or probability of intercept is obviously driven by the desire that the receiver should intercept the emissions of a radar at the earliest possible time in order to provide as much time as possible for developing a response. To design ESM equipment, we would like to know what control, if any, can be exerted over the design to ensure that the intercept time is as small as possible or the probability of intercept is as high as possible for likely threats.

The problem of estimating the period and phase of short, sparse and noisy record of TOA measurements is one which occurs in the processing of radar pulse trains by ESM equipment. The process of scanning through different sectors of space and through different carrier frequencies might cause the record of any individual pulse train to be incomplete. The measurements of the TOAs may not be precise due to the presence of noise. However, the pulse repetition interval (PRI) of a radar pulse train is one of the most important parameters to estimate. A PRI is often peculiar to the make of radar. It frequently gives information about the mode of operation. The mode of a radar varies depending upon the intentions of the platform (craft, vehicle, vessel or person) which carries it. For instance, navigational radars employ different modes depending on the maximum range of interest, varying the PRI accordingly. Importantly, radars which queue or guide weapon systems employ different modes depending upon what stage of the targetting process they are at. To understand what quality of data is required to reliably estimate the period and to improve the ability to compute it online are the aims of our investigations in this area.

Both of these problems rely heavily on the theory of Diophantine approximation and simultaneous Diophantine approximation for their solution. Since practitioners of signal processing are not generally well-versed in these areas and because of the need to develop new results for our purposes, we devote the greater amount of space in the thesis to these topics. Indeed, it is hoped that the contributions in these fields have merit in their own right.

Diophantine approximation is the study of the approximation of real numbers by rational numbers. It is so named in honour of DIOPHANTOS, who studied the solution of certain equations in rationals. Homogeneous Diophantine approximation of a real number α involves finding non-zero integers p and q that make $q\alpha - p$ or

$\alpha - p/q$ small in absolute value. Inhomogeneous approximation of α with respect to another real number β involves finding integers p and q that make $q\alpha - p - \beta$ small in absolute value. The smallest non-zero integers which give an approximation error less than a certain value are *best* approximations. We will concern ourselves with methods for calculating these best approximations.

Simultaneous Diophantine approximation extends the problem to simultaneous approximation of many real numbers by rationals. The multiplicity of numbers to be approximated determines the dimension of the problem. For higher dimensions, the way in which the approximation error is measured allows extra variability in the statement of the problem and as to how best approximations are defined. Simultaneous Diophantine approximation is closely related to *integer programming*. Integer programming is the problem of finding a set of integers which minimise a linear cost function with respect to a number of linear constraints. Whereas Diophantine approximation is a rather mature branch of number theory, simultaneous Diophantine approximation (and integer programming) are not yet as richly endowed with theoretical results. Coupled with this, or perhaps a consequence, is the apparent computational intractability of finding best simultaneous Diophantine approximations. It is the computational aspects of the problem we shall be most concerned with. For low dimensional problems, we explicitly formulate algorithms which are capable of finding complete sequences of best approximations, in a certain sense.

A convenient way of expressing problems of simultaneous Diophantine approximation is in terms of the approximation of linear forms by lattice points. A point lattice is a set of points which is generated from integer linear combinations of linearly independent basis vectors. The means by which we measure the closeness of approximation is through the application of norms and semi-norms. The metric balls associated with norms are convex bodies. The study of the relationship between point lattices and convex bodies is part of the geometry of numbers. For this reason, we find it appropriate to review the geometry of numbers as a precursor to our study of simultaneous Diophantine approximation. The main object of our investigations into the geometry of numbers are to introduce point lattices, convex bodies and the various notions of lattice reduction. Lattice reduction is the process of finding a lattice basis in a canonical form. The canonical form is usually stipulated in order to ensure that the basis vectors are short with respect to a certain norm. The problem of finding short lattice vectors is similar to the simultaneous Diophantine approximation. Indeed, the relatively recently discovered LLL algorithm for Lovász reduction of lattices is an important tool for simultaneous Diophantine approximation because of its ease of computation and its guarantee of finding reasonable short vectors (in a sense we will make precise). The importance of this algorithm to simultaneous Diophantine approximation and thence to our signal processing problems warrants that we should spend some time in understanding its operation.

2. Organisation of This Thesis

Our treatment of the subjects places heavy emphasis on algorithms for the solution of problems both in number theory and in signal processing. The thesis has two parts. The first part (Chapter 2 to Chapter 4) is concerned with number theory and is presented, on the whole, as pure mathematics with a computational bent. The second part is about the signal processing applications of probability of intercept (Chapter 5) and parameter estimation of periodic pulse trains (Chapter 6). We construct in the first part the algorithms and theoretical tools which are necessary in the second part. In the second part, we show that the intercept time problems and the parameter estimation problems are best understood as Diophantine approximation and simultaneous Diophantine approximation problems. We give details of how the algorithms from the first part can be modified and applied.

In Chapter 2, we introduce the subject of Diophantine approximation. We define the meanings of best Diophantine approximation, both homogeneous and inhomogeneous and in the absolute and relative sense. In order to appreciate the efficiencies of the algorithms of Euclid and Cassels, we present some naïve algorithms for finding best Diophantine approximations which essentially involve an exhaustive search over the integers. We then show how Euclid's algorithm can be employed to find homogeneous best approximations much more quickly. In fact, we show that the correspondence between the output of Euclid's algorithm and the sequence of best approximations for a given real number α is nearly one-to-one. We demonstrate the relationship of the best approximations and Euclid's algorithm with the *simple continued fraction* expansion of a real number. We show that the best approximations in the absolute sense correspond almost completely with the *convergents* of that expansion and that the best approximations in the relative sense are *intermediate fractions* of the expansion.

We then investigate Cassels' algorithm for computing best inhomogeneous Diophantine approximations. We show that the algorithm produces outputs, which we call *auxiliary convergents*, from which all the best inhomogeneous approximations can be found.

The application of best Diophantine approximations to other mathematical objects is then explored. We discover a relationship between best Diophantine approximations and the successive maxima of certain almost periodic functions constructed from *diagonal functions*. As an important example, we discover that the successive peaks of a *periodogram* of three time samples with positive amplitude can be interpreted in terms of the best approximations of the ratio of the differences of the sample times. We also show how best Diophantine approximations with a prescribed approximation error can be located in a *Farey series* of appropriate order.

In Chapter 3 we discuss the geometry of numbers. We introduce the point lattice and the convex body. We present Minkowski's first and second theorems. We

discuss and give an algorithm for finding the shortest vector in a lattice of arbitrary rank. We then discuss various notions of lattice reduction, namely those of Gauss, Minkowski, Hermite, Korkin & Zolotarev and Lovász. We present and analyse algorithms for reduction in the sense of Gauss, Hermite and Lovász. For Gaussian reduction, we demonstrate a relationship with the development of the centred continued fraction expansion of a complex number. The algorithm we present for Lovász reduction is a variant of the LLL algorithm of Lenstra *et al.* (1982).

Simultaneous Diophantine approximation is discussed in Chapter 4. In the first part of the chapter, we introduce the theory of (ρ, h)-*minimal sets* which, as we shall see, can be used to describe algorithms which are guaranteed of finding best simultaneous Diophantine approximations, according to our quite general definition. For lattices of rank two and three, we are able to realise these algorithms. For lattices of rank two, we discover that our algorithm can be made to operate like an additive version of Euclid's algorithm (which we discuss in Chapter 2) or Gauss' algorithm (which we discuss in Chapter 3), depending upon the inputs. The algorithm for lattices of rank three also possesses a rather simple, additive nature. We present numerical examples to demonstrate its ability to discover best approximations in a quite general class of problems. We demonstrate its ability to find best approximations to a line and to a plane in three dimensions, with respect to the Euclidean norm and with respect to the sup-norm. Furthermore, we are able to increase the speed of the algorithm by skipping some intermediate bases. The algorithm that we derive in this way — which we call the "accelerated" algorithm — can be regarded as a generalisation of, and was inspired by, an algorithm of Furtwängler (1927). With numerical examples, we demonstrate its equivalence with Furtwängler's algorithm under certain conditions, we demonstrate its ability to generate best approximations from any input basis and we present some evidence to suggest that, at least for some inputs, the algorithm is able to find best approximations with a prescribed maximum approximation error in a number of iterations which is proportional to the logarithm of the error.

Since the prospects of the existence of a computationally efficient algorithm for finding best approximations for lattices of arbitrary rank are thought to be poor, we then turn our attention to algorithms which are intended to produce good approximations with a moderate amount of computation. The first algorithm of this type we review is Brun's algorithm. We find that the algorithm has a rather natural geometrical interpretation. However, it can be shown that, for certain inputs, it does not produce good approximations. The invention of the LLL algorithm has provided a tool for the development of algorithms for simultaneous Diophantine approximation from which provably good approximations can be obtained without excessive computational effort. As an example of recent algorithms which are derived from the LLL algorithm, we study the HJLS algorithm of Hastad *et al.* (1989)

but we also mention the PSLQ algorithm of Ferguson & Bailey (1991) and algorithms of Just (1992) and Rössner & Schnorr (1996).

In Chapter 5, we study intercept time and probability of intercept problems. We begin by considering problems involving only two pulse trains. We consider problems in which the phase (time offset of the first pulse from the origin) of both pulse trains are known *a priori* and they are equal, where they are known and unequal and where one or both of the phases are unknown and assumed to be uniformly distributed over the range of the PRI. For the problem of *in phase* initial conditions, where both phases are known and equal, we formulate the problem as one of finding a best approximation with a certain approximation error, the error being determined by the sum of the pulse widths of the intercepting pulses. We show how Euclid's algorithm can be applied to obtain solutions. For *arbitrary phase* initial conditions, we find that the problem can be stated as a problem of best inhomogeneous Diophantine approximation and, consequently, Cassels' algorithm can be applied to obtain solutions. We also show that all further intercepts after the first can be obtained by means of a recurrence relation.

When one or both of the phases are random, we cannot compute the intercept time. Rather, we must be content with determining the probability of intercept over a time interval. For the *discrete time* probability of intercept problem, where one phase is known and the other is uniformly distributed, we calculate the probability of an intercept occurring with one of the first N pulses of the pulse train with known phase. The probability of intercept turns out to have a piecewise linear form with a maximum of four linear segments. We interpret the slopes and the positions of the boundaries of the segments in terms of best approximations and associated approximations and their approximation errors. For the *continuous time* probability of intercept, where neither of the phases are known beforehand, the probability is calculated over an observation interval, the length of which is a continuous variable. The form of the probability of intercept is again shown to be piecewise linear, with an additional quadratic segment. We find that, for small pulse widths, the expression for the continuous time probability is well-approximated by the discrete time probability. We also discuss how the probability of intercept varies as a function of the PRI of a pulse train, and show how this is related to adjacent elements of a Farey series of appropriate order.

For the intercept time and probability of intercept of multiple pulse trains, we are able to formulate the problem as a simultaneous Diophantine approximation problem. We find that the properties which enabled relatively simple expressions to be obtained for intercepts involving two pulse trains do not generalise in an obvious way. The only problem which we are able to solve satisfactorily is the in phase intercept time of three pulse trains, to which we can apply the additive or accelerated algorithm of Chapter 4. However, we are able to disprove the existence of a bounded number of linear segments in the expression for the continuous time

probability of intercept of more than two pulse trains. Also, we are able to give an expression for the probability of intercept over a short time interval.

We briefly review other approaches to intercept time problems that have appeared in the literature. We review methods which exploit properties of linear congruence and those which replace the assumption of strict periodicity with stochastic behaviour. We find that the exploitation of linear congruence properties is nearly identical to our approach. For the statistical description of pulse trains, we find that the well-known expression for probability of intercept amongst many pulse trains that was derived by Self & Smith (1985) is similar to our own expression for short observation intervals.

In Chapter 6, we discuss the problem of estimating the PRI and phase of a periodic pulse train of which only a short, sparse and noisy record exists of the TOAs. We propose two statistical models for the observed data. The first model, which we call "simple," assumes only that the measurement errors on the TOAs are independent and identically distributed (i.i.d.) zero mean normal random variables. It assumes nothing about the way in which pulses go missing from the record. The second model, which we call "extended," assumes that the differences in consecutive pulse indices of the observed pulses are drawn from a geometric distribution.

We consider the maximum likelihood estimation of the PRI and phase for data generated by these models. We find that the problem is a simultaneous Diophantine approximation problem and that, for the simple model, either no maximum likelihood estimates exist or there are an infinitude. For the extended model, we consider the *joint maximum likelihood estimation and association* (JMLEA) of PRI, phase and pulse indices. We show that the JMLEA exists and that a sensible approach to its calculation is through the application of a simultaneous Diophantine approximation algorithm derived from the LLL algorithm. We show that there is a strong connection with the maximisation of the periodogram of the observed TOAs, further strengthening the results of Chapter 2.

We present the results of extensive numerical simulations which show that the proposed algorithm is capable of correctly associating pulse indices to observations, and thereby generating statistically efficient estimates of PRI and phase, even when 99.9% of the pulses are missing from the record, only nine TOAs are recorded and the standard deviation of the measurement error is as high as 1% of the PRI. We also demonstrate through simulation its robustness to imprecise knowledge of the model parameters: measurement noise variance and missing pulse rate.

Finally, in Chapter 7, we summarise the findings of the thesis and the major original contributions it contains, as well as discussing possibilities for further research.

3. How to Read This Thesis

As one would expect, this thesis is intended to be read from start to finish. However, the thesis has been written with two (often) distinct audiences in mind: pure mathematicians and computer scientists with an inclination towards number theory and by engineers and applied mathematicians with an inclination towards signal processing.

It is hoped by the author that the first part of the thesis contains results which might interest number theorists; that the subjects presented in the early chapters are not merely a theoretical preparation for the signal processing applications in the second part but that they are interesting and useful in their own right. Those that have no interest in the practical application of the subject matter might well be content to read only Chapter 2 to Chapter 4.

On the other hand, practitioners of signal processing may not wish to delve deeply into the intricate workings of algorithms for Diophantine approximation, simultaneous Diophantine approximation and lattice reduction. However, there would be little comprehension of the proposed solution to the engineering problems of Chapter 5 and Chapter 6 without at least a superficial understanding of the theory and algorithms of the earlier chapters. Here, the advantages of the "theorem-proof" style of exposition in the fist part of the thesis are manifest. The important properties of the algorithms that are later relied upon can be quickly distilled from the text, highlighted as they are in the theorem statements. This allows for the swift absorption of major results without any need for a thorough understanding of the "how and why." Even so, large portions of the earlier chapters could be skipped entirely. The sections on Cassels' algorithm (Section 5 of Chapter 2), and the theory of (ρ, h)-minimal sets and derived algorithms (Section 3 to Section 5 of Chapter 4) are sections which could be missed by a reader who is not interested by such things, without having an adverse effect on the ability to make sense of the later chapters.

It is also appropriate that we mention here the philosophy applied by the author in deciding when to give proofs and when to omit them. Proofs have been given where they are either short or deemed important in the development of the subject matter. They are presented where the theorem statement is sufficiently different from other theorems which can be found in the literature. They are likewise omitted if the proof is long and not essential. This leads to choices which may appear puzzling at first. For example, in Chapter 3, the proofs of Minkowski's first and second theorem are not given, although they underpin the theory of the geometry of numbers. For our purposes, no knowledge is needed of the proofs and, in any case, they are abundant in the literature. Similarly, in Chapter 2, we take particular care to prove results about best Diophantine approximations in the absolute sense, going to greater lengths than many texts on the subject, but state without proof the theorem which relates the best Diophantine approximations of a real number

in the relative sense with the intermediate fractions of its simple continued fraction expansion. We justify this emphasis on the basis that a thorough understanding of best homogeneous approximations in the absolute sense is required to place us in a position to prove results about Cassels' algorithm and its ability to find best inhomogeneous approximations. This is one of the important original contributions of the thesis. Furthermore, we make much greater use of approximations in the absolute sense than we do of those in the relative sense, as is witnessed in Chapter 5. The author hopes that, in the context of the material presented, the choices will be seen to be appropriate by the time the reader reaches the end of the thesis.

4. The Presentation of Algorithms

This thesis places a heavy emphasis on presenting and analysing algorithms. The algorithms are presented with a Pascal-like syntax. However, they are not fully functional programs. Some details of the algorithms are not completely set out in the text. It is left to the reader to infer from the surrounding text the inputs and outputs of the algorithm, the types of each variable, and the number of arguments and return types of function and procedure calls. However, the author has tried to ensure that this is never a difficult task for the reader.

Our analysis of the algorithms, although at times quite detailed, may not seem to the computer scientist to have gone far enough. Our analysis with regard to the running time of an algorithm never goes further than understanding its ARITHMETIC COMPLEXITY. In this model of computation, each simple arithmetic operation on a real number (multiplication, division, addition, subtraction and rounding) has unit cost. Occasionally, we will be happy just to conclude that an algorithm terminates in a finite amount of time. At other times, we will analyse the algorithm only so far as to determine the number of iterations through a particular loop.

A more thorough approach would have required that we analyse the algorithms according to the BIT COMPLEXITY model. In this model, we analyse the running time of the algorithm according to the number of operations that must be performed by a Turing machine to complete the execution of the algorithm. While this gives a more realistic picture of the running time of an implementation on a real computer for large input sizes, the extra complication it introduces is not warranted here.

Finally, we explain what is meant when we make reference to algorithms being "computationally efficient" or to problems being "computationally infeasible" or "intractable." We recall that computational problems can be divided up into a number of classes, depending on their time and space requirements on a Turing machine or other criteria such as their ability to be parallellised. Two important classes of algorithms are $\mathfrak{P}$ and $\mathfrak{NP}$. A problem which requires a "true or false" answer is in $\mathfrak{P}$ if it can be resolved in the affirmative by a *deterministic* Turing machine in an amount of time which is bounded by a polynomial of the input size. Such a problem is in $\mathfrak{NP}$ if it can be resolved in the affirmative by a *non-deterministic*

Turing machine in an amount of time which is bounded by a polynomial of the input size. Without wishing to discuss in detail the differences between deterministic and non-deterministic Turing machines, we summarise by saying that if we present the algorithm for determination of the problem as a decision tree then it is in $\mathfrak{P}$ if the number of *nodes* in the tree which lead to resolution in the affirmative is bounded by a polynomial of the input size and it is in $\mathfrak{NP}$ in the *depth* of that part of the tree which leads to resolution in the affirmative is bounded in this way. If a problem is in $\mathfrak{P}$ then it is also in $\mathfrak{NP}$. A common assertion is that problems in $\mathfrak{P}$ are "computationally feasible" for execution on a realistic computer. Problems that are in $\mathfrak{NP}$ but not in $\mathfrak{P}$ are likewise "computationally infeasible." It is an open question in computer science as to whether $\mathfrak{P} = \mathfrak{NP}$. Again without giving a full explanation of the definition of the terms, problems which are $\mathfrak{NP}$-complete are thought to be computationally infeasible since if any such problem was found to be in $\mathfrak{P}$ then $\mathfrak{P} = \mathfrak{NP}$. The status of $\mathfrak{NP}$-hard problems is not as clear, but this is usually treated as an indication that the problem is quite possibly computationally infeasible. The reader is referred to, for example, Lewis & Papadimitriou (1981) for a discussion of these subjects.

5. Original Contributions

There are five areas in which the author believes this thesis makes an important, original contribution. In the field of number theory, the theory developed in Chapter 2 relating the outputs of Cassels' algorithm with the sequence of best inhomogeneous Diophantine approximations is, to the author's knowledge, the most comprehensive to have yet appeared.

The algorithms for best simultaneous Diophantine approximation in Chapter 4 with lattices of rank three are novel and able to produce sequences of best simultaneous Diophantine approximations for a more general class of approximation problems than previous algorithms which have been published. It can be used for both best approximation of a line by lattice points ("traditional" simultaneous Diophantine approximation) and approximation of a plane by lattice points (best approximate integer relations). The invention of the *extended norm* to ensure correct operation of the algorithm is also a novel feature of the presentation. Importantly for the applications we have in mind, it can also be used to solve certain coincidence problems involving three pulse trains.

As a bridge between the number theoretic work and the applications in signal processing, the findings regarding the relationship between best Diophantine approximations and the successive maxima of diagonal functions and, in particular, the periodogram are new and original. The findings may lead to a greater involvement of number theoretic algorithms in frequency and spectrum estimation.

The chapter on intercept time (Chapter 5) presents the intercept time problem as a Diophantine approximation problem. The application of that theory to the

interpretation of phenomena is an original contribution of this thesis. Furthermore, the results, as few as they are, for intercepts between more than two pulse trains have not previously appeared in the literature.

Finally, the statistical model of the process of sparse and noisy observation of a periodic pulse train and the subsequent application of a modern algorithm for simultaneous Diophantine approximation (presented in Chapter 6) is an original contribution. The use of an algorithm derived from the LLL algorithm has led to a considerable improvement in our ability to quickly recover information from short, sparse and noisy records.

CHAPTER 2

DIOPHANTINE APPROXIMATION

1. Approximation of a Real Number by Rational Numbers

We consider the problem of finding approximations, in an appropriate sense, to a single real number by rational numbers. This is known as the problem of DIOPHANTINE APPROXIMATION, in honour of DIOPHANTOS of Alexandria, who studied many problems involving rational numbers in his books, *Arithmetica*, written *c.* 300 A.D.

There are a number of ways in which we might define what is meant by a "good" approximation. We will consider three ways of approximating a real number α. In each we shall seek to minimise a function which represents the "nearness" of a rational number p/q from α while keeping the denominator q as small as possible. We call such a function the (absolute value of the) APPROXIMATION ERROR function. The first such function we consider is $\alpha - p/q$ and problems which involve minimisation of the absolute value of this function we call HOMOGENEOUS DIOPHANTINE APPROXIMATION IN THE RELATIVE SENSE. The second such function is $q\alpha - p$, and minimisation of its absolute value we shall refer to as HOMOGENEOUS DIOPHANTINE APPROXIMATION IN THE ABSOLUTE SENSE. Finally, given an additional real number β, we will also consider the minimisation of the absolute value of $q\alpha - p - \beta$, which we call INHOMOGENEOUS DIOPHANTINE APPROXIMATION (in the absolute sense).

If we write our chosen approximation error function $F(p/q)$ then we say that a rational number p/q with positive denominator is a BEST DIOPHANTINE APPROXIMATION (or simply best approximation) in the appropriate sense if, for all other rational numbers p'/q' with positive denominators,

$$q' \leqslant q \;\Rightarrow\; |F(p'/q')| \geqslant |F(p/q)| \tag{1.1}$$

and

$$|F(p'/q')| \leqslant |F(p/q)| \;\Rightarrow\; q' \geqslant q. \tag{1.2}$$

REMARK 1.1. Several variations of the definition of a best approximation are possible. Indeed, a significant generalisation for approximation of multiple real numbers is presented in Chapter 4. However, we remark at this point that it is usual to allow best approximations to have either positive or negative denominators, and to write the inequalities in (1.1) and (1.2) involving q and q' with the same inequalities in terms of $|q|$ and $|q'|$. For homogeneous approximation, this affords no extra

generality, since a rational number p/q can be a best approximation according to the expanded definition if and only if $(-p)/(-q)$ is. For inhomogeneous approximation, however, there is a non-trivial difference between the definitions. For the applications we have in mind in subsequent chapters, the definition we have given is the most appropriate.

For reasons which will become apparent, we will usually express best approximations not as a rational number p/q, but rather as the ordered pair of numerator and denominator (p, q).

We will begin our study of algorithms for Diophantine approximation in Section 2 by discussing some naïve algorithms for best Diophantine approximation. These are algorithms which any mathematically capable person could invent within minutes of learning the problem. Although we present much improved algorithms subsequently, it is nevertheless instructive to examine the behaviour of these obvious algorithms, if only to understand why the improved algorithms are desirable. We will then study Euclid's algorithm in Section 3, which we will show produces (with some trivial exceptions) each and every best homogeneous Diophantine approximation of its input α in the absolute sense, and calculates them very efficiently. In Section 4, we will introduce the *simple continued fraction* expansion of a real number and show that Euclid's algorithm produces the *convergents* of the simple continued fraction expansion of α. We will also obtain running time bounds on the algorithm and indicate how the algorithm can also be made to produce the best homogeneous Diophantine approximations in the relative sense. In Section 5, we will present Cassels' algorithm for inhomogeneous approximation. We will show that it can be used to efficiently find all the best inhomogeneous Diophantine approximations of α with respect another real number β.

At the end of this chapter, we devote two sections to the examination of some related problems. In Section 6, we discuss the successive maxima of certain *diagonal functions* in two variables (see Definition 6.1). We show that this is a generalisation of the Diophantine approximation problem. As an important example, we show how Euclid's algorithm can be used to find successive maxima in a periodogram of three samples with positive amplitudes (see Example 6.2 and subsequent discussion). This relationship is potentially quite important in signal processing because of its application to frequency estimation and spectrum estimation of irregularly sampled data. We will partially extend these results to periodograms with arbitrary numbers of samples in Chapter 6.

In the last section of this chapter, Section 7, we review some elementary properties of *Farey series* with special emphasis on their relationship with Diophantine approximation.

There are two main objectives in this chapter: to acquaint the reader with algorithms for Diophantine approximation as a foundation for our study of signal

processing applications in later chapters and to present original research. The original contributions, as set forth in Section 5 of Chapter 1, are chiefly those which relate to the calculation of best inhomogeneous Diophantine approximations with Cassels' algorithm and to the correspondence between successive maxima of certain diagonal functions and the best Diophantine approximations of a real number. This necessitates an emphasis on specific properties of best Diophantine approximations early in the chapter which is more detailed than would be required if our only purpose was to present a review of the theory.

2. Some Naïve Algorithms for Diophantine Approximation

Consider a naïve approach to finding best homogeneous approximations in the absolute sense to a real number α by a rational number p/q. Clearly, for any q, that value of p which minimises $|q\alpha - p|$ is $p = \lfloor q\alpha \rceil$. This suggests the following algorithm with the approximation error function $F(p,q) = q\alpha - p$.

ALGORITHM 2.1.

```
1  begin
2    q := 1;
3    p := ⌊qα⌉;
4    output(p, q);
5    η* := F(p, q);
6    while η* ≠ 0 do
7          q := q + 1;
8          p := ⌊qα⌉;
9          η := F(p, q);
10         if |η| < |η*| then η* := η; output(p, q); fi;
11   od;
12 end.
```

PROPOSITION 2.1. *If $F(p,q) = q\alpha - p$ where α is a real number then Algorithm 2.1 outputs all best homogeneous approximations to α in the absolute sense, unless α is a half-integer.*

REMARK 2.1. We mean by HALF-INTEGER a number of the form $k + \frac{1}{2}$, $k \in \mathbb{Z}$. Obviously, a half-integer $k + \frac{1}{2}$ has three best approximations: $(k, 1)$, $(k+1, 1)$ and $(2k+1, 2)$. Algorithm 2.1 will miss one of the first two of these, but it will find the other two and then terminate.

PROOF. The proof is by induction. The algorithm finds the best approximation for $q = 1$. If the algorithm has found all best approximations with denominators $n < q$, then on the q^{th} iteration it will find all best approximations with a denominator of q, since there can only be one if α is not a half-integer. Thus it has been proven by induction. □

By the same method of proof, we can prove the following proposition.

PROPOSITION 2.2. *If $F(p,q) = p/q - \alpha$ where α is a real number then Algorithm 2.1 outputs all best homogeneous approximations to α in the relative sense, unless α is a half-integer.*

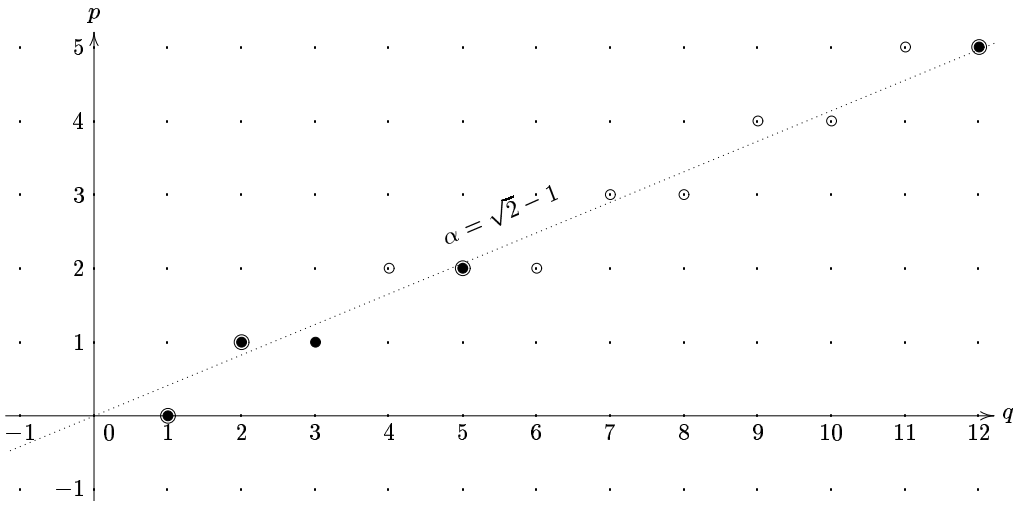


FIGURE 1. Graphical interpretation of the operation of Algorithm 2.1 on the input $\alpha = \sqrt{2}-1$ for homogeneous Diophantine approximation.

Figure 1 illustrates the operation of Algorithm 2.1 on the input $\alpha = \sqrt{2}-1$. Here we see that the real input α can be regarded as the slope of a line passing through the origin. Diophantine approximations to α can be regarded as points with integer coordinates which lie close to the line. The integer points are represented by dots ($\cdot$). Open circles ($\odot$) represent the points which are considered by Algorithm 2.1. Filled circles ($\bullet$) represent the best homogeneous Diophantine approximations to α in the relative sense. Open circles superimposed over filled circles represent the best homogeneous Diophantine approximations to α in the absolute sense.

We now inquire into the number of iterations required by the algorithm to find approximations to a certain accuracy. It is clear that if α is a rational number then the algorithm will terminate after a number of iterations equal to its denominator when expressed in lowest terms. Similarly, it is clear that the algorithm can never terminate if α is irrational. How many iterations of the algorithm are required to produce a best approximation with an absolute approximation error not greater than some $\epsilon > 0$? To answer this question, we present the following famous theorem of DIRICHLET.

THEOREM 2.1. *Given any real numbers $\alpha > 0$ and $Q > 1$, there exists a rational number p/q such that*

$$0 < q < Q \qquad \text{and} \qquad |q\alpha - p| \leqslant \frac{1}{Q}. \tag{2.1}$$

PROOF. The proof makes use of the pigeon-hole principle. Let $k = \lceil Q \rceil$. Consider the set of $k+1$ integer pairs which consists of $(0,0)$, $(-1,0)$ as well as $(\lfloor q\alpha \rfloor, q)$ for $q = 1, \ldots, k-1$. For any element (p,q) of the set, it is clear that $0 \leqslant q\alpha - p \leqslant 1$. We then divide the interval $[0,1]$ into k equal sub-intervals, each of length $1/k$. Since we have $k+1$ elements in our set of integers pairs, there must be a pair of pairs (p_1, q_1) and (p_2, q_2) such that

$$\frac{j}{k} \leqslant q_1\alpha - p_1 \leqslant q_2\alpha - p_2 \leqslant \frac{j+1}{k}$$

for some $0 \leqslant j < k$. Thus,

$$0 \leqslant (q_2 - q_1)\alpha - (p_2 - p_1) \leqslant \frac{1}{k} \leqslant \frac{1}{Q}$$

and $0 < |q_2 - q_1| < k$ which implies that $0 < |q_2 - q_1| < Q$. Therefore, the rational number $(p_2 - p_1)/(q_2 - q_1)$, expressed so that the denominator is positive, satisfies (2.1) and the theorem is proven. □

REMARK 2.2. No significant improvement is possible on the bound $q < Q$ in (2.1). For suppose

$$Q > 3 \qquad \text{and} \qquad \frac{1}{Q} < \alpha < \frac{1}{Q-1}. \tag{2.2}$$

In this case, it is easily shown that

$$\frac{1}{Q} < q\alpha < 1 - \frac{1}{Q}$$

whenever $0 < q < Q - 2$.

We might hope to replace the upper bound $q < Q$ in (2.1) with something substantially smaller when we consider homogeneous Diophantine approximation errors in the relative sense because the approximation errors are smaller by a factor of q. A factor of one half will have to suffice, as the next theorem shows.

THEOREM 2.2. *Given any real numbers $\alpha > 0$ and $Q > 1$, there exists a rational number p/q such that*

$$0 < q < \tfrac{1}{2}Q + 1 \qquad \textit{and} \qquad \left|\alpha - \frac{p}{q}\right| \leqslant \frac{1}{Q}. \tag{2.3}$$

PROOF. If we set $q = \lceil \frac{1}{2}Q \rceil < \frac{1}{2}Q + 1$ and $p = \lfloor q\alpha \rceil$ then

$$\left|\alpha - \frac{p}{q}\right| = \frac{|q\alpha - \lfloor q\alpha \rceil|}{q} \leqslant \frac{1}{2\lceil \frac{1}{2}Q \rceil} \leqslant \frac{1}{Q}.$$

□

REMARK 2.3. Once again, no significant improvement on the upper bound $q < \frac{1}{2}Q + 1$ is possible. For suppose again that we have (2.2). It is then easily shown that

$$\frac{|q\alpha - p|}{q} > \frac{1}{Q}$$

whenever $0 < q < \frac{1}{2}(Q-1)$.

Dirichlet's theorem (Theorem 2.1) implies that we can be sure of finding a best homogeneous approximation to α in the absolute sense with an absolute approximation error of less than ϵ with $0 < \epsilon < 1$ in less than $1/\epsilon$ iterations with $F(p,q) = |q\alpha - p|$. Similarly, we are assured of finding a best homogeneous approximation to α in the relative sense with an absolute approximation error of less than ϵ within $1/(2\epsilon) + 1$ iterations when $F(p,q) = |p/q - \alpha|$.

We now consider the modification of Algorithm 2.1 to make it suitable for inhomogeneous Diophantine approximation of a real number α. If we were to set our approximation error function $F(p,q)$ to $q\alpha - p - \beta$ for real numbers α and β ($\neq 0$) and also to change our expression for choosing p to $p := \lfloor q\alpha - \beta \rceil$ on lines 3 and 8 then the resulting algorithm can be shown to find all best inhomogeneous Diophantine approximations to α with respect to β, except when $\alpha - \beta$ is a half-integer (in which case one and only one is missed). The following theorem implies that the number of best approximations may be infinite.

THEOREM 2.3. *If α is an irrational number and $\beta \neq 0$ and $\epsilon > 0$ are real numbers then there exists a pair of integers (p,q) with $q > 0$ such that*

$$|q\alpha - p - \beta| \leqslant \epsilon \tag{2.4}$$

PROOF. From Dirichlet's theorem, we can find a pair of integers (p', q') such that $|q'\alpha - p'| \leqslant \epsilon$ and $q' > 0$. Let $\delta = q'\alpha - p'$. Now, $\delta \neq 0$ since α is irrational. If δ and β have the same sign then set $k = \lfloor \beta/\delta \rfloor + 1 \geqslant 1$. Then

$$|kq'\alpha - kp' - \beta| \leqslant \epsilon$$

and so $(p,q) = (kp', kq')$ satisfies (2.4). If δ and β have opposite sign then set $k' = \lceil \beta \rceil$ if β is positive or $k' = \lfloor \beta \rfloor$ otherwise. With $k = \lfloor (\beta - k')/\delta \rfloor + 1 \geqslant 1$ we find that

$$|kq'\alpha - kp' + k' - \beta| \leqslant \epsilon.$$

Therefore $(kp' - k', kq')$ satisfies (2.4). □

Theorem 2.3 implies that our modification of Algorithm 2.1 might not terminate if α is irrational. However, whereas we are guaranteed that Algorithm 2.1 as stated for homogeneous approximation will terminate if α is rational, this is not the case for the modification to this algorithm just discussed for inhomogeneous approximations. Consider the case where α is rational. In this case, $q\alpha - \beta - \lfloor q\alpha - \beta \rceil$ can take on only finitely many values. If none of these values are zero then it is clear that our modified Algorithm 2.1 will never terminate while producing only a finite number of best approximations. To overcome this difficulty, we propose the following algorithm, which incorporates some further minor modifications to Algorithm 2.1.

ALGORITHM 2.2.

```
1  begin
2    q := 1;
3    p := ⌊qα⌉; P := ⌊qα − β⌉;
4    output(P, q);
5    η := qα − p; ζ* := qα − P − β;
6    while η ≠ 0 ∧ ζ* ≠ 0 do
7          q := q + 1;
8          p := ⌊qα⌉; P := ⌊qα − β⌉;
9          η := qα − p; ζ := qα − P − β;
10         if |ζ| < |ζ*| then ζ* := ζ; output(P, q); fi;
11   od;
12 end.
```

Observe that Algorithm 2.2 keeps track of both the inhomogeneous approximation error (ζ) as well as the homogeneous approximation error (η). The algorithm terminates as soon as either approximation error becomes zero. We are now assured that Algorithm 2.2 will continue to iterate and find best approximations with successively smaller approximation errors and it will eventually terminate if the number of such best approximations is finite.

If the algorithm terminates with $\eta = 0$ then all the best inhomogeneous approximations have been found. For suppose there were some best inhomogeneous approximation (P', q') with $q' > q$, where q has that value for which the algorithm terminates with $\eta = 0$. Then the pair of integers $(P' - p, q' - q)$ must have the same inhomogeneous approximation error as (P', q') and, clearly, $0 < q' - q < q'$. Therefore, (P', q') cannot possibly be a best inhomogeneous approximation.

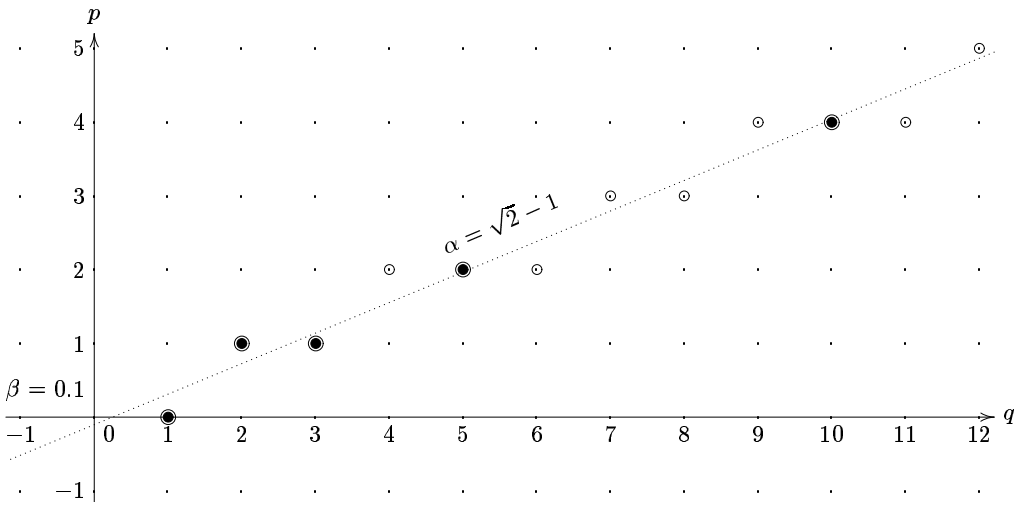


FIGURE 2. Graphical interpretation of the operation of Algorithm 2.2 on the inputs $\alpha = \sqrt{2}-1$ and $\beta = 0.1$ for inhomogeneous Diophantine approximation.

Figure 2 illustrates the operation of Algorithm 2.2 on the input $\alpha = \sqrt{2} - 1$ when $\beta = 0.1$, using the same geometrical interpretation as we used to illustrate the operation of Algorithm 2.1 in Figure 1. The line representing α no longer runs through the origin, as it did in the homogeneous case, but is now offset by the amount β. Open circles superimposed over filled circles represent the best inhomogeneous Diophantine approximations to α with respect to β. Comparing their positions with the positions of the best homogeneous approximations in Figure 1, we can see that they have little in common.

For homogeneous approximation, we have Dirichlet's theorem to provide an *a priori* upper bound on the number of iterations required to produce a sufficiently good approximation, as defined by a maximum approximation error ϵ. Unfortunately, no such upper bound can be given in the inhomogeneous case, as shown by the following theorem, due to KHINCHIN. For a proof, see CASSELS (1957), Theorem III, p. 51.

THEOREM 2.4. *Let $\varphi(q)$ be any positive function of the integer variable q such that*

$$\lim_{q \to \infty} \varphi(q) \to 0.$$

Then there is an irrational α and a real number β such that the pair of inequalities

$$0 < q \leqslant Q \qquad \textit{and} \qquad |q\alpha - p - \beta| < \varphi(Q)$$

has no solution in integers (p, q) for infinitely many values of Q.

3. Euclid's Algorithm

The following is an algorithm which was originally described by EUCLID in Propositions 1 and 2 of Book VII of his *Elements* (HEATH, 1908), although it is almost certainly of earlier origin (KNUTH, 1981).

ALGORITHM 3.1.

```
1 begin
2   r := x; s := y;
3   while r > 0 do
4         if r ⩾ s then r := r − s;
5         else swap(r, s);
6         fi;
7   od;
8 end.
```

Algorithm 3.1 is familiar to most students of mathematics as the greatest common divisor algorithm. Given two positive integer inputs x and y, this algorithm eventually terminates with $r = 0$ and $s = \gcd(x, y)$.

In Propositions 2 and 3 of Book X of *Elements*, EUCLID proposes essentially the same algorithm to determine the "greatest common measure" of two "commensurable magnitudes," if one exists. In modern terms, what is meant is that Algorithm 3.1 can be given positive real inputs x and y and if there exists some (greatest) real number δ such that

$$m\delta = x \qquad \text{and} \qquad n\delta = y \tag{3.1}$$

where m and n are both integers then the algorithm will terminate with $r = 0$ and $s = \delta$. The number δ is then the greatest common measure. The algorithm terminates if and only if x and y have a greatest common measure. This statement is equivalent to the statement that the algorithm terminates if and only if x/y is a rational number.

Clearly, the algorithm could be made significantly faster if we replaced the repeated subtraction with division. Also, Algorithm 3.1 makes no attempt to calculate the integers m and n as defined in (3.1) which relate the greatest common measure to the inputs x and y. The algorithm we now present can be regarded as an improved version of Algorithm 3.1 which remedies these two deficiencies for the inputs $x = \alpha$ and $y = 1$. The remainder of this section will be devoted to showing that the intermediate calculations performed by the algorithm yield integers which have special properties with regard to the Diophantine approximation of α.

ALGORITHM 3.2.

begin
  $\eta_{-1} := -1;\ \eta_{-2} = \alpha;$
  $p_{-1} := 1;\ p_{-2} := 0;$
  $q_{-1} := 0;\ q_{-2} := 1;$
  $n := 0;$
  while $\eta_{n-1} \neq 0$ **do**
    $a_n := \left\lfloor \dfrac{-\eta_{n-2}}{\eta_{n-1}} \right\rfloor;$
    $p_n := p_{n-2} + a_n p_{n-1};$
    $q_n := q_{n-2} + a_n q_{n-1};$
    $\eta_n := \eta_{n-2} + a_n \eta_{n-1};$
    $n := n + 1;$
  od;
end.

PROPOSITION 3.1. *For each $n \geqslant 0$ for which $\eta_n \neq 0$, Algorithm 3.2 calculates values for η_n, a_n, p_n and q_n so that*

(*i*) $\eta_n = q_n\alpha - p_n$,
(*ii*) $\eta_n(\eta_{n-1} + k\eta_n) \leqslant 0$ *if and only if* $k \leqslant a_{n+1}$, $k \in \mathbb{Z}$,
(*iii*) $\eta_n(\eta_n + \eta_{n-1}) < 0$,

(*iv*) $\eta_n\eta_{n-1} < 0$,
(*v*) $a_{n+1} > 0$,
(*vi*) $q_n > 0$,
(*vii*) $\eta_{n+1}\big(\eta_{n+1} - \frac{1}{2}\eta_{n-1}\big) < 0$ *unless* $\eta_{n+1} = 0$ *and*
(*viii*) $p_{n+1}q_n - p_nq_{n+1} = \eta_nq_{n+1} - \eta_{n+1}q_n = (-1)^n$.

PROOF. The proof of (*i*) is by inspection of Algorithm 3.2. Clearly, $\eta_{-2} = q_{-2}\alpha - p_{-2}$ and $\eta_{-1} = q_{-1}\alpha - p_{-1}$. The expressions for updating p_n, q_n and η_n all have the same form, so the condition $\eta_n = q_n\alpha - p_n$ will always be maintained (even if η_n becomes zero).

Consider statement (*ii*). If $k \leqslant a_{n+1}$ then $k \leqslant -\eta_{n-1}/\eta_n$ and so $\eta_n(\eta_{n-1} + k\eta_n) \leqslant 0$. If $k > a_n$ then $k > -\eta_{n-1}/\eta_n$ and so $\eta_n(\eta_{n-1} + k\eta_n) > 0$. Thus statement (*ii*) is true. Notice that it is also true when $n = -1$.

Using the obvious inequality

$$\frac{-\eta_{n-2}}{\eta_{n-1}} - 1 < a_n \leqslant \frac{-\eta_{n-2}}{\eta_{n-1}}, \tag{3.2}$$

we find that, for $n \geqslant 0$,

$$\eta_{n-1}\eta_n = \eta_{n-1}\eta_{n-2} + a_n\eta_{n-1}^2 < 0 \tag{3.3}$$

and similarly

$$\eta_{n-1}(\eta_n + \eta_{n-1}) > 0. \tag{3.4}$$

From (3.3) and (3.4) we have $\eta_{n-1}^2\eta_n(\eta_n + \eta_{n-1}) < 0$ and therefore statement (*iii*) is true for any $n \geqslant 0$.

Statement (*iii*) implies (*iv*) and (*ii*) and (*iii*) together imply (*v*). Furthermore, (*v*) together with the observation that $q_0 = 1$ implies (*vi*).

Consider (*vii*). If $a_{n+1} = 1$ then (3.2) implies that $-\eta_{n-1}/\eta_n < 2$ which implies that $\eta_n^2 > -\frac{1}{2}\eta_n\eta_{n-1}$. Now, if $a_{n+1} = 1$ then $\eta_{n+1} = \eta_{n-1} + \eta_n$ which implies that

$$\eta_n\big(\eta_{n+1} - \tfrac{1}{2}\eta_{n-1}\big) = \eta_n^2 + \tfrac{1}{2}\eta_n\eta_{n-1} > 0. \tag{3.5}$$

By application of (*iv*) with n replaced by $n + 1$ we see that, unless $\eta_{n+1} = 0$, $\eta_n\eta_{n+1} < 0$ so, in combination with (3.5), we have

$$\eta_n^2\eta_{n+1}\big(\eta_{n+1} - \tfrac{1}{2}\eta_{n-1}\big) < 0 \tag{3.6}$$

which implies (*vii*) since $\eta_n^2 > 0$.

On the other hand, if $a_{n+1} > 1$ then (3.2) implies that $-\eta_{n-1}/\eta_n \geqslant 2$ which implies that $\eta_n^2 \leqslant -\frac{1}{2}\eta_n\eta_{n-1}$. Therefore,

$$\begin{aligned}\eta_n\eta_{n+1} &= \eta_n(\eta_{n-1} + a_{n+1}\eta_n)\\ &> \eta_n\eta_{n-1} - \left(\frac{\eta_{n-1}}{\eta_n} + 1\right)\eta_n^2\\ &= -\eta_n^2 \geqslant \tfrac{1}{2}\eta_n\eta_{n-1}\end{aligned}$$

and so again

$$\eta_n\big(\eta_{n+1} - \tfrac{1}{2}\eta_{n-1}\big) > 0$$

which, through (3.6), implies (*vii*).

We prove statement (*viii*) by induction. The statement is clearly true for $n = -2$. Suppose it is true for all $n < N$ and $N > -2$. Then

$$\begin{aligned} p_{N+1}q_N - p_Nq_{N+1} &= (p_{N-1} + a_{N+1}p_N)q_N - p_N(q_{N-1} + a_{N+1}q_N) \\ &= -(p_Nq_{N-1} - p_{N-1}q_N) \end{aligned}$$

and

$$\begin{aligned} \eta_Nq_{N+1} - \eta_{N+1}q_N &= \eta_N(q_{N-1} + a_{N+1}q_N) - (\eta_{N-1} + a_{N+1}\eta_N)q_N \\ &= -(\eta_{N-1}q_N - \eta_Nq_{N-1}) \end{aligned}$$

and so the statement is true for $n = N$ also. □

COROLLARY 3.1. *For each $n \geqslant 0$ for which $\eta_n \neq 0$, Algorithm 3.2 calculates values for p_n, q_n and a_n such that*

$$\left(\alpha - \frac{p_n}{q_n}\right)\left(\alpha - \frac{p_{n-1} + kp_n}{q_{n-1} + kq_n}\right) \leqslant 0$$

whenever $0 < k \leqslant a_{n+1}$.

PROOF. The corollary is a direct consequence of statements (*ii*) and (*vi*) of Proposition 3.1. □

We shall have frequent recourse to the following fact.

FACT 3.1. *If x, a, b and c are real numbers such that $ac \leqslant 0$, $bc \geqslant 0$ and*

$$(x - a)(x - b) \leqslant 0 \tag{3.7}$$

then

$$(x - a)(x - b - c) \leqslant (x - a)(x - b) \tag{3.8}$$

PROOF. The proof is trivial if $c = 0$ so suppose $c \neq 0$. Suppose the fact is untrue and (3.7) is satisfied but (3.8) is not. By subtraction of the left hand side of (3.8) from the right hand side, we find that $c(x - a) < 0$ which implies that

$$cx < ac \leqslant 0. \tag{3.9}$$

But

$$c^2(x - a)(x - b) = (cx - ca)(cx - cb) \leqslant 0$$

which implies that $c(x - b) \geqslant 0$ and therefore $cx \geqslant bc \geqslant 0$, contradicting (3.9). □

PROPOSITION 3.2. *If Algorithm 3.2 is run on an input real number α then, for each $n \geqslant 0$ for which $\eta_n \neq 0$, it is true that*

(*i*) $\eta_n > 0$ *if* n *is even,*
(*ii*) $\eta_n < 0$ *if* n *is odd,*
(*iii*) $|\eta_n| < |\eta_{n-1}|$ *and*
(*iv*) $|\eta_{n+1}| < \frac{1}{2}|\eta_{n-1}|$.

PROOF. Statements (*i*) and (*ii*) follow from statement (*iv*) of Proposition 3.1, observing that $\eta_{-1} = -1$.

Using Fact 3.1 with $x = \eta_n$, $a = -\eta_{n-1}$, $b = 0$ and $c = \eta_{n-1}$ and using (*iii*) of Proposition 3.1, we see that $\eta_n(\eta_n + \eta_{n-1}) < 0$ implies that $(\eta_n - \eta_{n-1})(\eta_n + \eta_{n-1}) < 0$. Thus, $\eta_n^2 - \eta_{n-1}^2 < 0$ and (*iii*) follows.

We use Fact 3.1 and statement (*vii*) of Proposition 3.1 in an analogous fashion to show (*iv*). □

COROLLARY 3.2. *If Algorithm 3.2 is run on an input real number α then it either terminates after a finite number of iterations with $\eta_n = 0$ or $\lim_{n\to\infty} \eta_n = 0$.*

We can now see that Algorithm 3.2 very quickly finds good homogeneous Diophantine approximations of α in the absolute sense. Clearly,

$$|\eta_n| < 2^{-\lceil n/2 \rceil},$$

so to find a pair of integers (p, q) with an approximation error, $|q\alpha - p| \leqslant \epsilon$, for some $\epsilon > 0$, requires at most $2\lceil \log_2(1/\epsilon) \rceil$ iterations of Algorithm 3.2. In Section 4, we will show that an even smaller bound is possible.

COROLLARY 3.3. *If Algorithm 3.2 is run on an input real number α and it terminates with $\eta_n = 0$ and $n > 0$ then $a_n \geqslant 2$.*

PROOF. From statement (*v*) of Proposition 3.1, we know that $a_n \geqslant 1$. If $a_n = 1$ then $\eta_n = \eta_{n-2} + \eta_{n-1} = 0$ but $|\eta_{n-2}| > |\eta_{n-1}|$ from statement (*iii*) of Proposition 3.2, so this is impossible. □

We now enquire into the relationship between the integers (p_n, q_n) generated by the algorithm and the best homogeneous Diophantine approximations in the absolute sense. We will find that the relationship is almost one-to-one.

PROPOSITION 3.3. *Suppose Algorithm 3.2 is executed with the real number α as its input. Suppose (p, q) are integers which are not both zero and let $\eta = q\alpha - p$. If $\eta_n \neq 0$ and*

$$(\eta - \eta_n)(\eta - \eta_{n-1} - k\eta_n) < 0 \tag{3.10}$$

for some $n \geqslant 0$ and $0 \leqslant k \leqslant a_{n+1}$ then either $q < 0$ or $q \geqslant q_{n-1} + (k+1)q_n$.

Before proving Proposition 3.3 we make some important observations about the proposition in the form of a lemma.

LEMMA 3.1. *Consider the statement of Proposition 3.3 and the approximation errors, η_n, calculated by Algorithm 3.2 for the real input α. It is true that*

(*i*) *the statement of Proposition 3.3 for $n = N > 0$ and $k = 0$ is equivalent to its statement for $n = N - 1$ and $k = a_N$,*

(*ii*) *satisfaction of* (3.10) *for $n = N \geqslant 0$ and $k = K$ where $0 < K \leqslant a_{N+1}$ implies satisfaction of* (3.10) *for $n = N$ and $k = K - 1$,*

(*iii*) *if*

$$\eta(\eta + \eta_{N-1} + \kappa\eta_N) < 0 \tag{3.11}$$

for $N \geqslant 0$ and $0 \leqslant \kappa \leqslant a_{n+1}$ then $q \neq 0$ and (3.10) *is satisfied for $n = N-1$ and $k = a_N - 1$.*

PROOF. We see that (*i*) follows from the fact that $\eta_N = \eta_{N-2} + a_N\eta_N$ and $q_{N-2} + (a_N + 1)q_{N-1} = q_N + q_{N-1}$.

That $(\eta - \eta_{N-1} - K\eta_N)(\eta - \eta_N) < 0$ implies $[\eta - \eta_{N-1} - (K-1)\eta_N](\eta - \eta_N) < 0$ can be confirmed by application of Fact 3.1 with $x = \eta$, $a = \eta_N$, $b = \eta_{N-1} + K\eta_N$ and $c = -\eta_N$. Clearly, $ac < 0$ and $bc > 0$ by virtue of statement (*iii*) of Proposition 3.1 since $K \leqslant a_{N+1}$. Thus, we have shown that (*ii*) is true.

Consider (*iii*). Using Fact 3.1 with $x = \eta$, $a = 0$, $b = -\eta_{N-1} - \kappa\eta_N$ and $c = \kappa\eta_N$, we see that (3.11) implies that $\eta(\eta + \eta_{N-1}) < 0$, which in turn implies that $|\eta| < |\eta_{N-1}| \leqslant 1$ and so $q \neq 0$.

Similarly, we can use Fact 3.1 with $x = \eta$, $a = 0$, $b = -\eta_{N-1} - \kappa\eta_N$ and $c = (\kappa + 1)\eta_N$ to show that (3.11) implies that $\eta(\eta + \eta_{N-1} - \eta_N) < 0$ which is equivalent to

$$\eta[\eta - \eta_{N-2} - (a_N - 1)\eta_{N-1}] < 0. \tag{3.12}$$

Using Fact 3.1 again with $x = \eta$, $a = \eta_{N-2} + (a_N - 1)\eta_{N-1}$, $b = 0$ and $c = \eta_{N-1}$ we see that (3.12) implies that $[\eta - \eta_{n-2} - (a_N - 1)\eta_{N-1}](\eta - \eta_{N-1}) < 0$, which is simply (3.10) for $n = N - 1$ and $k = a_N - 1$. Therefore, we have proved (*iii*). □

PROOF OF PROPOSITION 3.3. The proof is by induction. The proposition is obviously true for $n = 0$ and $k = 0$ because satisfaction of (3.10) is equivalent to satisfaction of $-1 < \eta < \alpha - \lfloor\alpha\rfloor < 1$. Therefore $q \neq 0$ because p and q are not both zero, which is the implication of the proposition.

Now, suppose the proposition is true for all $0 \leqslant k < K$ when $n = N \geqslant 0$ and $0 < K \leqslant a_{N+1}$ and for all $n < N$ if $N > 0$. Suppose (p, q) is an integer pair which satisfies the conditions of the proposition for $n = N$ and $k = K$. Because of statement (*ii*) of Lemma 3.1, we know that $q < 0$ or $q \geqslant q_{N-1} + Kq_N$. Let

$$(p', q') = (p - p_{N-1} - Kp_N, q - q_{N-1} - Kq_N).$$

Clearly, p' and q' are not both zero. By substitution of $\eta' = q'\alpha - p' = \eta - \eta_{N-1} - K\eta_N$ into (3.10), we find that

$$\eta'[\eta' + \eta_{N-1} + (K-1)\eta_N] < 0.$$

We observe that this is just (3.11) from statement (iii) of Lemma 3.1 with η replaced by η' and κ replaced by $K-1$. Therefore $q' \neq 0$. If $N = 0$ then this implies that $q < 0$ or $q \geqslant q_{-1} + (K+1)q_0$ and the proposition is true for $k = K$ also. If $N > 0$ then, because statement (iii) implies that (3.10) is satisfied with η replaced by η' for $n = N-1$ and $k = a_N - 1$, we can see that the proposition can be applied to (p', q') with $n = N-1$ and $k = a_N - 1$ and therefore $q' \geqslant q_{N-2} + a_N q_{N-1} = q_N$. Hence, $q < 0$ or $q \geqslant q_{N-1} + (K+1)q_N$ and the proposition is true for $k = K$ also.

To complete the induction, we observe statement (i) of Lemma 3.1. □

PROPOSITION 3.4. *Suppose Algorithm 3.2 is executed with the real number α as its input. Suppose (p, q) are integers with $q > 0$ and let $\eta = q\alpha - p$. If $\eta_n \neq 0$ and* (3.10) *is satisfied for some $n \geqslant 0$ and $0 \leqslant k < a_{n+1}$ then either $(p, q) = (p^*, q^*)$, where (p^*, q^*) are defined as*

$$(p^*, q^*) = (p_{n-1} + (k+1)p_n, q_{n-1} + (k+1)q_n),$$

or $q > q^*$.

PROOF. First we show that (p^*, q^*) yields a solution to (3.10). With $\eta^* = q^*\alpha - p^*$ we find that

$$\eta^* = \eta_{n-1} + (k+1)\eta_n$$

and so

$$(\eta^* - \eta_{n-1} - k\eta_n)(\eta^* - \eta_n) = \eta_n(\eta_{n-1} + k\eta_n) < 0$$

because of statement (ii) of Proposition 3.1 and because $k < a_{n+1}$.

Now, Proposition 3.3 implies that $q \geqslant q^*$. So if we let

$$(p', q') = (p - p^*, q - q^*)$$

then $q' \geqslant 0$. After substitution of η' for η in (3.10) we have

$$(\eta' + \eta_n)(\eta' + \eta_{n-1} + k\eta_n) < 0. \tag{3.13}$$

Using Fact 3.1 successively, we find that (3.13) implies that $(\eta' + \eta_n)(\eta' + \eta_{n-1}) < 0$ which implies that $(\eta' - \eta_{n-1})(\eta' + \eta_{n-1}) < 0$. Thus $|\eta'| < |\eta_{n-1}| \leqslant 1$. If $q' = 0$ then $|\eta'| = |p'| > 1$ unless $p' = 0$ also. Therefore $(p, q) = (p^*, q^*)$ or $q > q^*$ and the proposition is proved. □

PROPOSITION 3.5. *Suppose Algorithm 3.2 is executed with the real number α as its input. Suppose (p, q) are integers which are not both zero and let $\eta = q\alpha - p$. If $n \geqslant 0$ and $|\eta| < |\eta_n|$ then $(p, q) = (p_{n+1}, q_{n+1})$ or $q < 0$ or $q > q_{n+1}$.*

PROOF. Now, $|\eta| < |\eta_n|$ implies that

$$(\eta - \eta_n)(\eta + \eta_n) < 0.$$

Using Fact 3.1 with $x = \eta$, $a = \eta_n$, $b = -\eta_n$ and $c = \eta_{n-1} + a_{n+1}\eta_n = \eta_{n+1}$ we find that

$$(\eta - \eta_n)[\eta - \eta_{n-1} - (a_{n+1} - 1)\eta_n] < 0.$$

Hence, the conditions of Proposition 3.3 and Proposition 3.4 are satisfied so $(p, q) = (p_{n+1}, q_{n+1})$ or $q < 0$ or $q > q_{n+1}$. □

PROPOSITION 3.6. *If (p, q) is a best homogeneous Diophantine approximation of α in the absolute sense and α is not a half-integer then $(p, q) = (p_n, q_n)$ for some $n \geqslant 0$, where the (p_n, q_n) are those which are calculated by Algorithm 3.2 for the input α. Furthermore, if (p_n, q_n) is a pair of integers calculated by Algorithm 3.2 for some real input α and $n > 0$ then (p_n, q_n) is a best approximation of the aforementioned type.*

PROOF. Suppose (p, q) is a best approximation but not one of the (p_n, q_n). Let $\eta = q\alpha - p$. Now, $|\eta| \leqslant \alpha - \lfloor\alpha\rfloor$ otherwise $(p_0, q_0) = (\lfloor\alpha\rfloor, 1)$ has a smaller absolute approximation error and $q_0 \leqslant q$ since, by definition, $q > 0$. Let N be the largest index such that $|\eta| < |\eta_{N-1}|$. Clearly, $N \geqslant 0$. If $N > 0$ then Proposition 3.5 furnishes a contradiction. If $N = 0$ and $(p, q) \neq (p_0, q_0)$ is a best approximation then $(p, q) = (\lceil\alpha\rceil, 1)$. Since $\eta = \eta_0$, we conclude that $\eta = \eta_0 = \frac{1}{2}$ and so α must be a half-integer.

Now, suppose (p_N, q_N) is not a best approximation for some $N > 0$. Therefore, there must exist some pair of integers (p, q), with $q > 0$ and $\eta = q\alpha - p$, such that

$$q \leqslant q_N \qquad \text{and} \qquad |\eta| \leqslant |\eta_N| \tag{3.14}$$

and one of these inequalities must be satisfied strictly. The right-hand inequality of (3.14), together with statement (*iii*) of Proposition 3.2, implies that $|\eta| < |\eta_{N-1}|$. Proposition 3.5 again furnishes a contradiction. □

REMARK 3.1. Consider the behaviour of Algorithm 3.2 when presented with a half-integer, say $\alpha = m + \frac{1}{2}$. The algorithm terminates after two iterations and calculates $(p_0, q_0) = (m, 1)$, $(p_1, q_1) = (2m + 1, 2)$. Note that $(m + 1, 1)$ is a best approximation of α, but does not appear amongst the (p_n, q_n), $n = 0, 1$.

REMARK 3.2. We remark that (p_0, q_0) might not necessarily be a best approximation of α. If $\alpha - \lfloor\alpha\rfloor > \frac{1}{2}$ then $(p_0, q_0) = (\lfloor\alpha\rfloor, 1)$ is not a best approximation since $(p_1, q_1) = (\lceil\alpha\rceil, 1)$ is better.

In this section, we have shown that Algorithm 3.2 produces a sequence of integer pairs (p_n, q_n) which, with some minor exceptions noted in Proposition 3.6 and

subsequent Remarks, consists of each and every best homogeneous Diophantine approximation in the absolute sense to its input real number. We have also shown in Proposition 3.2 (and subsequent discussion) that the number of iterations required to produce a pair of integers yielding an absolute approximation error not greater than some specified number ϵ is logarithmic in $1/\epsilon$. Clearly, this is a great improvement over the naïve Algorithm 2.1 of the previous section, which requires up to $1/\epsilon$ iterations for the same task.

4. Simple Continued Fractions

Consider a (possibly infinitely) continued fraction of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}} \tag{4.1}$$

where the a_i are independent variables called the PARTIAL QUOTIENTS. We will occasionally use the notation $[a_0, a_1, a_2, \ldots]$ as a convenient way of expressing a continued fraction in the form of (4.1) in terms of the prescribed partial quotients.

In the case where $a_0 \in \mathbb{Z}$ and $a_i \in \mathbb{N}$ for all $i > 0$, the resulting continued fraction is called a SIMPLE CONTINUED FRACTION(S.C.F.).

If we truncate the s.c.f. at the n^{th} partial quotient and express the resulting fraction in its lowest terms then the rational number, p_n/q_n, which results is called the n^{th} CONVERGENT (for reasons which will soon become clear). We say that an s.c.f. is an EXPANSION of a real number α if it evaluates to α when the s.c.f. consists of only a finite number of partial quotients or if

$$\lim_{n\to\infty} \frac{p_n}{q_n} = \alpha$$

otherwise. In our discussion of simple continued fractions, we impose the additional restriction that an s.c.f. which terminates but does not simply consist of the first term a_0 shall have a final partial quotient greater than one. This allows us to state the following theorem, which summarises a number of classical results which can all be found in, for example, KHINCHIN (1964) or HARDY & WRIGHT (1979).

THEOREM 4.1. *Each real number has a unique s.c.f. expansion. Each s.c.f. evaluates or converges to a unique real number. Furthermore, every rational number has a finite s.c.f. expansion and every infinite s.c.f. converges to an irrational number.*

We adopt the convention that the $(-1)^{\text{th}}$ and $(-2)^{\text{th}}$ convergents of any s.c.f. are given by $p_{-1} = q_{-2} = 1$ and $p_{-2} = q_{-1} = 0$. This may seem somewhat arbitrary (and reminiscent of Algorithm 3.2!), but it allows for significant simplification in many discussions of the s.c.f. and we demonstrate this in the following theorem.

THEOREM 4.2. *If we express the s.c.f. of α in the form of* (4.1) *then the n^{th} convergent of α, p_n/q_n, satisfies*

$$\frac{p_n}{q_n} = \frac{p_{n-2} + a_n p_{n-1}}{q_{n-2} + a_n q_{n-1}} \tag{4.2}$$

for all $n \geqslant 0$.

PROOF. The theorem is obviously true for $n = 0$. We complete the proof by induction. Suppose the statement of the theorem is true for all $0 \leqslant n < N$ then the N^{th} convergent can be obtained by using (4.2) for the $(N-1)^{\text{th}}$ convergent with the partial quotient $a_{N-1} + (1/a_N)$. This gives

$$\begin{aligned}
\frac{p_N}{q_N} &= \frac{p_{N-3} + \left(a_{N-1} + \dfrac{1}{a_N}\right) p_{N-2}}{q_{N-3} + \left(a_{N-1} + \dfrac{1}{a_N}\right) q_{N-2}} \\
&= \frac{p_{N-2} + a_N p_{N-3} + a_{N-1} a_N p_{N-2}}{q_{N-2} + a_N q_{N-1} + a_{N-1} a_N p_{N-2}} \\
&= \frac{p_{N-2} + a_N p_{N-1}}{q_{N-2} + a_N q_{N-1}},
\end{aligned}$$

which is simply (4.2) for $n = N$. □

PROPOSITION 4.1. *Algorithm 3.2 generates the complete sequence of partial quotients, a_n, and convergents, p_n/q_n, of the s.c.f. expansion of α.*

PROOF. If we make the observation that initially we can write

$$\alpha = [\alpha] = \left[k, \frac{1}{\alpha - k}\right],$$

where we use the $[\cdot]$ notation to denote the continued fraction with the prescribed partial quotients, $k \in \mathbb{Z}$ and $k \neq \alpha$, then by choosing $a_0 = k = \lfloor \alpha \rfloor$ we can ensure that $1/(\alpha - a_0) > 1$. If α is an integer then the s.c.f. expansion for α is trivial. Provided α is not an integer, and setting $\xi_1 = 1/(\alpha - a_0)$, we can repeat the procedure by choosing $a_1 = \lfloor \xi_1 \rfloor$. If ξ is an integer, then $\alpha = [a_0, a_1]$. Otherwise, we find that

$$\alpha = \left[a_0, a_1, \frac{1}{\xi_1 - a_1}\right]$$

and the last partial quotient is assigned to ξ_2 and again $\xi_2 > 1$. This procedure can be repeated, finding successive ξ_n and forming the partial quotients from their integer parts until, for some n, ξ_n is an integer, should this ever occur. Up to the terminating step,

$$\alpha = [a_0, a_1, \ldots, a_{n-1}, \xi_n].$$

We now show that $\xi_n = -\eta_{n-2}/\eta_{n-1}$, where the η_n are those which are calculated by Algorithm 3.2 for the input α. If we make the assignment $\xi_0 = \alpha$ then it is clearly true for $n = 0$. If it is true for all $n < N$ then

$$\begin{aligned}\xi_N &= \frac{1}{\xi_{N-1} - a_{N-1}} \\ &= \frac{-\eta_{N-2}}{\eta_{N-3} + a_{N-1}\eta_{N-2}} \\ &= \frac{-\eta_{N-2}}{\eta_{N-1}}\end{aligned}$$

and thus it is true for $n = N$ also, and so it is shown by induction.

We use Theorem 4.2 to show that expressions for p_n and q_n are then correct. It remains to show that the s.c.f. so developed converges to α. This is a direct consequence of Corollary 3.2. □

Consider the growth rate of the convergents and the rate of decay of the approximation errors from one iteration of Algorithm 3.2 to the next for positive real inputs, $\alpha > 0$. To understand the growth rate of the convergents, we introduce the FIBONACCI NUMBERS F_n. These are defined by $F_0 = 0$, $F_1 = 1$ and $F_n = F_{n-2}+F_{n-1}$ for $n > 1$.

LEMMA 4.1. *The n^{th} Fibonacci number, F_n, can be expressed as*

$$F_n = \frac{\gamma^n - (-\gamma)^{-n}}{\sqrt{5}} \tag{4.3}$$

where

$$\gamma = \frac{\sqrt{5}+1}{2}$$

is the GOLDEN RATIO.

PROOF. We can verify this directly for $n = 0$ and $n = 1$. To show (4.3) holds for all n we use induction. Suppose it were true for all $n < N$. Making use of the identity $\gamma = 1 + \gamma^{-1}$, we find that

$$\begin{aligned}F_N &= F_{N-1} + F_{N-2} \\ &= \frac{\gamma^{N-1} - (-\gamma)^{1-N} + \gamma^{N-2} - \left(-\gamma^{2-N}\right)}{\sqrt{5}} \\ &= \frac{\gamma^{N-1}(1+\gamma^{-1}) - (-\gamma)^{1-N}(1-\gamma)}{\sqrt{5}} \\ &= \frac{\gamma^N - (-\gamma)^{-N}}{\sqrt{5}}\end{aligned}$$

and so (4.3) is satisfied for $n = N$ also and by induction, for all $n \geqslant 0$. □

THEOREM 4.3. *For any $\alpha > 0$, the numerator, p_n, and denominator, q_n, of the n^{th} partial quotient, $n \geqslant 0$, satisfy*

$$p_n \geqslant F_n \qquad \text{and} \qquad q_n \geqslant F_{n+1}$$

where F_n is the n^{th} Fibonacci number.

PROOF. The proof follows directly from consideration of (4.2) in Theorem 4.2. Since $a_0 \in \mathbb{N}_0$ and $a_n \in \mathbb{N}$ for all $n > 0$, we minimise the growth of the numerator and denominator at each stage if the convergent a_0 is 0 and $a_n = 1$ for each $n > 0$. This implies $p_n = p_{n-2} + p_{n-1}$ and $q_n = q_{n-2} + q_{n-1}$ for all $n > 0$ and $p_0 = F_0$, $p_1 = q_0 = F_1$ and $q_1 = F_2$. The proof is therefore complete. □

REMARK 4.1. The partial quotients described in the proof of Theorem 4.3 belong to the s.c.f. which converges to γ^{-1}. That is,

$$\gamma^{-1} = [0, 1, 1, 1, \ldots].$$

In order to understand the rate of decay of the approximation errors, we first recall DIRICHLET's theorem (Theorem 2.1). The following two theorems are related to that theorem but they offer an improved bound. They also lead to an improved upper bound on the running time required by Algorithm 3.2 to produce best approximations with a specified maximum absolute approximation error. The theorems are not proved here since they are somewhat incidental. The proofs can be found in HARDY & WRIGHT (1979).

THEOREM 4.4. *For any real number $\alpha > 0$ there is an infinite number of rational numbers p/q which satisfy the inequality*

$$|q\alpha - p| \leqslant \frac{1}{\sqrt{5}q}. \tag{4.4}$$

THEOREM 4.5. *Of any three consecutive convergents in the s.c.f. expansion of a real number $\alpha > 0$, at least one satisfies* (4.4).

We can combine the result of Theorem 4.5 with that of Theorem 4.3 to obtain the following corollary.

COROLLARY 4.1. *Any three consecutive approximation errors, η_n, in the s.c.f. expansion of a real number, $\alpha > 0$, satisfy*

$$\min\{F_{n+1}|\eta_n|, F_{n+2}|\eta_{n+1}|, F_{n+3}|\eta_{n+2}|\} \leqslant \frac{1}{\sqrt{5}}. \tag{4.5}$$

REMARK 4.2. The approximation errors in the s.c.f. expansion of γ^{-1} asymptotically achieves the upper bound prescribed by (4.5). Therefore, the constant $1/\sqrt{5}$ cannot be replaced by any smaller number. To see this, we merely observe that the

approximation errors $|\eta_n|$ for the s.c.f. expansion of γ^{-1} are given by $|\eta_n| = \gamma^{-n+1}$ (which is easily proved by induction) and therefore

$$F_{n+1}|\eta_n| = \frac{1 - (-\gamma)^{-2n-2}}{\sqrt{5}}.$$

We can use Corollary 4.1 to improve on the bound we found in the previous section for the number of iterations required to find best approximations with an absolute approximation error not greater than $\epsilon > 0$. From Corollary 4.1, the number of iterations required is $N+3$ (after accounting for the fact that one iteration is required to calculate the 0^{th} convergent) where N is an integer such that

$$\frac{1}{\sqrt{5}F_{N+1}} \leqslant \epsilon.$$

This implies that

$$\gamma^{N+1} - (-\gamma)^{-N-1} \geqslant \frac{1}{\epsilon}.$$

Since $(-\gamma)^{-N-1} < 1$ for all $N \geqslant 0$, we have

$$N \geqslant \log_\gamma\left(\frac{1}{\epsilon} + 1\right) - 1.$$

Therefore, we require $\left\lceil \log_\gamma(\epsilon^{-1} + 1)\right\rceil + 2$ iterations. This implies that the number of iterations is approximately $2.078\ln(1/\epsilon)+2$ when ϵ is small. The bound we found in the previous section, that the number of iterations is at most $2\lceil\log_2(1/\epsilon)\rceil$, is approximately $2.885\ln(1/\epsilon)$. Thus, our new upper bound is appreciably smaller for small ϵ (and close to smallest possible).

For completeness, we conclude this section by stating (without proof) the following result from KHINCHIN (1964) concerning the relationship between the convergents of the s.c.f. and best homogeneous Diophantine approximations in the relative sense. Before the statement of the theorem, we define the INTERMEDIATE FRACTIONS (or INTERMEDIATE CONVERGENTS) of a s.c.f. These are fractions of the form

$$\frac{p_{n-2} + kp_{n-1}}{q_{n-2} + kp_{n-1}}$$

and they are defined for all $n \geqslant 0$ for which $a_n > 1$ and, of those, for all $0 < k < a_n$. (We have already encountered these fractions. They appeared in Proposition 3.3, which allowed us to prove the nearly complete correspondence between the convergents and the best homogeneous approximations in the absolute sense in Proposition 3.6.)

THEOREM 4.6. *Every best homogeneous Diophantine approximation in the relative sense to a real number α is a convergent or an intermediate fraction of the s.c.f. expansion of α.*

REMARK 4.3. It should be noted that the converse is not true in general. Consider the s.c.f. expansion of π,

$$\pi = [3, 7, 15, 1, 292, 1, \ldots].$$

The first few (non-negatively indexed) convergents are

$$\frac{p_0}{q_0} = \frac{3}{1}, \quad \frac{p_1}{q_1} = \frac{22}{7} \quad \text{and} \quad \frac{p_2}{q_2} = \frac{333}{106}.$$

Consider the intermediate fractions between the 1st and 2nd convergent. These are the fractions

$$\frac{p_1 + p_2}{q_1 + q_2} = \frac{25}{8}, \quad \frac{p_1 + 2p_2}{q_1 + 2q_2} = \frac{47}{15}, \quad \ldots, \quad \frac{p_1 + 14p_2}{p_2 + 14q_2} = \frac{311}{99}. \tag{4.6}$$

Straightforward calculation shows that

$$\left|\pi - \frac{22}{7}\right| = 0.00126\cdots < \left|\pi - \frac{25}{8}\right| = 0.0166\cdots,$$

so the intermediate fraction $25/8$ is not a best homogeneous approximation of π in the relative sense. However, all the intermediate fractions in the sequence $(p_1 + kp_2)/(q_1 + kq_2)$ are best approximations of this type when $7 < k < 15 = a_3$.

In this section, we have demonstrated the connection between Algorithm 3.2 and the simple continued fraction expansion of a real number. We have called upon results from that theory to improve the running time bound on the algorithm for calculation of best approximations over that which we found in the previous section (although only by an asymptotically constant factor). We have also presented a theorem (Theorem 4.6) which relates best homogeneous Diophantine approximations of a real number in the relative sense to the convergents and intermediate fractions of the s.c.f. expansion of that number.

5. Cassels' Algorithm

Our aim in this section is to describe an algorithm due to CASSELS (1954) and to examine in the relationship of its outputs with best inhomogeneous Diophantine approximations of the inputs.[1] We begin by presenting the algorithm in full in Algorithm 5.1. The algorithm initially appears quite complex and, from the algorithm alone, we would have some difficulty in deducing what relationship it has, if any, with best inhomogeneous approximations. This will be explained, little by little, in the subsequent series of propositions which culminates in Theorem 5.1. The reader is therefore advised to read the algorithm in a cursory manner at first, and to refer back to it as necessary to confirm the statements of the propositions.

[1] In CLARKSON *et al.* (1996), the author made the mistake of attributing a similar algorithm to DESCOMBES (1956).

ALGORITHM 5.1.

1 **begin**
2 $\eta_{-1} := -1;\ \eta_{-2} = \alpha;\ \zeta_{-1} := -\beta;$
3 $p_{-1} := 1;\ p_{-2} := 0;\ P_{-1} := 0;$
4 $q_{-1} := 0;\ q_{-2} := 1;\ Q_{-1} := 0;$
5 $n := 0;$
6 **while** $\eta_{n-1} \neq 0 \ \wedge\ \zeta_{n-1} \neq 0$ **do**
7 $a_n := \left\lfloor \dfrac{-\eta_{n-2}}{\eta_{n-1}} \right\rfloor;$
8 $p_n := p_{n-2} + a_n p_{n-1};\ q_n := q_{n-2} + a_n q_{n-1};$
9 $\eta_n := \eta_{n-2} + a_n \eta_{n-1};$
10 **if** $Q_{n-1} \leqslant q_{n-1}$ **then**
11 $b_n := \left\lfloor \dfrac{-\zeta_{n-1} - \eta_{n-2}}{\eta_{n-1}} \right\rfloor;$
12 $P_n := P_{n-1} + p_{n-2} + b_n p_{n-1};\ Q_n := Q_{n-1} + q_{n-2} + b_n q_{n-1};$
13 $\zeta_n := \zeta_{n-1} + \eta_{n-2} + b_n \eta_{n-1};$
14 **else**
15 $P_n := P_{n-1} - p_{n-1};\ Q_n := Q_{n-1} - q_{n-1};$
16 $\zeta_n := \zeta_{n-1} - \eta_{n-1};$
17 **fi**;
18 $n := n + 1;$
19 **od**;
20 **end**

We observe that this algorithm is "built on top of" Algorithm 3.2, in that the values of η_n, a_n, p_n and q_n are those which would be calculated by that algorithm for the same input α. Therefore we bring to bear the results of the previous two sections to prove the following proposition, which is also due to CASSELS, and subsequent results in this section.

PROPOSITION 5.1. *Suppose Algorithm 5.1 is executed on the real inputs α and β. For each $n \geqslant 0$,*

$$\zeta_n = Q_n\alpha - P_n - \beta. \tag{5.1}$$

If, additionally, $\eta_n \neq 0$ and $\zeta_n \neq 0$ then either

$$0 < Q_n \leqslant q_n \qquad \text{and} \qquad \zeta_n(\zeta_n + \eta_{n-1}) < 0 \tag{$\mathcal{A}_n$}$$

or

$$q_n < Q_n \leqslant q_n + q_{n-1} \qquad \text{and} \qquad \zeta_n(\zeta_n - \eta_n) < 0 \tag{$\mathcal{B}_n$}$$

PROOF. The truth of (5.1) is apparent from inspection of Algorithm 5.1. We see that $\zeta_{-1} = Q_{-1}\alpha - P_{-1} - \beta$. The expressions for updating P_n, Q_n and ζ_n everywhere

have the same form. Bearing in mind statement (i) of Proposition 3.1, we find that (5.1) must always be true.

The proof of the remainder of the proposition is by induction. For $n = 0$ we have $Q_{-1} = q_{-1} = 0$ and so Algorithm 5.1 produces $P_0 = b_0 = \lfloor \alpha - \beta \rfloor$ and $Q_0 = q_0 = 1$. Furthermore, $0 < \zeta_0 = \alpha - \beta - \lfloor \alpha - \beta \rfloor < 1 = -\eta_{-1}$ and so $\zeta_0(\zeta_0 + \eta_{-1}) < 0$. Thus, a pair (P_0, Q_0) has been calculated by Algorithm 5.1 which satisfies $(\mathcal{A}_n)$ for $n = 0$.

Suppose the statement of the theorem is true for all $0 \leqslant n < N$. If $(\mathcal{B}_n)$ holds for $n = N - 1$ then Algorithm 5.1 sets $Q_N = Q_{N-1} - q_{N-1}$ and $\zeta_N = \zeta_{N-1} - \eta_{N-1}$ and so $(\mathcal{A}_n)$ holds for $n = N$ unless $\eta_N = 0$ or $\zeta_N = 0$.

Suppose $(\mathcal{A}_n)$ holds for $n = N - 1$. Algorithm 5.1 will then set

$$Q_N = Q_{N-1} + q_{N-2} + b_N q_{N-1} \qquad \text{and} \qquad \zeta_N = \zeta_{N-1} + \eta_{N-2} + b_N \eta_{N-1}.$$

Now,

$$\frac{-\zeta_{N-1} - \eta_{N-2}}{\eta_{N-1}} = \frac{\zeta_{N-1}(\zeta_{N-1} + \eta_{N-2})}{-\zeta_{N-1}\eta_{N-1}}. \tag{5.2}$$

The numerator of the right-hand side of (5.2) is negative because of $(\mathcal{A}_n)$ for $n = N-1$, which also implies that $\zeta_{N-1}\eta_{N-2} < 0$. From statement (iv) of Proposition 3.1, we find that $\zeta_{N-1}\eta_{N-1} > 0$. Hence the expressions of (5.2) are positive. By the same reasoning,

$$\frac{-\zeta_{N-1} - \eta_{N-2}}{\eta_{N-1}} < \frac{-\eta_{N-2}}{\eta_{N-1}}.$$

Hence, $0 \leqslant b_N \leqslant a_N$.

If $b_N < a_N$ then $0 < Q_N \leqslant q_N$. Now, using the inequalities

$$\frac{-\zeta_{N-1} - \eta_{N-2}}{\eta_{N-1}} - 1 < b_N \leqslant \frac{-\zeta_{N-1} - \eta_{N-2}}{\eta_{N-1}} \tag{5.3}$$

together with the assumption that $\eta_{N-1} \neq 0$, we find that

$$-\eta_{N-1}^2 < \zeta_N \eta_{N-1} < 0 \qquad \text{and} \qquad 0 < (\zeta_N + \eta_{N-1})\eta_{N-1} < \eta_{N-1}^2 \tag{5.4}$$

when $\zeta_N \neq 0$ which implies that $\zeta_N(\zeta_N + \eta_{N-1}) < 0$. Thus, if $b_N < a_N$ then Algorithm 5.1 finds an integer couple (P_N, Q_N) which satisfies $(\mathcal{A}_n)$ for $n = N$.

If $b_N = a_N$ then $q_N < Q_N \leqslant q_N + q_{N-1}$. In this case,

$$b_N = a_N = \frac{\eta_N - \eta_{N-2}}{\eta_{N-1}} > \frac{\eta_N - \eta_{N-2} - \zeta_{N-1}}{\eta_{N-1}}. \tag{5.5}$$

We can use (5.5) to replace the left-hand inequality of (5.3) in order to improve (5.4) so that

$$\eta_{N-1}\eta_N < \zeta_N \eta_{N-1} < 0 \qquad \text{and} \qquad 0 < (\zeta_N - \eta_N)\eta_{N-1} < -\eta_{N-1}\eta_N$$

(provided neither $\zeta_N = 0$ nor $\eta_N = 0$) which implies that $\zeta_N(\zeta_N - \eta_N) < 0$. Therefore, $(\mathcal{B}_n)$ holds for $n = N$. □

REMARK 5.1. We note some implications which have come to light in the proof of Proposition 5.1. We now see that we can regard Algorithm 5.1 as being in one of three states when $n \geqslant 0$: the state in which (P_n, Q_n) and ζ_n satisfy $(\mathcal{A}_n)$, in which they satisfy $(\mathcal{B}_n)$ or in which the algorithm is about to terminate with $\zeta_n = 0$ or $\eta_n = 0$. We will refer to either of the first two states simply by the equation numbers $(\mathcal{A}_n)$ or $(\mathcal{B}_n)$.

Furthermore, for every iteration $n \geqslant 0$ we observe and emphasise that

(*i*) regardless of the algorithm's state, we will always have $0 < Q_n \leqslant q_n + q_{n-1}$,
(*ii*) unless the algorithm is terminating with $\zeta_n = 0$, we have $\zeta_n(\zeta_n + \eta_{n-1}) < 0$,
(*iii*) the algorithm is in state $(\mathcal{A}_n)$ or the terminating state when $n = 0$,
(*iv*) if the algorithm is in state $(\mathcal{A}_n)$ then $0 \leqslant b_{n+1} \leqslant a_{n+1}$ and if $k \in \mathbb{Z}$, $k \leqslant b_{n+1}$ then $\eta_n(\zeta_n + \eta_{n-1} + k\eta_n) \leqslant 0$,
(*v*) the algorithm is in state $(\mathcal{B}_n)$ if and only if it was in state $(\mathcal{A}_n)$ on the previous iteration and $b_n = a_n$ and
(*vi*) if it is in state $(\mathcal{B}_n)$ then $n > 0$, $(P_{n+1}, Q_{n+1}) = (P_{n-1}, Q_{n-1})$ and the algorithm will either be in state $(\mathcal{A}_n)$ or in the terminating state with $\eta_{n+1} = 0$ on the subsequent iteration.

These remarks lead to the following proposition and corollary.

PROPOSITION 5.2. *If Algorithm 5.1 is executed on the real inputs α and β then, for each $n \geqslant 0$ for which $\zeta_n \neq 0$, it is true that*

(*i*) $\zeta_n > 0$ *if n is even,*
(*ii*) $\zeta_n < 0$ *if n is odd and*
(*iii*) $|\zeta_n| < |\eta_{n-1}|$.

PROOF. Statements (*i*) and (*ii*) follow from the fact that $\zeta_n \eta_{n-1} < 0$, which is a consequence of statement (*ii*) of Remark 5.1, and because $\zeta_0 = \alpha - \beta - \lfloor \alpha - \beta \rfloor \geqslant 0$.

Statement (*iii*) is also a consequence of statement (*ii*) of Remark 5.1. □

COROLLARY 5.1. *If Algorithm 5.1 is run on the real inputs α and β then the algorithm either terminates after a finite number of iterations with $\eta_n = 0$ or $\zeta_n = 0$ or*

$$\lim_{n\to\infty} \zeta_n = \lim_{n\to\infty} \eta_n = 0.$$

From Proposition 5.2, we can already see that Algorithm 5.1 produces good inhomogeneous Diophantine approximations very quickly because $|\zeta_n| \leqslant |\eta_{n-1}|$. Indeed, we can use the results of Theorem 4.1 and subsequent discussion from the previous section to deduce that the number of iterations of Algorithm 5.1 required to produce a pair of integers with an absolute approximation error less than some $\epsilon > 0$ is at most $\lceil \log_\gamma(\epsilon^{-1} + 1) \rceil + 3$ (the extra iteration being required because $|\zeta_n| \leqslant |\eta_{n-1}|$ rather than $|\zeta_n| \leqslant |\eta_n|$).

However, we can show that the integer pairs produced by Algorithm 5.1 yield all the best inhomogeneous Diophantine approximations of α with respect to β. We show this in the following propositions.

PROPOSITION 5.3. *Suppose Algorithm 5.1 is executed with the real numbers α and β as its inputs. Suppose (P, Q) are integers and let $\zeta = Q\alpha - P - \beta$. If, for some $n \geqslant 0$, the algorithm is in state $(\mathcal{A}_n)$ and*

$$(\zeta - \zeta_n)(\zeta - \zeta_n - \eta_{n-1} - k\eta_n) < 0 \tag{5.6}$$

and $0 \leqslant k \leqslant b_{n+1}$ then either $Q \leqslant 0$ or $Q \geqslant Q_n + q_{n-1} + (k+1)q_n$. Furthermore, the same is also true if the algorithm is in the terminating state with $\zeta_n \neq 0$ and $k = 0$.

PROOF. The proof is by induction on n and k. The proposition is true for $n = 0$ and $k = 0$ because, in this case, satisfaction of (5.6) is equivalent to satisfaction of

$$\alpha - \beta - \lfloor \alpha - \beta \rfloor - 1 < \zeta < \alpha - \beta - \lfloor \alpha - \beta \rfloor. \tag{5.7}$$

If $Q = 1$ then $\zeta = \alpha - \beta - P$ and (5.7) cannot be satisfied. Therefore $Q \leqslant 0$ or $Q \geqslant Q_0 + q_{-1} + q_0 = 2$.

Suppose the proposition is true for all $0 \leqslant k < K$ when $n = N \geqslant 0$ and $0 < K \leqslant b_{N+1}$ and for all $n < N$ if $N > 0$. Suppose (P, Q) is an integer pair which satisfies the conditions of the proposition for $n = N$ and $k = K$. (Note that this condition excludes the possibility that the algorithm is in the terminating state.) We will call upon Fact 3.1 with $x = \zeta$, $a = \zeta_N$, $b = \zeta_N + \eta_{N-1} + K\eta_N$ and $c = -\eta_N$ to show that (5.6) implies that

$$(\zeta - \zeta_N)[\zeta - \zeta_N - \eta_{N-1} - (K-1)\eta_N] < 0. \tag{5.8}$$

That Fact 3.1 can be applied is not clear until we show that $ac \leqslant 0$ and $bc \geqslant 0$. To see that $bc \geqslant 0$, recall statement (iv) of Remark 5.1. Now, $\zeta_N \eta_{N-1} < 0$ because of $(\mathcal{A}_n)$ and therefore

$$\zeta_N \eta_{N-1} \eta_N (\zeta_N + \eta_{N-1} + K\eta_N) \geqslant 0.$$

Recalling statement (iv) of Proposition 3.1 we now see that $ac \leqslant 0$.

Having established (5.8), we can now apply the proposition to (P, Q) for $n = N$ and $k = K - 1$ to find that $Q \leqslant 0$ or $Q \geqslant Q_N + q_{N-1} + Kq_N$. Let

$$(p, q) = (P - P_N - p_{N-1} - Kp_N, Q - Q_N - q_{N-1} - Kq_N)$$

and let $\eta = q\alpha - p$. Clearly, p and q are not both zero. We have

$$\eta = \zeta - \zeta_N - \eta_{N-1} - K\eta_N$$

which implies that $(\mathcal{A}_n)$ can be rewritten as

$$\eta(\eta + \eta_{N-1} + K\eta_N) < 0.$$

From statement (iii) of Lemma 3.1 we find that $q \neq 0$ and (3.10) is satisfied for $n = N - 1$ and $k = a_N - 1$. If $N = 0$ this implies that $Q \leqslant 0$ or $Q \geqslant Q_0 + q_{-1} + (K+1)q_0 = K + 2$. If $N > 0$ then we can apply Proposition 3.3 to show that $q < 0$ or $q \geqslant q_N$ which implies that $Q \leqslant 0$ or $Q \geqslant Q_N + q_{N-1} + (K+1)q_N$.

To complete the induction on n and k, it remains to show that if the proposition holds for all $0 \leqslant n < N$ then it holds for $n = N$ and $k = 0$. (Note that our discussion now includes the possibility that the algorithm is in the terminating state with $\zeta_N \neq 0$.) Suppose (P, Q) is an integer pair which satisfies the conditions of the proposition for $n = N$ and $k = 0$.

Suppose the algorithm was in state $(\mathcal{B}_n)$ on the previous iteration, $n = N - 1$. Then, from statement (vi) of Remark 5.1, we see that $N \geqslant 2$. We also note that (5.6) for $n = N$, $k = 0$ reduces to

$$\begin{aligned}(\zeta - \zeta_N)(\zeta - \zeta_N - \eta_{N-1}) &= (\zeta - \zeta_{N-2})(\zeta - \zeta_{N-1}) \\ &= (\zeta - \zeta_{N-2})(\zeta - \zeta_{N-2} - \eta_{N-3} - b_{N-1}\eta_{N-2}) < 0.\end{aligned}$$

But this is simply (5.6) for $n = N - 2$, $k = b_{N-1}$. We conclude that $Q \leqslant 0$ or $Q \geqslant Q_{N-1} + q_{N-2} = Q_N + q_{N-2} + q_{N-1} > Q_N + q_{N-1}$.

On the other hand, suppose the algorithm was in state $(\mathcal{A}_n)$ on the previous iteration, $n = N - 1$. We intend to apply Fact 3.1 to show that, in this case, (5.6) implies that

$$(\zeta - \zeta_N)(\zeta - \zeta_{N-1}) < 0. \tag{5.9}$$

To do this, we set $x = \zeta$, $a = \zeta_N$, $b = \zeta_N + \eta_{N-1} = \zeta_{N-1} + \eta_{N-2} + (b_N + 1)\eta_{N-1}$ and $c = -[\eta_{N-2} + (b_N + 1)\eta_{N-1}]$. Recalling from statement (v) of Remark 5.1 that $b_N < a_N$, we see that

$$\eta_{N-1}[\eta_{N-2} + (b_N + 1)\eta_{N-1}] \leqslant 0$$

because of statement (ii) of Proposition 3.1. Also, statement (ii) of Remark 5.1 implies that $\zeta_N \eta_{N-1} < 0$. Therefore,

$$\eta_{N-1}^2 \zeta_N [\eta_{N-2} + (b_N + 1)\eta_{N-1}] \geqslant 0 \tag{5.10}$$

and, because $\eta_{N-1}^2 > 0$, we find that $ac \leqslant 0$. Again using statement (ii) of Remark 5.1, we find from (5.10) that

$$\eta_{N-1}^2 \zeta_N^2 (\zeta_N + \eta_{N-1})[\eta_{N-2} + (b_N + 1)\eta_{N-1}] \leqslant 0.$$

As $\eta_{N-1}^2 \zeta_N^2 > 0$, we see that $bc \geqslant 0$. Therefore, we have shown the validity of (5.9) in this case. We observe that (5.9) is just (5.6) when $n = N - 1$ and $k = b_N$. Therefore, we can apply the proposition to show that $Q \leqslant 0$ or $Q \geqslant Q_{N-1} + q_{N-2} + (b_N + 1)q_{N-1} = Q_N + q_{N-1}$.

Thus, regardless of whether the algorithm was in state $(\mathcal{A}_n)$ or $(\mathcal{B}_n)$ on the previous iteration $n = N - 1$, we have found that either $Q \leqslant 0$ or $Q \geqslant Q_N + q_{N-1}$.

Let

$$(p, q) = (P - P_N - p_{N-1}, Q - Q_N - q_{N-1}).$$

Clearly, p and q are not both zero. Let

$$\eta = q\alpha - p = \zeta - \zeta_N - \eta_{N-1}.$$

Substitution into (5.6) reveals that

$$\eta(\eta + \eta_{N-1}) < 0. \tag{5.11}$$

Witness that this is (3.11) of statement (iii) of Lemma 3.1 with $\kappa = 0$ which implies that Proposition 3.10 can be applied. We then find that either $q < 0$ or $q \geqslant q_N$. Hence, either $Q \leqslant 0$ or $Q \geqslant Q_{N-1} + q_N = Q_N + q_{N-1} + q_N$, in agreement with the proposition. Therefore, the induction on n and k is complete and the proposition is proved. □

PROPOSITION 5.4. *Suppose Algorithm 5.1 is executed with the real inputs α and β. Suppose (P, Q) is a pair of integers with $Q > 0$ and let $\zeta = Q\alpha - P - \beta$. If, for some $n \geqslant 0$, the algorithm is in state $(\mathcal{A}_n)$ and* (5.6) *of Proposition 5.3 is satisfied for some $0 \leqslant k \leqslant b_{n+1}$ and $k < a_{n+1}$ and*

$$\eta_{n-1} + (k+1)\eta_n \neq 0 \tag{5.12}$$

then either $(P, Q) = (P^*, Q^*)$, where (P^*, Q^*) is defined by

$$(P^*, Q^*) = (P_n + p_{n-1} + (k+1)p_n, Q_n + q_{n-1} + (k+1)q_n),$$

or $Q > Q^*$.

PROOF. Firstly, we show that (P^*, Q^*) satisfies (5.6). With $\zeta^* = Q^*\alpha - P^* - \beta = \zeta_n + \eta_{n-1} + (k+1)\eta_n$, we find that

$$(\zeta^* - \zeta_n)(\zeta^* - \zeta_n - \eta_{n-1} - k\eta_n) = \eta_n[\eta_{n-1} + (k+1)\eta_n] < 0$$

by virtue of (5.12) and statement (ii) of Proposition 3.1 because $k + 1 \leqslant a_{n+1}$.

Now, from Proposition 5.3, we know that $Q \geqslant Q^*$. Let

$$(p, q) = (P - P^*, Q - Q^*)$$

and suppose that p and q are not both zero. Let

$$\eta = q\alpha - p = \zeta - \zeta^* = \zeta - \zeta_n - \eta_{n-1} - (k+1)\eta_n.$$

Substitution into (5.6) yields

$$[\eta + \eta_{n-1} + (k+1)\eta_n](\eta + \eta_n) < 0. \tag{5.13}$$

Successive application of Fact 3.1 shows that (5.13) implies that $(\eta + \eta_{n-1})(\eta + \eta_n) < 0$ which implies that $(\eta + \eta_{n-1})(\eta - \eta_{n-1}) < 0$. Thus, $|\eta| < |\eta_{n-1}| \leqslant 1$ which implies that $q \neq 0$. Hence, if $(P, Q) \neq (P^*, Q^*)$ then $Q > Q^*$. □

We now see that the inhomogeneous approximation errors of the (P_n, Q_n) generated by Algorithm 5.1 quickly vanish and, in the sense implied by Proposition 5.4, they enjoy a uniqueness property. For this reason, we call the (P_n, Q_n) for $n \geqslant 0$ the AUXILIARY CONVERGENTS of α with respect to β. Similarly, the b_n, where defined, are AUXILIARY PARTIAL QUOTIENTS. We also define the INTERMEDIATE AUXILIARY CONVERGENTS between the n^{th} and $(n+1)^{\text{th}}$ auxiliary convergents as those pairs of integers of the form

$$(P_n + p_{n-1} + kp_n, Q_n + q_{n-1} + kq_n) \tag{5.14}$$

with $0 \leqslant k < b_{n+1}$ and $n \geqslant 0$. Obviously, this is only meaningful if (P_n, Q_n) satisfies $(\mathcal{A}_n)$. However, we also define an intermediate auxiliary convergent which follows the final auxiliary convergent if, on the terminating step, $\zeta_n \neq 0$. This final intermediate auxiliary convergent has the form of (5.14) with $k = 0$.

We can now state and prove the main result of this section.

THEOREM 5.1. *If (P, Q) is a best inhomogeneous Diophantine approximation of α with respect to β then it is an auxiliary convergent or intermediate auxiliary convergent of α with respect to β.*

PROOF. Let $\zeta = Q\alpha - P - \beta$. Suppose there are only a finite number of auxiliary convergents and intermediate auxiliary convergents (which will occur if Algorithm 5.1 terminates) and each has an absolute approximation error greater than $|\zeta|$. This means that $|\zeta| < |\zeta_N|$ and $|\zeta| < |\zeta_n + \eta_{N-1}|$ and that $\zeta_N \neq 0$ and $\eta_N = 0$, where N is the index of the final auxiliary convergent (the iteration on which the algorithm enters the terminating state). Hence, starting with either

$$(\zeta - \zeta_N)(\zeta + \zeta_N) < 0 \qquad \text{or} \qquad (\zeta - \zeta_N - \eta_{N-1})(\zeta + \zeta_N + \eta_{N-1}) < 0$$

depending on whether $|\zeta_N| < |\zeta_N + \eta_{N-1}|$ or not, respectively, we can apply Fact 3.1 to show that

$$(\zeta - \zeta_N)(\zeta - \zeta_N - \eta_{N-1}) < 0$$

because of statement (ii) of Remark 5.1. We can apply Proposition 5.3 to show that $Q \geqslant Q_N + q_{N-1} + q_N$. But the integer pair $(P - p_N, Q - q_N)$ has the same approximation error as (P, Q) since $\eta_N = 0$ and $0 < Q - q_N < Q$. Thus, (P, Q) cannot possibly be a best inhomogeneous Diophantine approximation and the theorem is true in this case.

On the other hand, suppose there exists some auxiliary convergent or intermediate auxiliary convergent which has an absolute approximation error not greater than $|\zeta|$. This is automatic if there are an infinite number of auxiliary convergents because of Corollary 5.1. Now, arrange the auxiliary convergents and intermediate auxiliary convergents, starting with (P_0, Q_0), in order of increasing index, in the case of auxiliary convergents, and with the intermediate auxiliary convergents ordered in the obvious way between the auxiliary convergents with which they are associated.

Choose the first integer pair from this sequence with an absolute approximation error not greater than $|\zeta|$. Let (P^*, Q^*) be this element. We will show that we can express (P^*, Q^*) as either

$$(P^*, Q^*) = (P_0, Q_0) \tag{5.15}$$

or

$$(P^*, Q^*) = (P_N + p_{N-1} + Kp_N, Q_N + q_{N-1} + Kq_N) \tag{5.16}$$

for some $N \geqslant 0$ and $0 \leqslant K \leqslant b_{N+1}$. This is obvious if (P^*, Q^*) is an intermediate auxiliary convergent but it requires some explanation otherwise. Suppose (P^*, Q^*) is an auxiliary convergent, say (P_m, Q_m) for some $m \geqslant 0$. If $m = 0$ then obviously we have (5.15). If $m > 0$ and (P_{m-1}, Q_{m-1}) satisfies $(\mathcal{A}_n)$ for $n = m - 1$ then (5.16) is obviously valid with $N = m - 1$ and $K = b_m$. If (P_{m-1}, Q_{m-1}) satisfies $(\mathcal{B}_n)$ then, from statement (vi) of Remark 5.1, $m \geqslant 2$ and $(P_{m-2}, Q_{m-2}) = (P_m, Q_m)$ and so there is an earlier element in the sequence with the same approximation error, contrary to our assumption. Furthermore, we note that if (P^*, Q^*) takes the form of (5.16) with $K = 0$ then (P_{N-1}, Q_{N-1}) cannot have satisfied $(\mathcal{B}_n)$ for, if it had, we would have $(P_N + p_{N-1}, Q_N + q_{N-1}) = (P_{N-1}, Q_{N-1})$ and (P_{N-1}, Q_{N-1}) occurs earlier in the sequence.

Now that we have established the validity of (5.15) and (5.16), let us set $\zeta^* = Q^*\alpha - P^* - \beta$. By assumption, $|\zeta^*| \leqslant |\zeta|$.

Suppose $(P^*, Q^*) = (P_0, Q_0) = (\lfloor \alpha - \beta \rfloor, 1)$. Now, we must have $\zeta_0 = \frac{1}{2} = -\zeta$ and $(P, Q) = (\lceil \alpha - \beta \rceil, 1)$, otherwise (P, Q) could not be a best approximation. But $(P, Q) = (P_0 + p_{-1}, Q_0 + q_{-1})$ in this case, which is an intermediate auxiliary convergent, and so the theorem holds.

If we have (5.16) with $K > 0$ then we know that $|\zeta| < |\zeta_N|$ and also that $|\zeta| < |\zeta_N + \eta_{N-1} + (K-1)\eta_N|$ because both integer pairs to which these approximation errors correspond occur earlier in the sequence than (P^*, Q^*) and therefore

$$(\zeta - \zeta_N)[\zeta - \zeta_N - \eta_{N-1} - (K-1)\eta_N] < 0.$$

We can then apply Proposition 5.4 with $n = N$ and $k = K - 1$ to show that $Q > Q^*$ unless $(P, Q) = (P^*, Q^*)$. If $Q > Q^*$ then (P, Q) cannot possibly be a best approximation because $|\zeta| \geqslant |\zeta^*|$. Therefore, the theorem is true in this case.

The last possibility we need to consider is that (P^*, Q^*) takes the form of (5.16) with $K = 0$. If $N = 0$ then we have $|\zeta| < |\zeta_0|$. This implies that $(P, Q) = (P^*, Q^*) = (P_0 - 1, 1)$ or that $Q > 1 = Q^*$ in which case (P, Q) cannot possibly be a best approximation since $|\zeta| \geqslant |\zeta^*|$. If $N > 0$ then, as we discussed above, we know that (P_{N-1}, Q_{N-1}) satisfies $(\mathcal{A}_n)$ for $n = N - 1$. Now, $|\zeta| < |\zeta_N|$ and $|\zeta| < |\zeta_{N-1}|$ because both integer pairs to which these approximation errors correspond occur

earlier in the sequence than (P^*, Q^*). Therefore,

$$(\zeta - \zeta_{N-1})(\zeta - \zeta_N) = (\zeta - \zeta_{N-1})(\zeta - \zeta_{N-1} - \eta_{N-2} - b_N\eta_{N-1}) < 0.$$

We can then apply Proposition 5.4 with $n = N - 1$ and $k = b_N$ to show that either $(P, Q) = (P^*, Q^*)$ or $Q > Q^*$. As before, we see that if $Q > Q^*$ then (P, Q) cannot possibly be a best approximation. Therefore, the proof is complete. □

6. Successive Maxima of Certain Diagonal Functions

In this section we will examine the relationship between certain diagonal functions and Diophantine approximation. A diagonal function is a type of almost periodic function (Levitan & Zhikov, 1982) which arises from taking the value of a periodic function in n variables, $n > 1$, along a "diagonal" or line passing through the origin (see Definition 6.1). For certain diagonal functions arising from periodic functions of two variables, we will show in Theorem 6.1 that there is a close relationship between the successive maxima of the function and the best homogeneous Diophantine approximations in the absolute sense of the slope of the diagonal. We will then illustrate the theorem by applying it to periodograms of the form

$$F(\omega) = \left| \sum_{j=1}^{3} A_j e^{-i\omega t_j} \right|^2$$

where the A_j are positive real numbers and the t_j are real numbers with $t_1 \leqslant t_2 \leqslant t_3$. We will see that the positions of the successive peaks in a periodogram of this type are dictated by the best homogeneous Diophantine approximations of the ratio $(t_2 - t_1)/(t_3 - t_1)$. This result is potentially significant for signal processing, where the successive maxima of the periodogram are important for frequency estimation and spectral estimation.

Definition 6.1. A function $F : \mathbb{R} \to \mathbb{R}$ is called a diagonal function if f can be expressed $F(x) = f(\alpha_1 x, \alpha_2 x, \ldots, \alpha_n x)$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a function which is periodic in each of its variables and $\alpha_i \in \mathbb{R}$ for each $i = 1, \ldots, n$.

Definition 6.2. A function $f : \mathbb{R}^2 \to \mathbb{R}$ is $\mathbb{Z}^2$-periodic if $f(\mathbf{x} + \mathbf{k}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{k} \in \mathbb{Z}^2$.

We use $\mathrm{Nint}(x)$ to denote the set of nearest integers to a real number x. Obviously, this will contain only one element unless x is a half-integer, in which case it will contain two elements.

Theorem 6.1. *Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is a symmetric, $\mathbb{Z}^2$-periodic function and f is bounded above. Suppose that the restriction of f to the unit square $\mathcal{T} = \left[-\frac{1}{2}, \frac{1}{2}\right] \times \left[-\frac{1}{2}, \frac{1}{2}\right]$ attains its maximum at $\mathbf{0}$ and nowhere else and nor are there any other local*

maxima. If x and α are real numbers such that $x \geqslant \frac{1}{2}$, $|\alpha| \leqslant 1$ and for all $y \in \mathbb{R}$, $y \geqslant \frac{1}{2}$,

$$y \leqslant x \;\Rightarrow\; f(y\alpha, y) \leqslant f(x\alpha, x) \tag{6.1}$$

and

$$f(y\alpha, y) \geqslant f(x\alpha, x) \;\Rightarrow\; y \geqslant x \tag{6.2}$$

then the Cartesian product $\mathcal{N} = \mathrm{Nint}(x\alpha) \times \mathrm{Nint}(x)$ contains either a best homogeneous Diophantine approximation of α in the absolute sense or the point $(0, 1)$.

Proof. If $x = \frac{1}{2}$ then the proof is obvious because $(0, 1) \in \mathcal{N}$. If $|\alpha| = 1$ then $(\alpha, 1)$ is the only best approximation of α. However, there must be an element of the form $(k\alpha, k) \in \mathcal{N}$ with $k \in \mathbb{N}$. If $k > 1$ then $f(\alpha, 1) \geqslant f(x\alpha, x)$ and $1 < x$, which contradicts (6.2). Therefore, for the rest of the proof, we suppose that $x > \frac{1}{2}$ and $|\alpha| < 1$.

Suppose that neither a best approximation nor $(0, 1)$ are elements of $\mathcal{N}$. Let (p, q) be that element of $\mathcal{N}$ with least absolute value for each coordinate. Clearly, $q > 0$ since $x > \frac{1}{2}$. Let $\eta = q\alpha - p$. Now, there must exist some best approximation (p^*, q^*) of α such that

$$q^* \leqslant q \qquad \text{and} \qquad |\eta^*| \leqslant |\eta|$$

and at least one of these inequalities must be satisfied strictly. If there is more than one best approximation that satisfies these conditions then choose that best approximation which gives the smallest $q^* > 0$.

Suppose $\eta^* = 0$. In this case, $(q^*\alpha, q^*) \in \mathbb{Z}^2$ and so $f(q\alpha, q) \geqslant f(x\alpha, x)$. If $q^* < q$ then $q^* < x$ which would contradict (6.2). If $q^* = q$ then

$$q^* - \tfrac{1}{2} < x \leqslant q^* + \tfrac{1}{2}.$$

Therefore,

$$p^* - \tfrac{1}{2}|\alpha| < x\alpha \leqslant p^* + \tfrac{1}{2}|\alpha|$$

but this implies that $(p^*, q^*) \in \mathcal{N}$ because $|\alpha| < 1$, contrary to our assumption.

Suppose $0 < |\eta^*| = |\eta|$. In this case, $0 < q^* < q$, otherwise (p, q) would also be a best approximation. Now $f(x\alpha, x) = f(x\alpha - p, x - q)$ because f is $\mathbb{Z}^2$-periodic. Also, $-\frac{1}{2} < x - q \leqslant \frac{1}{2}$. If $\eta^* = \eta$ then, using the fact that

$$x\alpha - p = (q^* + x - q)\alpha - p^*,$$

we see that

$$f(x\alpha, x) = f((q^* + x - q)\alpha, q^* + x - q).$$

But $\frac{1}{2} < q^* + x - q \leqslant q^* + \frac{1}{2} < x$ and this contradicts (6.2). Similarly, if $\eta^* = -\eta$ then

$$f(x\alpha, x) = f((q^* + q - x)\alpha, q^* + q - x),$$

again contradicting (6.2) since $\frac{1}{2} \leqslant q^* + q - x < x$.

Finally, suppose $0 < |\eta^*| < |\eta|$. Let $\mathbf{x} = (x\alpha - p, x - q) \in \mathcal{T}$. Consider the sets

$$\mathcal{P}_1 = \{(a, b) \in \mathcal{T} \mid a\alpha - b < |\eta^*|\}$$
$$\mathcal{P}_2 = \{(a, b) \in \mathcal{T} \mid a\alpha - b > |\eta^*|\}$$

and

$$\mathcal{P}_3 = \{(a, b) \in \mathcal{T} \mid a\alpha - b = |\eta^*|\}$$

Clearly, $\mathcal{P}_1 \cup \mathcal{P}_2 \cup \mathcal{P}_3 = \mathcal{T}$ and $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ whenever $i \neq j$. Also, $\mathbf{0} \in \mathcal{P}_1$ and either $\mathbf{x} \in \mathcal{P}_2$ or $-\mathbf{x} \in \mathcal{P}_2$ and both $\mathcal{P}_1$ and $\mathcal{P}_2$ are open in $\mathcal{T}$. Consider the set

$$\mathcal{C} = \{\mathbf{z} \in \mathcal{T} \mid f(\mathbf{z}) \geqslant f(\mathbf{x})\}.$$

We know that $\mathcal{C} \cap \mathcal{P}_i \neq \emptyset$ when $i = 1$ or $i = 2$. Suppose $\mathcal{C} \cap \mathcal{P}_3 = \emptyset$. This would imply that $\mathcal{C}$ is disconnected and in turn this would imply that the restriction of f to $\mathcal{T}$ has a local maximum on $\mathcal{P}_2$. However, we have assumed in the theorem statement that f has no such local maximum in $\mathcal{P}_2$. Therefore, there is some point $\mathbf{z} \in \mathcal{P}_3$ such that $f(\mathbf{z}) \geqslant f(\mathbf{x})$. Thus, there exists some $\delta \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ such that

$$f(\eta^* + \delta\alpha, \delta) \geqslant f(x\alpha, x)$$

and, therefore,

$$f((q^* + \delta)\alpha, q^* + \delta) \geqslant f(x\alpha, x).$$

If $q^* + \delta < x$ then we contradict (6.2).

Suppose $q^* + \delta \geqslant x$ which implies that $q^* = q$. Now, $x > q - \frac{1}{2}$. Thus, both x and $q^* + \delta$ lie in the interval $(q^* - \frac{1}{2}, q^* + \frac{1}{2}]$. Consider any real z in this interval. Because $|\alpha| < 1$, there are only two possible integers which can be elements of $\mathrm{Nint}(z\alpha)$ and they differ by one. Clearly, $p_0 = \lfloor q^*\alpha \rceil$ is one such value and $|q^*\alpha - p_0| \leqslant \frac{1}{2}$. Let p_1 be the other possible element of $\mathrm{Nint}(z\alpha)$. Thus, $|p_1 - p_0| = 1$ and $|q^*\alpha - p_1| \geqslant \frac{1}{2}$. We see that both p and p^* are distinct elements of the set $\{p_0, p_1\}$. Now, $|\eta^*| \leqslant \frac{1}{2}$ because (p^*, q^*) is a best approximation. If $|\eta^*| = \frac{1}{2}$ then $|\eta| = \frac{1}{2}$, but we have assumed that $|\eta^*| < |\eta|$, so $|\eta^*| < \frac{1}{2}$, $|\eta| > \frac{1}{2}$, $p^* = p_0$ and $p = p_1$. However, this implies that $q = q^* = 1$, since we chose that best approximation with smallest $q^* > 0$ and all best approximations have approximation errors less than (or equal to) $\frac{1}{2}$. In this case, one of the values p_0 or p_1 must be zero. Also, it is easily verified that $|p^*| \leqslant 1$. Now, because $x \leqslant q^* + \delta$, we must have $|x\alpha| \leqslant |(q^* + \delta)\alpha|$ which implies that $|p| \leqslant |p^*|$. Thus, either $(p, q) = (0, 1)$ or $(p, q) = (p^*, q^*)$. □

We illustrate Theorem 6.1 with two examples.

EXAMPLE 6.1. Let

$$f(\mathbf{x}) = -\min_{\mathbf{k}\in\mathbb{Z}^2} g(\mathbf{x}-\mathbf{k})$$

where

$$g(\mathbf{z}) = \left(|z_1|^p + |z_2|^p\right)^{1/p}$$

for some real constant $p \geqslant 1$ (g is p-norm; see Definition 5.2 in the next chapter). It is easily verified that $\|\mathbf{x}-\mathbf{k}\|_p$ attains its minimum for any $\mathbf{x}$ whenever $\mathbf{k} = (\lfloor x_1 \rceil, \lfloor x_2 \rceil)$. Furthermore, it is clear that $f(\mathbf{x})$ is $\mathbb{Z}^2$-periodic and bounded above and the restriction of $f(\mathbf{x})$ to the unit square $\mathcal{T}$ has only one local maximum at $\mathbf{0}$. Thus, Theorem 6.1 can be applied to functions of this type.

EXAMPLE 6.2. Consider functions of the form

$$\begin{aligned} f(x_1,x_2) &= \left|A_1 + A_2 e^{-i2\pi x_1} + A_3 e^{-i2\pi x_2}\right|^2 \\ &= A_1^2 + A_2^2 + A_3^2 \\ &\quad + 2A_1A_2\cos 2\pi x_1 + 2A_1A_3\cos 2\pi x_2 + 2A_2A_3\cos 2\pi(x_1-x_2) \end{aligned} \tag{6.3}$$

where the A_i, $i = 1,2,3$, are positive real numbers. A plot of $f(x_1,x_2)$ over $\mathcal{T}$

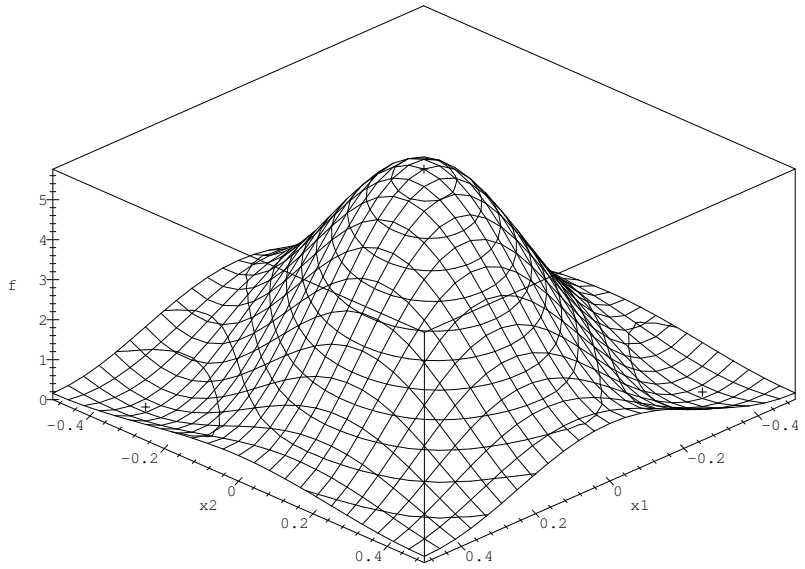


FIGURE 3. A plot of $f(x_1,x_2)$, as defined in (6.3), over the unit square $[-\frac{1}{2},\frac{1}{2}] \times [-\frac{1}{2},\frac{1}{2}]$.

appears in Figure 3. The values $A_1 = 1$, $A_2 = 0.8$ and $A_3 = 0.6$ were used to generate the plot. The local maxima and minima are marked with a plus sign (+). We now seek expressions for the positions of the local maxima and minima in terms of the values of A_1, A_2 and A_3.

The function is obviously $\mathbb{Z}^2$-periodic. The partial derivatives of the first and second orders are

$$\frac{\partial f}{\partial x_1} = -4\pi A_2[A_1 \sin 2\pi x_1 + A_3 \sin 2\pi(x_1 - x_2)], \tag{6.4}$$

$$\frac{\partial f}{\partial x_2} = -4\pi A_3[A_1 \sin 2\pi x_2 - A_2 \sin 2\pi(x_1 - x_2)], \tag{6.5}$$

$$\frac{\partial^2 f}{\partial x_1^2} = -8\pi^2 A_2[A_1 \cos 2\pi x_1 + A_3 \cos 2\pi(x_1 - x_2)], \tag{6.6}$$

$$\frac{\partial^2 f}{\partial x_2^2} = -8\pi^2 A_3[A_1 \cos 2\pi x_2 + A_2 \cos 2\pi(x_1 - x_2)] \tag{6.7}$$

and

$$\frac{\partial^2 f}{\partial x_1 x_2} = \frac{\partial^2 f}{\partial x_2 x_1} = 8\pi^2 A_2 A_3 \cos 2\pi(x_1 - x_2). \tag{6.8}$$

Consider the local maxima of f on $\mathbb{R}^2$. Recall that, because f has continuous second order partial derivatives, the local maxima of f occur wherever

$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial x_2} = 0, \tag{6.9}$$

$$\frac{\partial^2 f}{\partial x_1^2} \leqslant 0, \frac{\partial^2 f}{\partial x_2^2} \leqslant 0, \tag{6.10}$$

and

$$\frac{\partial^2 f}{\partial x_1^2}\frac{\partial^2 f}{\partial x_2^2} \geqslant \left(\frac{\partial^2 f}{\partial x_1 x_2}\right)^2. \tag{6.11}$$

We firstly consider trivial solutions of (6.9). Namely, these are the solutions of the form $(x_1, x_2) = (m/2, n/2)$ where $m, n \in \mathbb{Z}$. If both m and n are even then (6.10) and (6.10) and (6.11) are satisfied strictly. Thus, f has absolute local maxima at all points in $\mathbb{Z}^2$. It is also readily apparent that these local maxima are also the global maxima. If both m and n are odd and (6.10) is satisfied then

$$-\frac{\partial^2 f}{\partial x_1 x_2} < \frac{\partial^2 f}{\partial x_i^2} \leqslant 0$$

for $i = 1, 2$, so (6.11) cannot be satisfied. If one of m or n is even and the other odd then one of the inequalities of (6.10) cannot be satisfied. Thus, of the trivial solutions of (6.9), only solutions consisting of integer pairs yield local maxima, and these are absolute and global.

We now seek non-trivial solutions of (6.9). We can rewrite (6.4) as

$$\frac{\partial f}{\partial x_1} = -4\pi M A_2 \sin 2\pi(x_1 - \theta) \tag{6.12}$$

where

$$M = \left[(A_1 + A_3 \cos 2\pi x_2)^2 + A_3^2 \sin^2 2\pi x_2\right]^{1/2} = \left[A_1^2 + A_3^2 + 2A_1 A_3 \cos 2\pi x_2\right]^{1/2}$$

and (when $M \neq 0$)

$$\cos 2\pi\theta = \frac{A_1 + A_3 \cos 2\pi x_2}{M} \qquad \text{and} \qquad \sin 2\pi\theta = \frac{A_3 \sin 2\pi x_2}{M}. \tag{6.13}$$

Obviously, $M \geqslant 0$. Suppose $M = 0$. This implies that $x_2 = n/2$ where $n \in \mathbb{Z}$ and $A_1 = A_3$. Clearly, $\partial f/\partial x_1 = 0$. However, consideration of (6.5) shows that (6.9) is only satisfied for the trivial solutions when $M = 0$. Therefore, suppose $M > 0$.

From (6.12), we see that $\partial f/\partial x_1 = 0$ whenever

$$x_1 = \theta + k/2, \tag{6.14}$$

where $k \in \mathbb{Z}$. Suppose k is even, in which case

$$\cos 2\pi x_1 = \cos 2\pi\theta \qquad \text{and} \qquad \sin 2\pi x_1 = \sin 2\pi\theta.$$

To satisfy (6.9), we require that

$$\begin{aligned}\frac{-1}{4\pi A_3}\frac{\partial f}{\partial x_2} &= A_1 \sin 2\pi x_2 - A_2 \sin 2\pi(x_1 - x_2) \\ &= A_1 \sin 2\pi x_2 + \frac{A_2(A_1 + A_3 \cos 2\pi x_2)\sin 2\pi x_2 - A_2 A_3 \sin 2\pi x_2 \cos 2\pi x_2}{M} \\ &= \left(A_1 + \frac{A_1 A_2}{M}\right)\sin 2\pi x_2 = 0.\end{aligned}$$

Since A_1, A_2 and M are positive, we must have $x_2 = n/2$. But this leads to the trivial solutions again.

Suppose k is odd in (6.14). Now,

$$\cos 2\pi x_1 = -\cos 2\pi\theta \qquad \text{and} \qquad \sin 2\pi x_1 = -\sin 2\pi\theta$$

and satisfaction of (6.9) requires that

$$\frac{-1}{4\pi A_3}\frac{\partial f}{\partial x_2} = \left(A_1 - \frac{A_1 A_2}{M}\right)\sin 2\pi x_2 = 0.$$

Apart from the trivial solutions arising from $\sin 2\pi x_2 = 0$, we see that a non-trivial solution may exist when $M = A_2$. Hence,

$$A_1^2 + A_2^2 + 2A_1 A_3 \cos 2\pi x_2 = A_2^2$$

and thus

$$\cos 2\pi x_2 = \frac{-A_1^2 + A_2^2 - A_3^2}{2A_1 A_3}. \tag{6.15}$$

A solution exists when the absolute value of the right hand side of (6.15) is less than or equal to one. Assuming this condition is satisfied then, after substituting (6.13), we find that

$$\cos 2\pi x_1 = \frac{-A_1^2 - A_2^2 + A_3^2}{2A_1 A_2} \tag{6.16}$$

and so there exist non-trivial solutions of (6.9) at

$$\mathbf{x} = \left(\pm\frac{1}{2\pi} \arccos \frac{-A_1^2 - A_2^2 + A_3^2}{2A_1A_2},\ \mp\frac{1}{2\pi} \arccos \frac{-A_1^2 + A_2^2 - A_3^2}{2A_1A_3} \right) + \mathbf{k}$$

where $\mathbf{k} \in \mathbb{Z}^2$ and the range of arccos is assumed to be $[0, \pi)$ and the opposite signs of the first and second coordinates are dictated by the opposite signs of $\sin 2\pi x_1$ and $\sin 2\pi x_2$.

Now, with $\cos 2\pi x_1$ and $\cos 2\pi x_2$ as given by (6.16) and (6.15), respectively, we find that

$$\begin{aligned} \cos 2\pi(x_1 - x_2) &= \cos 2\pi x_1 \cos 2\pi x_2 + \sin 2\pi x_1 \sin 2\pi x_2 \\ &= \frac{-1}{A_2}\big[(A_1 + A_3 \cos 2\pi x_2) \cos 2\pi x_2 + A_3 \sin^2 2\pi x_2\big] \\ &= \frac{-1}{A_2}\big[(A_1 + A_3 \cos 2\pi x_2) \cos 2\pi x_2 + A_3\big(1 - \cos^2 2\pi x_2\big)\big] \\ &= -\frac{A_1 \cos 2\pi x_2 + A_3}{A_2} \\ &= \frac{A_1^2 - A_2^2 - A_3^2}{2A_2A_3}. \end{aligned} \tag{6.17}$$

Substitution of this equality, together with (6.15) and (6.16), into (6.6) and (6.7) yields

$$\frac{\partial^2 f}{\partial x_1^2} = 8\pi^2 A_2^2 > 0$$

and

$$\frac{\partial^2 f}{\partial x_2^2} = 8\pi^2 A_3^2 > 0$$

so these non-trivial solutions are not local maxima. Similar substitution into (6.8) yields

$$\frac{\partial^2 f}{\partial x_1 x_2} = 4\pi^2\big(A_1^2 - A_2^2 - A_3^2\big).$$

Now,

$$\left(\frac{\partial^2 f}{\partial x_1 x_2}\right)^2 \left(\frac{\partial^2 f}{\partial x_1^2}\frac{\partial^2 f}{\partial x_2^2}\right)^{-1} = \left(\frac{A_1^2 - A_2^2 - A_3^2}{2A_1A_2}\right)^2 = \cos^2 2\pi(x_1 - x_2) \leqslant 1$$

so (6.11) is satisfied whenever the non-trivial solutions exist. Thus, if the non-trivial solutions exist then they are local minima. Furthermore, substitution of the expressions for $\cos 2\pi x_1$, $\cos 2\pi x_2$ and $\cos 2\pi(x_1 - x_2)$ in (6.16), (6.15) and (6.17) into that for f in (6.3) reveals that $f(x_1, x_2) = 0$ for these values of x_1 and x_2. Since f is non-negative, it is clear that, whenever these non-trivial solutions exist, they yield global minima.

We have now shown that the only local maxima on $\mathbb{R}^2$ are the elements of $\mathbb{Z}^2$. However, we need to consider the restriction of f to the unit square $\mathcal{T}$. Having found

the local maxima in $\mathbb{R}^2$, we need only examine the behaviour of f on the boundary of $\mathcal{T}$. On the boundary of $\mathcal{T}$, f takes the form of a cosine in the free variable with a non-negative coefficient. Therefore, the only maxima on the boundary which require investigation are those which occur at $\left(\pm\frac{1}{2}, 0\right)$ and $\left(0, \pm\frac{1}{2}\right)$. However, these are trivial solutions of (6.9) and we know from our discussion above that they are not local maxima in $\mathbb{R}^2$. Hence, they cannot be local maxima on $\mathcal{T}$.

Therefore, Theorem 6.1 can be applied to functions in the form of (6.3).

Consider the way in which the above example might be used to locate the positions of peaks in a periodogram. We define a PERIODOGRAM as any function of the form

$$F(\omega) = \left| \sum_{j=1}^{n} A_j e^{-i\omega t_j} \right|^2$$

where the $A_j \in \mathbb{C}$ are called AMPLITUDES and the $t_j \in \mathbb{R}$ are called SAMPLE TIMES. Consider periodograms of 3 points (that is, $n = 3$) with real, positive amplitudes and distinct sample times ordered so that $t_1 < t_2 < t_3$. In this case, we have

$$\begin{aligned} F(\omega) &= \left| A_1 e^{-i\omega t_1} + A_2 e^{-i\omega t_2} + A_3 e^{-i\omega t_3} \right|^2 \\ &= \left| A_1 + A_2 e^{-i\omega (t_2 - t_1)} + A_3 e^{-i\omega (t_3 - t_1)} \right|^2 \\ &= \left| A_1 + A_2 e^{-i2\pi x \alpha} + A_3 e^{-i2\pi x} \right|^2 \end{aligned}$$

where

$$x = \frac{\omega}{2\pi(t_3 - t_1)} \qquad \text{and} \qquad \alpha = \frac{t_2 - t_1}{t_3 - t_1}.$$

Clearly, we have transformed the periodogram into a diagonal function of the type discussed in Example 6.2.

For example, consider the specific case where the sample times are $t_1 = 1$, $t_2 = \sqrt{2}$ and $t_3 = 2$ and the amplitudes are $A_1 = 1$, $A_2 = 0.8$ and $A_3 = 0.6$. A graph of the diagonal function $f(x\alpha, x)$ which results is shown in Figure 4. Successive maxima in the sense of (6.1) and (6.2) in Theorem 6.1 are indicated by heavy lines. The horizontal dotted lines which connect the occurrences of successive maxima indicate that none have been missed. The vertical dotted lines indicate the boundaries of intervals in which $(\lfloor x\alpha \rceil, \lfloor x \rceil)$ have constant values. We see that, as Theorem 6.1 implies, the successive maxima occur where $(\lfloor x\alpha \rceil, \lfloor x \rceil)$ are best homogeneous Diophantine approximations of α (*cf.* Figure 1).

7. Farey Series

The Farey series of order n, $\mathfrak{F}_n$, is simply the series of fractions in lowest terms in ascending order, such that the denominators of each are positive and less than or equal to n. Some authors impose the restriction that the series consist only of fractions between and including 0 and 1. We do not impose this restriction here. Table 1 lists the Farey series between 0 and 1 for orders one to five.

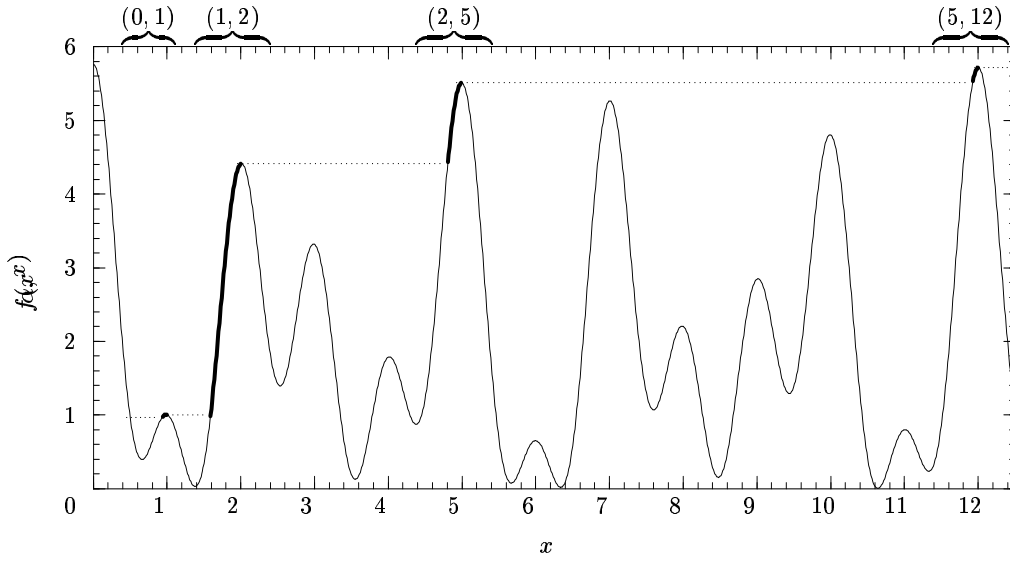


FIGURE 4. Plot of $f(x\alpha, x)$ for Example 6.2 with $\alpha = \sqrt{2}-1$, $A_1 = 1$, $A_2 = 0.8$ and $A_3 = 0.6$.

TABLE 1. The Farey series up to order five between 0 and 1.

$$
\begin{array}{ccccccccccc}
\frac{0}{1} & & & & & & & & & & \frac{1}{1} \\
\frac{0}{1} & & & & & \frac{1}{2} & & & & & \frac{1}{1} \\
\frac{0}{1} & & & \frac{1}{3} & & \frac{1}{2} & & \frac{2}{3} & & & \frac{1}{1} \\
\frac{0}{1} & & \frac{1}{4} & \frac{1}{3} & & \frac{1}{2} & & \frac{2}{3} & \frac{3}{4} & & \frac{1}{1} \\
\frac{0}{1} & \frac{1}{5} & \frac{1}{4} & \frac{1}{3} & \frac{2}{5} & \frac{1}{2} & \frac{3}{5} & \frac{2}{3} & \frac{3}{4} & \frac{4}{5} & \frac{1}{1}
\end{array}
$$

As a point of history, we note that the Farey series were, in fact, first investigated by HAROS in 1802, but rediscovered by FAREY in 1816 (DICKSON, 1919; HARDY & WRIGHT, 1979).

We begin this section by stating an elementary theorem of Farey series. The theorem and proof are adapted from HARDY & WRIGHT (1979) and NIVEN & ZUCKERMAN (1980). We then explore the relationship between the elements of the Farey series with the convergents and intermediate fractions of the s.c.f. expansions of real numbers.

DEFINITION 7.1. The MEDIANT of two fractions h/k and h'/k' is $(h+h')/k+k'$.

THEOREM 7.1. *If $h/k < h'/k'$ are adjacent elements of $\mathfrak{F}_n$ then*

$$h'k - hk' = 1. \tag{7.1}$$

Moreover, the fractions are also adjacent in the Farey series of higher order up to but not including $\mathfrak{F}_{(k+k')}$, in which the mediant, $(h+h')/(k+k')$, is the sole element separating them.

PROOF. The proof is by induction on n. Clearly, (7.1) is satisfied for $n = 1$. Suppose it is satisfied for all adjacent elements in $\mathfrak{F}_n$ for $1 \leqslant n < N$. Let H/K be a new element which occurs in $\mathfrak{F}_N$ which lies between $h/k < h'/k'$, adjacent elements in $\mathfrak{F}_{N-1}$. Obviously, $K = N$.

Consider the differences

$$\frac{H}{K} - \frac{h}{k} = \frac{Hk - hK}{kK} > 0 \qquad \text{and} \qquad \frac{h'}{k'} - \frac{H}{K} = \frac{h'K - Hk'}{k'K} > 0.$$

Let the numerators of these fractions be r and s, respectively. That is,

$$r = Hk - hK > 0 \qquad \text{and} \qquad s = h'K - Hk' > 0.$$

We solve for H and K, bearing in mind (7.1), to discover that

$$H = rh' + sh \qquad \text{and} \qquad K = rk' + sk.$$

Now consider the set of all fractions of the form

$$\frac{ah + bh'}{ak + bk'}$$

with $a, b \in \mathbb{N}$. If u/v is one such fraction then $h/k < u/v < h'/k'$ and that fraction with (uniquely) least denominator is the mediant, $(h+h')/(k+k')$. Thus

$$\frac{H}{K} = \frac{h + h'}{k + k'}$$

and therefore the mediant must be the sole element in $\mathfrak{F}_N$ separating h/k and h'/k'. Also we see again that

$$Hk - hK = (h + h')k - h(k + k') = 1$$

and

$$h'K - Hk' = h'(k + k') - (h + h')k' = 1$$

so the theorem statement is true for $n = N$ also. The process of induction is therefore complete. □

Theorem 7.1 suggests the following recursive procedure for computing successive elements of the Farey series of a prescribed order.

ALGORITHM 7.1.

```
1 proc farey(h, k, h', k', n) ≡
2   if k + k' ⩽ n then
3     farey(h, k, h + h', k + k', n);
4     output((h + h')/(k + k'));
5     farey(h + h', k + k', h', k', n);
6   fi.
```

PROPOSITION 7.1. *If $h/k < h'/k'$ are adjacent elements of the Farey series $\mathfrak{F}_m$, for some m, then the procedure* farey *of Algorithm 7.1 recursively calculates and outputs in ascending order the elements of $\mathfrak{F}_n$ which lie strictly between h/k and h'/k'.*

PROOF. The proof is immediate from inspection of Algorithm 7.1 and consideration of Theorem 7.1. □

LEMMA 7.1. *Suppose $h/k < h'/k'$ are adjacent in $\mathfrak{F}_n$ and $n < 1/\epsilon$ for some $\epsilon \in \mathbb{R}$. If $\alpha \in [h/k, h'/k']$ and $H/K \in \mathfrak{F}_n$ then*

$$|K\alpha - H| \leqslant \epsilon \;\Rightarrow\; \frac{H}{K} \in \left\{\frac{h}{k}, \frac{h'}{k'}\right\}. \tag{7.2}$$

PROOF. Suppose (7.2) is not satisfied. Assuming, without loss of generality, that $H/K < \alpha$, we have

$$|K\alpha - H| \leqslant \epsilon$$

and there is an adjacent element H'/K' such that

$$\frac{H}{K} < \frac{H'}{K'} \leqslant \alpha.$$

This implies that

$$\left|\alpha - \frac{H}{K}\right| \geqslant \left|\frac{H'}{K'} - \frac{H}{K}\right| = \frac{H'K - HK'}{KK'} = \frac{1}{KK'}$$

from Theorem 7.1. After multiplication throughout by K, we see that

$$|K\alpha - H| \geqslant \frac{1}{K'} \geqslant \frac{1}{n} > \epsilon,$$

contrary to our assumption. □

LEMMA 7.2. *Suppose $h/k < h'/k'$ are adjacent in $\mathfrak{F}_n$ and $n + 1 \geqslant 1/\epsilon$ for some $\epsilon \in \mathbb{R}$. If $\alpha \in [h/k, h'/k']$ then either*

$$|k\alpha - h| \leqslant \epsilon \qquad \textit{or} \qquad |k'\alpha - h'| \leqslant \epsilon.$$

PROOF. Suppose that

$$\frac{h}{k} \leqslant \alpha \leqslant \frac{h + h'}{k + k'}. \tag{7.3}$$

This implies that

$$\left|\alpha - \frac{h}{k}\right| \leqslant \left|\frac{h + h'}{k + k'} - \frac{h}{k}\right| = \frac{1}{k(k + k')}$$

and so

$$|k\alpha - h| \leqslant \frac{1}{k + k'} \leqslant \frac{1}{n + 1} \leqslant \epsilon.$$

If we suppose, instead of (7.3), that

$$\frac{h + h'}{k + k'} \leqslant \alpha \leqslant \frac{h'}{k'}$$

then we find that $|k'\alpha - h'| \leqslant \epsilon$. □

THEOREM 7.2. *Suppose $h/k < h'/k'$ are adjacent elements of the Farey series $\mathfrak{F}_n$ for some $n > 0$ and p_{i-1}/q_{i-1} and p_i/q_i, $i > 0$, are the $(i-1)^{th}$ and i^{th} convergents of the s.c.f. expansion of some real number α. Let*

$$\eta_{i-1} = q_{i-1}\alpha - p_{i-1} \qquad \textit{and} \qquad \eta_i = q_i\alpha - p_i.$$

If

$$\frac{h}{k} \leqslant \alpha \leqslant \frac{h'}{k'} \qquad \textit{and} \qquad \frac{1}{|\eta_{i-1}|} - 1 \leqslant n < \frac{1}{|\eta_i|} \tag{7.4}$$

then

$$\frac{p_i}{q_i} \in \left\{\frac{h}{k}, \frac{h'}{k'}\right\}.$$

REMARK 7.1. Firstly, we remark that, for any α satisfying the left-hand inequalities of (7.4), there must be some integer n satisfying the right-hand inequalities of (7.4) because, from statement (*iii*) of Proposition 3.2, $|\eta_i| < |\eta_{i-1}|$.

PROOF OF THEOREM 7.2. Lemma 7.1 implies that if $p_i/q_i \in \mathfrak{F}_n$ then it is an element of $\{h/k, h'/k'\}$. Lemma 7.2 implies that there is an element of this set, H/K, which satisfies $|K\alpha - H| \leqslant |\eta_{i-1}|$. Proposition 3.5 implies that either $(H, K) = (p_i, q_i)$ or $K > q_i$. Thus, p_i/q_i must be a member of the set. □

THEOREM 7.3. *The convergent p_i/q_i and the intermediate fraction*

$$\frac{p_{i-1} + kp_i}{q_{i-1} + kq_i}$$

of the s.c.f. expansion of a real number α are adjacent in the series $\mathfrak{F}_n$ for $i \geqslant 0$, $0 < k \leqslant a_{i+1}$ and $q_{i-1} + kq_i \leqslant n < q_{i-1} + (k+1)q_i$.

PROOF. We will prove the theorem by induction. Firstly, we see that the theorem is true for $i = 0$ and $k = 1$ since $(p_{-1}, q_{-1}) = (1, 0)$ and $(p_0, q_0) = a_0, 1$ and $a_0/1$ and $(a_0 + 1)/1$ are adjacent in $\mathfrak{F}_1$.

Assume the theorem is true for $i = I$ and $0 < k < K \leqslant a_{I+1}$. Assume also that the theorem is true for all $0 \leqslant i < I$. Therefore, the convergent p_I/q_I and the intermediate fraction $[p_{I-1} + (K-1)p_I]/[q_{I-1} + (K-1)q_I]$ are adjacent in $\mathfrak{F}_n$ for $q_{I-1} + (K-1)q_I \leqslant n \leqslant q_{I-1} + Kq_I$. Clearly, the mediant of these fractions is the next intermediate fraction, $(p_{I-1} + Kp_I)/(q_{I-1} + Kq_I)$. From Theorem 7.1, we see that the theorem is true for $i = I$ and $k = K$ also.

To complete the induction, we observe that if the theorem is true for all $0 < i \leqslant I$ and $k = a_{I+1}$ then p_I/q_I and p_{I+1}/q_{I+1} are adjacent in $\mathfrak{F}_n$ where $q_{I+1} \leqslant n \leqslant q_I + q_{I+1}$. Therefore, the theorem will again be true for $i = I + 1$ and $k = 1$. □

REMARK 7.2. From Corollary 3.1, we know that α lies between the adjacent elements of the Farey series referred to in the previous theorem.

THEOREM 7.4. *If $h/k < h'/k'$ are adjacent elements in a Farey series such that their mediant is p_i/q_i, the i^{th} convergent, $i > 0$, of the s.c.f. expansion of a real number $\alpha \neq p_i/q_i$, then*

$$\frac{p_{i-1}}{q_{i-1}} = \begin{cases} \dfrac{h}{k} & \text{if } \alpha < \dfrac{p_i}{q_i}, \\ \dfrac{h'}{k'} & \text{otherwise.} \end{cases}$$

PROOF. Theorem 7.3 implies that one of the elements of the set $\{h/k, h'/k'\}$ is the convergent p_i/q_i and the other is the intermediate fraction

$$\frac{p_{i-2} + (a_i - 1)p_{i-1}}{q_{i-2} + (a_i - 1)q_{i-1}}.$$

Corollary 3.1 can then be applied for $k = a_i - 1$ and $k = a_i$ to complete the proof. □

REMARK 7.3. If $\alpha = p_i/q_i$ in the statement of the previous theorem then p_{i-1}/q_{i-1} is whichever of the fractions $\{h/k, h'/k'\}$ has the lesser denominator. This occurs because $\eta_i = q_i\alpha - p_i = 0$ and Corollary 3.3 implies that $a_i \geqslant 2$. Thus, $q_{i-2} + (a_i - 1)q_{i-1} > q_{i-1}$ and hence the convergent p_{i-1}/q_{i-1} has the lesser denominator.

In this section, we have seen that the Farey series are closely related to homogeneous Diophantine approximation of a real number. We have seen, in Theorem 7.2, that the best approximations of a given real number within a certain approximation error can be easily located in the Farey series of appropriate order. We have also found, in Theorem 7.3, that the elements of a Farey series immediately surrounding a given real number are always convergents or intermediate fractions of the s.c.f. expansion of that number.

CHAPTER 3

GEOMETRY OF NUMBERS

1. Historical Remarks

The geometry of numbers, as its name suggests, has a distinctly geometric "flavour." However, it is important to bear in mind that many of the results which can be reinterpreted in the light of this theory, such as GAUSS' reduction algorithm, precede it and are couched in terms of the theory of quadratic forms. So, by way of introduction to this section, we make some historical remarks about the development of the geometry of numbers. These remarks are mainly drawn from SCHARLAU & OPOLKA (1985).

One of the origins of the geometry of numbers is in the study of the representation of natural numbers by sums of squares. DIOPHANTOS knew some basic theorems and FERMAT developed the theory in the first half of the 17^{th} century, stating (but not proving) the famous "two-square" and "four-square" theorems.

THEOREM 1.1 (Two-Square Theorem). *Every prime number of the form $4k+1$ can be written uniquely as a sum of two squares.*

THEOREM 1.2 (Four-Square Theorem). *Every natural number is a sum of four squares of natural numbers (where zero is allowed as a summand).*

FERMAT was also interested in certain specific quadratic forms, namely x^2+y^2, x^2+2y^2, x^2+3y^2 and x^2-dy^2 (this last form being often associated with PELL). In particular, he was interested in which (prime) numbers could be expressed by these forms. However, it was LAGRANGE (1773) who systematically developed the theory of the representation of numbers by binary quadratic forms. A BINARY QUADRATIC FORM is a function of two integer variables,

$$q(x,y) = ax^2 + bxy + cy^2,$$

where $a, b, c \in \mathbb{Z}$ and a and c not zero. LAGRANGE considered the special form

$$Q(x,y) = ax^2 + 2bxy + cy^2.$$

LAGRANGE studied the equivalence classes between binary quadratic forms of this type. Two forms $ax^2+2bxy+cy^2$ and $AX^2+2BXY+CY^2$ are EQUIVALENT if there is an invertible integral linear substitution of variables from (x,y) to (X,Y)

which transforms the one form into the other. Thus, the two forms are equivalent if

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{M} \begin{pmatrix} X \\ Y \end{pmatrix}$$

where $\mathbf{M}$ is a 2×2 integral matrix with $\det \mathbf{M} = \pm 1$ (a *unimodular* matrix; see Definition 2.4). The two forms are said to be PROPERLY EQUIVALENT if $\det \mathbf{M} = 1$. Clearly, any number which can be represented by a certain binary quadratic form can be represented by any equivalent binary quadratic form. Furthermore, the discriminant of the binary quadratic form Q, $\Delta = 4(ac - b^2)$, is preserved under these transformations. Thus, each equivalence class has associated with it a discriminant. A binary quadratic form $Q(x, y)$ is called POSITIVE if $Q(x, y) > 0$ for all $x, y \in \mathbb{Z}$ apart from $x = y = 0$. A binary quadratic form is positive if and only if its discriminant is positive and the coefficients of the squares (a and c) are positive.

LAGRANGE identified a member of each proper equivalence class of positive binary quadratic forms, a so-called REDUCED form.

THEOREM 1.3. *A positive binary quadratic form* $Q'(X, Y) = AX^2 + 2BXY + CY^2$ *is properly equivalent to a reduced form* $Q(x, y) = ax^2 + 2bxy + cy^2$. *By reduced we mean that the coefficients* a, b *and* c *satisfy*

$$\frac{-a}{2} < b \leqslant \frac{a}{2}, \quad 0 < a \leqslant c \quad \textit{and} \quad a \leqslant \sqrt{\frac{\Delta}{3}} \tag{1.1}$$

where Δ *is the discriminant of* Q *and* Q'. *Furthermore, if* $a = c$ *then* $0 \leqslant b \leqslant a/2$.

REMARK 1.1. The rightmost inequality in (1.1) is implied by the other inequalities because

$$\frac{\Delta}{3} = \frac{4(ac - b^2)}{3} \geqslant \frac{4(a^2 - b^2)}{3} \geqslant \frac{4\left(a^2 - \frac{1}{4}a^2\right)}{3} \geqslant a^2.$$

LAGRANGE was then able to determine that the number of proper equivalence classes of positive binary quadratic forms for a given discriminant must be finite and he was able to tabulate these classes on the basis of discriminant and its reduced form.

GAUSS (1801) was the first to emphasise algorithms, describing a method by which quadratic forms can be reduced. For this reason the lattice basis reduction algorithm in two dimensions, which we discuss in Section 6.1 is named after him.

However, it was MINKOWSKI (1896*b*) who invented the geometry of numbers. He fully understood and exploited the idea of the point lattice to develop the theory. For the purpose of this thesis, we only require a tiny fraction of this theory. Of his results, we require only his First and Second Theorems and his notion of a reduced lattice basis.

We do not set out to detail the results of original research in this chapter. The purpose is provide a link between the material of the previous chapter on Diophantine approximation and that of the next chapter on simultaneous Diophantine

approximation, which is most conveniently and elegantly expressed in the language of the geometry of numbers.

In the next section, we introduce point lattices. We will then define convex bodies and state MINKOWSKI's First or Fundamental Theorem in Section 3. After briefly reviewing the properties of the **QR** decomposition and the Cholesky decomposition of a matrix in Section 4, we immediately apply the properties to the formulation of an algorithm for finding shortest vectors in a lattice in Section 5. However, the computational infeasibility of this algorithm prompts a review of lattice reduction in Section 6. To conclude this chapter, we discuss a fast algorithm for lattice reduction, the so-called LLL algorithm of LENSTRA *et al.* (1982) in Section 7.

2. Point Lattices

DEFINITION 2.1. Consider a set $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ of linearly independent points in $\mathbb{R}^m$, $m \geqslant n$. The set

$$\Omega = \{a_1\mathbf{b}_1 + a_2\mathbf{b}_2 + \cdots + a_n\mathbf{b}_n \mid a_1, a_2, \ldots, a_n \in \mathbb{Z}\}$$

is a POINT LATTICE (or simply lattice) of RANK n in $\mathbb{R}^m$ and $\mathcal{B}$ is a BASIS of Ω.

DEFINITION 2.2. A BASIS MATRIX of a lattice Ω is a matrix consisting of column vectors which together form a basis of Ω.

DEFINITION 2.3. A FUNDAMENTAL PARALLELEPIPED of a lattice Ω is any parallelepiped constructed from a set of basis vectors of Ω.

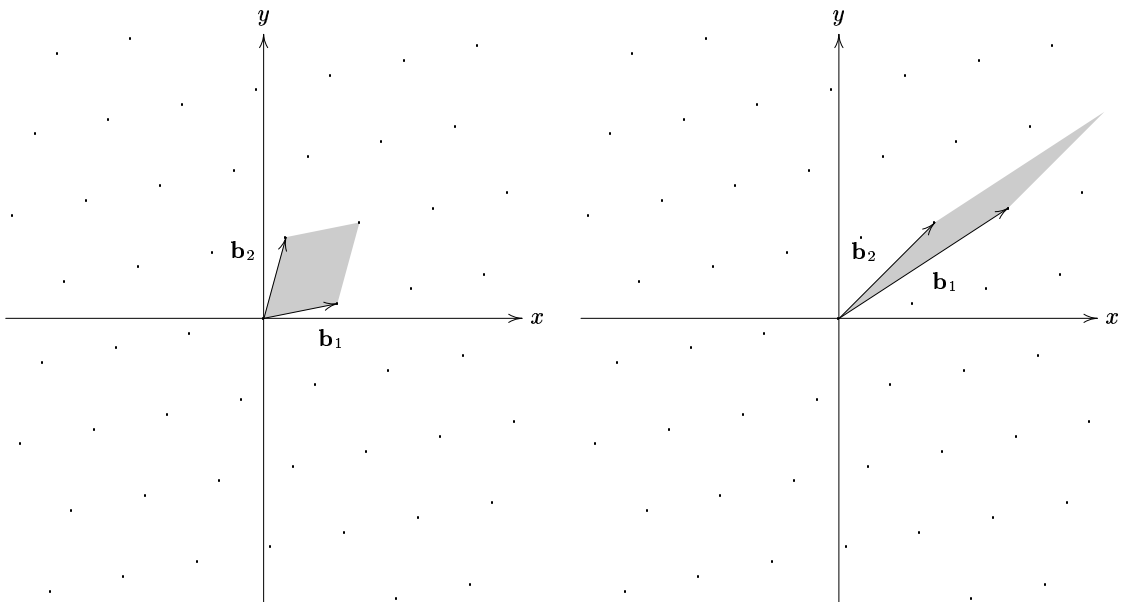


FIGURE 1. Two bases (indicated by vectors) and fundamental parallelograms (indicated by shading) of a lattice of rank two.

Figure 1 shows an examples of a lattice of rank two. The same lattice is depicted in both diagrams, but different bases are used. The corresponding fundamental parallelepipeds (in this case, parallelograms) are indicated by the shaded regions.

Theorems 2.1 to 2.4 are adapted from SIEGEL (1989).

THEOREM 2.1. *Let* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ *be* n *linearly independent vectors from a lattice* Ω *of rank* n*. There exists a basis* $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ *for* Ω *such that*

$$\begin{aligned}
\mathbf{b}_1 &= c_{11}\mathbf{v}_1, \\
\mathbf{b}_2 &= c_{12}\mathbf{v}_1 + c_{22}\mathbf{v}_2, \\
&\vdots \\
\mathbf{b}_n &= c_{1n}\mathbf{v}_1 + c_{2n}\mathbf{v}_2 + \cdots + c_{nn}\mathbf{v}_n.
\end{aligned}$$

where $c_{ij} \in \mathbb{Q}$ *for* $1 \leqslant i \leqslant j \leqslant n$ *and* $1/c_{ii} \in \mathbb{N}$ *for* $1 \leqslant i \leqslant n$*.*

The basis of a lattice is not unique. The definition and theorems which follow describe how other bases can be obtained from a given one.

DEFINITION 2.4. A square $n \times n$ integer matrix is UNIMODULAR if the value of its determinant is ± 1.

THEOREM 2.2. *The inverse of a unimodular matrix is unimodular.*

THEOREM 2.3. *Let* $\mathbf{B}, \mathbf{B}' \in \mathbb{R}^{m\times n}$ *be basis matrices of the lattices* Ω *and* Ω'*, respectively, both of rank* n *in* $\mathbb{R}^m$*. The lattices are equal if and only if there exists some unimodular* $\mathbf{M} \in \mathbb{Z}^{n\times n}$ *such that* $\mathbf{B} = \mathbf{B}'\mathbf{M}$*.*

DEFINITION 2.5. The REAL SPAN of a lattice Ω is the set of all linear combinations with *real* coefficients of the vectors of any basis of Ω.

Similarly, we could refer to the lattice itself as the INTEGER SPAN of any of its bases, according to a similar definition.

DEFINITION 2.6. The set of linearly independent points $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ is a PRIMITIVE BASIS of the lattice Ω of rank $n \geqslant q$ if every point of Ω which also lies in the real span of $\mathcal{V}$ is a point of the lattice generated by $\mathcal{V}$.

THEOREM 2.4. *If* $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ *is a primitive basis of a lattice* Ω *of rank* $n > q$ *then there exist* $n - q$ *vectors* $\mathbf{b}_{q+1}, \mathbf{b}_{q+2}, \ldots, \mathbf{b}_n$ *such that*

$$\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q, \mathbf{b}_{q+1}, \mathbf{b}_{q+2}, \ldots, \mathbf{b}_n\}$$

is a basis of Ω*.*

DEFINITION 2.7. Consider a set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ of linearly independent points in $\mathbb{R}^n$. The set

$$\mathcal{O}(\mathcal{V}) = \{\mu_1\mathbf{v}_1 + \mu_2\mathbf{v}_2 + \cdots + \mu_q\mathbf{v}_q \mid \mu_1, \mu_2, \ldots, \mu_q \in \mathbb{R}; \\ |\mu_1| + |\mu_2| + \cdots + |\mu_q| \leqslant 1\}$$

is the HYPEROCTAHEDRAL of $\mathcal{V}$.

DEFINITION 2.8. The hyperoctahedral $\mathcal{O}$ of $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$ is PERFECT in a lattice Ω if

$$\mathcal{O}(\mathcal{V}) \cap \Omega \subseteq \{0, \pm\mathbf{v}_1, \pm\mathbf{v}_2, \ldots, \pm\mathbf{v}_q\},$$

where $\pm$ designates both its argument and its additive inverse (symmetric point).

THEOREM 2.5. *If $\mathbf{v}_1$ and $\mathbf{v}_2$ are points in a lattice Ω and the hyperoctahedral of $\{\mathbf{v}_1, \mathbf{v}_2\}$ is perfect in Ω then $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a primitive basis of Ω.*

REMARK 2.1. Obviously, in this case, the hyperoctahedral of $\{\mathbf{v}_1, \mathbf{v}_2\}$ is simply a parallelogram with vertices at $\pm\mathbf{v}_1$ and $\pm\mathbf{v}_2$.

PROOF OF THEOREM 2.5. If $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$ do not form a primitive basis of Ω then there exists a non-zero point $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \in \Omega$ with $c_1, c_2 \in \mathbb{Q}$ and $0 \leqslant c_1, c_2 < 1$. It follows that

$$(c_1 - \lfloor c_1 \rceil)\mathbf{v}_1 + (c_2 - \lfloor c_2 \rceil)\mathbf{v}_2 \in \mathcal{O}(\mathcal{V}) \cap \Omega,$$

contradicting the assumption that $\mathcal{O}(\mathcal{V})$ is perfect in Ω. □

The following theorem is due to FURTWÄNGLER (1927).

THEOREM 2.6. *If $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are linearly independent points in a lattice Ω and the hyperoctahedral of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is perfect in Ω then either $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ or $\{\mathbf{v}_1, \mathbf{v}_2, (\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3)/2\}$ is a primitive basis of Ω.*

PROOF. Every lattice point of Ω in the real span of $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ can be expressed as a linear combination of these vectors with rational coefficients. Let us discuss the conditions under which $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ can define a perfect octahedral and yet not form a primitive basis of Ω. Suppose $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ does not form a primitive basis, so there must exist some $\mathbf{u} \in \Omega$ that can be expressed

$$\mathbf{u} = \frac{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3}{d}$$

where $c_1, c_2, c_3, d \in \mathbb{Z}$ and this fraction is expressed in its lowest terms with $d > 1$. This lattice point will lie within the octahedral if $|c_1| + |c_2| + |c_3| \leqslant d$.

Now, assume there is no integer q for which

$$qc_1 \equiv 1 \pmod{d}. \tag{2.1}$$

Then there exists an $r = d/\gcd(c_1, d) < d$ such that

$$rc_1 \equiv 0 \pmod{d}.$$

We can write

$$r\mathbf{u} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \frac{b_2\mathbf{v}_2 + b_3\mathbf{v}_3}{d}$$

where $a_1, a_2, a_3 \in \mathbb{Z}$ are chosen so that $-d/2 < b_2, b_3 \leqslant d/2$ and b_2 and b_3 are not both zero because otherwise $\mathbf{u}$ would not have been expressed in its lowest terms. Hence,

$$\mathbf{u}' = \frac{b_2\mathbf{v}_2 + b_3\mathbf{v}_3}{d}$$

is a lattice point and $|b_2| + |b_3| \leqslant d$ so $\mathbf{u}' \in \mathcal{O}(\mathcal{V})$, contrary to our assumption.

On the other hand, suppose (2.1) holds for some q. Then

$$q\mathbf{u} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 + \frac{\mathbf{v}_1 + b_2\mathbf{v}_2 + b_3\mathbf{v}_3}{d}$$

where, once again, $a_1, a_2, a_3 \in \mathbb{Z}$ are chosen so that $-d/2 < b_2, b_3 \leqslant d/2$. Again,

$$\mathbf{u}' = \frac{\mathbf{v}_1 + b_2\mathbf{v}_2 + b_3\mathbf{v}_3}{d}$$

must also be a lattice point in Ω.

Unless $b_2 = b_3 = d/2$, we will have $1 + |b_2| + |b_3| \leqslant d$ and therefore $\mathbf{u}' \in \mathcal{O}(\mathcal{V})$ and the octahedral is not perfect. However, if $b_2 = b_3 = d/2$ then

$$2\mathbf{u}' = \frac{2\mathbf{v}_1}{d} + \mathbf{v}_2 + \mathbf{v}_3$$

and $2\mathbf{v}_1/d$ will also be a lattice point. If $d > 2$ then $2\mathbf{u}' \in \mathcal{O}(\mathcal{V})$. This leaves only one possibility.

Therefore, unless $(\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3)/2$ is a lattice point, $\mathcal{V}$ forms a primitive basis of Ω. If $(\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3)/2$ is a lattice point then a set consisting of this point and any two of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ forms a primitive basis. □

3. Convex Bodies and Minkowski's Theorem

The presentation of definitions and theorems in this subsection is essentially a summary of Lectures I–III of Siegel (1989).

Definition 3.1. A non-empty set $\mathcal{S}$ in $\mathbb{R}^n$ is convex if, for every pair of elements $\mathbf{x}$ and $\mathbf{y}$ of $\mathcal{S}$,

$$\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in \mathcal{S}$$

for all $0 \leqslant \lambda \leqslant 1$, $\lambda \in \mathbb{R}$.

Definition 3.2. A point $\mathbf{x}$ in a set $\mathcal{S}$ in $\mathbb{R}^n$ is an interior point of $\mathcal{S}$ if there exists a metric ball centred at $\mathbf{x}$ which is contained in $\mathcal{S}$.

Definition 3.3. A set $\mathcal{S}$ in $\mathbb{R}^n$ is open if it consists only of interior points.

Definition 3.4. The interior of a set $\mathcal{S}$ in $\mathbb{R}^n$ is the set of interior points of $\mathcal{S}$.

We use the notation $\operatorname{Int}\mathcal{S}$ to denote the interior of $\mathcal{S}$.

Definition 3.5. A convex body is a set in $\mathbb{R}^n$ which is convex, open and bounded.

DEFINITION 3.6. A set $\mathcal{S}$ in $\mathbb{R}^n$ is CENTRALLY SYMMETRIC if there exists a point $\mathbf{c} \in \mathcal{S}$, the CENTRE of $\mathcal{S}$, such that, for all $\mathbf{x} \in \mathcal{S}$, $2\mathbf{c} - \mathbf{x} \in \mathcal{S}$.

DEFINITION 3.7. The CHARACTERISTIC FUNCTION of a set $\mathcal{S}$ in $\mathbb{R}^n$ is mapping from $\mathbb{R}^n$ to $\mathbb{R}$ taking the value 1 when the argument is an element of $\mathcal{S}$ and 0 otherwise.

The following definition of volume is not the most general possible, but it will be sufficient for our purposes.

DEFINITION 3.8. If the characteristic function of a set $\mathcal{S}$ in $\mathbb{R}^n$ is Riemann integrable then its integral is the VOLUME of $\mathcal{S}$.

We will use the notation $\operatorname{vol}\mathcal{S}$ to denote the volume of $\mathcal{S}$.

THEOREM 3.1. *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be a basis matrix of a lattice Ω of rank n in $\mathbb{R}^n$. Then the volume of any fundamental parallelepiped of Ω is* $\det \mathbf{B}$.

We are now in a position to state MINKOWSKI's First (or Fundamental) Theorem.

THEOREM 3.2. *Let $\mathcal{S}$ be a centrally symmetric convex body about the origin in $\mathbb{R}^n$ and let Ω be a lattice of rank n in $\mathbb{R}^n$. Let Δ be the volume of a fundamental parallelepiped of Ω. If $\operatorname{vol}\mathcal{S} > 2^n\Delta$ then there exists a non-zero lattice point of Ω in $\mathcal{S}$.*

4. The QR decomposition and the Cholesky decomposition

Although the **QR** decomposition and the Cholesky decomposition of matrices are more often associated with linear algebra than with the geometry of numbers, we nevertheless find it necessary to recall these basic results.

Recall the following definitions of a (column) orthogonal, an upper (lower) triangular and a positive definite matrix.

DEFINITION 4.1. A matrix $\mathbf{Q} \in \mathbb{R}^{m\times n}$ is COLUMN ORTHOGONAL if $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. A matrix is ORTHOGONAL if it is square and column orthogonal.

DEFINITION 4.2. A matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ is UPPER (LOWER) TRIANGULAR if $r_{ij} = 0$ whenever $i > j$ $(i < j)$.

DEFINITION 4.3. A matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ is POSITIVE DEFINITE if $\mathbf{x}^T\mathbf{P}\mathbf{x} > 0$ for all non-zero $\mathbf{x} \in \mathbb{R}^n$.

Consider the following algorithm, operating on an input $m \times n$ matrix $\mathbf{B}$ of full column rank to produce an $m \times n$ matrix $\mathbf{Q}$ and an $n \times n$ matrix $\mathbf{R}$.

ALGORITHM 4.1 (Gram-Schmidt orthonormalisation).

```
1  begin
2    R := 0;
3    for k := 1 to n do
4        q_k := b_k;
5        for j := 1 to k − 1 do
6            r_jk := q_j · b_k;
7            q_k := q_k − r_jk q_j;
8        od;
9        r_kk := √(q_k · q_k);
10       q_k := q_k / r_kk;
11   od;
12   output(Q, R);
13 end.
```

This algorithm serves as proof of

THEOREM 4.1 ("Skinny" **QR** decomposition). *A matrix* $\mathbf{B} \in \mathbb{R}^{m \times n}$ *of full column rank can be uniquely expressed as*

$$\mathbf{B} = \mathbf{QR}$$

where $\mathbf{Q}$ *is an* $m \times n$ *column orthogonal matrix and* $\mathbf{R}$ *is an* $n \times n$ *upper triangular matrix with positive diagonal entries.*

Inspection of the algorithm reveals that it requires $O(n^2)$ iterations through the inner loop on lines 5–8. On each of these iterations, operations are performed on vectors of m elements and so the total number of arithmetic operations is $O(mn^2)$.

The Gram-Schmidt orthonormalisation procedure as described in Algorithm 4.1 is not the best method for performing the **QR** decomposition of a matrix from the point of view of numerical accuracy (GOLUB & VAN LOAN, 1989).

For any matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ of full column rank it is clear that

$$\mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x}) \cdot (\mathbf{B}\mathbf{x}) > 0$$

for all non-zero $\mathbf{x} \in \mathbb{R}^n$. If we write $\mathbf{P} = \mathbf{B}^T\mathbf{B}$ then $\mathbf{P}$ is positive definite. From the **QR** decomposition of $\mathbf{B}$ as $\mathbf{B} = \mathbf{QR}$ we find that we can also express $\mathbf{P}$ in terms of upper triangular matrices so that $\mathbf{P} = \mathbf{R}^T\mathbf{R}$. The following theorem shows that all symmetric positive definite matrices can be decomposed in this way.

THEOREM 4.2 (Cholesky decomposition). *If* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *is a symmetric, positive definite matrix then there exists a unique upper triangular matrix* $\mathbf{R} \in \mathbb{R}^{n \times n}$ *such that* $\mathbf{P} = \mathbf{R}^T\mathbf{R}$.

Although we will not bother to state and prove an algorithm for Cholesky decomposition here, it is not surprising that algorithms for Cholesky decomposition are very similar in principle to algorithms for **QR** decomposition.

Now, let us briefly consider the extension of **QR** decomposition to matrices which do not have full column rank. If $\mathbf{B}$ is an $m \times n$ matrix with column rank d then Algorithm 4.1 will produce a $m \times n$ matrix $\mathbf{Q}$ which will have $n - d$ columns consisting only of zeros. Similarly, there will be $n - d$ rows of the $n \times n$ matrix $\mathbf{R}$ which will consist only of zeros. We must also take care to avoid the problem of division of zero by zero. Where this would occur, we assign zero as the result. The **QR** decomposition of a matrix without full column rank obtained this way does not therefore have the property that $\mathbf{R}$ has positive diagonal entries, although they will be non-negative. Furthermore, the uniqueness of the decomposition no longer holds since, for example, any row of $\mathbf{R}$ which consists only of zeros can be replaced by a row of arbitrary values.

5. Finding Short Vectors in a Lattice

Consider the problem of finding short or shortest vectors in a lattice. We know from Theorem 2.3 that, for any given lattice of rank greater than one, there exist infinitely many bases. How do we find the shortest vector from a given basis? We now discuss this problem, following Cohen (1993), §2.7.3. To define exactly what we mean by "short," we require the concept of a norm. Firstly, we recall the definitions of a norm, a p-norm and the sup-norm.

Definition 5.1. A norm $\|\cdot\|$ in dimension n is a map from $\mathbb{R}^n$ to $\mathbb{R}$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, and for all $\lambda \in \mathbb{R}$,

(i) $\|\mathbf{x}\| > 0$,
(ii) $\|\lambda\mathbf{x}\| = |\lambda| \, \|\mathbf{x}\|$ and
(iii) $\|\mathbf{x} + \mathbf{x}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$

All norms are similar in a sense which is made precise by the following theorem.

Theorem 5.1. *For any two norms $\|\cdot\|$ and $\|\cdot\|'$ in dimension n there exists $\mu_1, \mu_2 \in \mathbb{R}$, $0 < \mu_1 < \mu_2$ such that, for all $\mathbf{x} \in \mathbb{R}^n$,*

$$\mu_1 \|\mathbf{x}\| \leqslant \|\mathbf{x}\|' \leqslant \mu_2 \|\mathbf{x}\| .$$

Definition 5.2. A p-norm $\|\cdot\|_p$ for some $p \in \mathbb{R}$, $p \geqslant 1$, in dimension n is a map from $\mathbb{R}^n$ to $\mathbb{R}$ which can be expressed for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}.$$

Definition 5.3. The sup-norm $\|\cdot\|_\infty$ in dimension n is a map from $\mathbb{R}^n$ to $\mathbb{R}$ which can be expressed for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}.$$

THEOREM 5.2. *A p-norm and the* sup*-norm are norms in any dimension.*

We use the notation $\|\cdot\|_\infty$ for the sup-norm because, for any $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p .$$

We will usually refer to the 2-norm as the EUCLIDEAN NORM.

Let us now devise a naïve algorithm for finding the shortest vector in $\mathbb{R}^2$ from a lattice Ω of rank 2 with respect to the Euclidean norm, $\|\cdot\|_2$. Given a basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ of Ω, we arrange the basis so that $\|\mathbf{b}_1\|_2 \geqslant \|\mathbf{b}_2\|_2$. We seek all lattice points $\mathbf{v} \in \Omega$ such that $\|\mathbf{v}\|_2 < \|\mathbf{b}_2\|_2$. Obviously, there are only finitely many. Consider the function

$$f(\lambda_1, \lambda_2) = \|\lambda_1\mathbf{b}_1 + \lambda_2\mathbf{b}_2\|_2^2 .$$

From partial differentiation of f with respect to λ_1, we find that f is minimised with respect to λ_1 when

$$\lambda_1 = \frac{-\lambda_2\mathbf{b}_1 \cdot \mathbf{b}_2}{\mathbf{b}_1 \cdot \mathbf{b}_1}$$

and thus we can deduce that $f(\lambda_1, \lambda_2) \geqslant \|\mathbf{b}_2\|_2^2 = f(0, 1)$ if

$$\lambda_2 \geqslant \frac{1}{\sin\theta} \tag{5.1}$$

where

$$\theta = \arccos \frac{\mathbf{b}_1 \cdot \mathbf{b}_2}{\|\mathbf{b}_1\|_2 \|\mathbf{b}_2\|_2} \tag{5.2}$$

is the acute angle between $\mathbf{b}_1$ and $\mathbf{b}_2$ (that is, $0 \leqslant \theta < \pi$).

For a given λ_1 which does not satisfy (5.1), we find that

$$f(\lambda_1, \lambda_2) < \|\mathbf{b}_2\|_2^2 \tag{5.3}$$

whenever

$$(\|\mathbf{b}_1\|_2 \lambda_1 + \|\mathbf{b}_2\|_2 \lambda_2 \cos\theta)^2 < \|\mathbf{b}_2\|_2^2 \left(1 - \lambda_2^2 \sin^2\theta\right).$$

This suggests the following algorithm and proposition.

ALGORITHM 5.1.

```
1 begin
2   if ‖b1‖2 < ‖b2‖2 then swap(b1, b2) fi;
3   v := b2;
4   c := (b1 · b2)/(‖b1‖2 ‖b2‖2);
5   s := √(1 − c²);
6   r := ‖b2‖2 / ‖b1‖2 ;
7   for k := 1 to ⌊1/s⌋ do
8       A := −kc;
9       B := √(1 − k²s²);
```

```
10        for j := ⌈r(A − B)⌉ to ⌊r(A + B)⌋ do
11            if ‖jb1 + kb2‖2 < ‖v‖2 then v := jb1 + kb2; fi;
12        od;
13    od;
14    output(v);
15 end.
```

PROPOSITION 5.1. *If Algorithm 5.1 is executed with the basis* $\{\mathbf{b}_1, \mathbf{b}_2\}$ *of a lattice* Ω *of rank* 2 *in* $\mathbb{R}^2$ *as its input then, after a finite number of iterations, the algorithm terminates yielding a lattice point* $\mathbf{v} \in \Omega$ *such that, for all* $\mathbf{0} \neq \mathbf{w} \in \Omega$, $\|\mathbf{v}\|_2 \leqslant \|\mathbf{w}\|_2$.

PROOF. We identify c with $\cos\theta$ and s with $\sin\theta$ as defined in (5.2). The proof is then by inspection of the algorithm and the preceding discussion. □

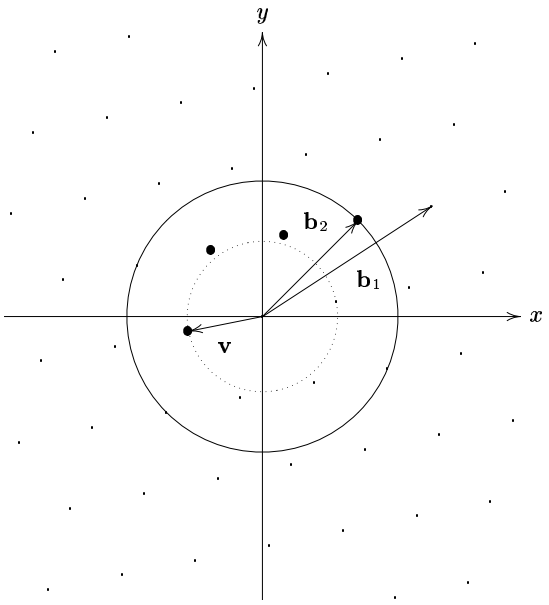


FIGURE 2. Graphical interpretation of the operation of Algorithm 5.1 on a lattice specified by the basis vectors $\{\mathbf{b}_1, \mathbf{b}_2\}$.

Figure 16 illustrates the operation of Algorithm 5.1 on a lattice which has been specified by the basis vectors $\{\mathbf{b}_1, \mathbf{b}_2\}$. The outer circle, depicted with a solid line, is the circle of radius $\|\mathbf{b}_2\|_2$ about the origin. The algorithm then tests each of the lattice points indicated by a bullet (•), comparing their norm against the shortest found so far. Finally, the algorithm outputs the vector $\mathbf{v}$. The inner circle, depicted with a dotted line, indicates the metric ball of radius $\|\mathbf{v}\|_2$ about the origin. Clearly, there are no lattice points other than the origin within the inner circle.

We see that the number of iterations through the outer loop of Algorithm 5.1 is governed by the inverse of the sine of the angle between the initial basis vectors. Moreover, because the algorithm tests every lattice point with norm less than $\|\mathbf{b}_2\|_2$ (after the *swap*), the time required by the algorithm is proportional to $\|\mathbf{b}_2\|_2^2$.

We could easily adapt Algorithm 5.1 to find the shortest vector with respect to other norms. For the p-norms, the algorithm can be adapted by reworking the formulae to find roots of polynomials of degree p, in place of the quadratic polynomial resulting from (5.3). For $p = 1, 2, 3, 4$ (and for the sup-norm) this can be done without fuss, but for integer $p > 4$ or non-integer p we need to resort to numerical methods. However, we only need to know the roots with sufficient accuracy to locate the nearby integer, so we expect that the computational overhead for such methods would not be great. Generalised to any norm $\|\cdot\|$, the total number of iterations through the inner loop of Algorithm 5.1 is still $O\big(\|\mathbf{b}_2\|^2\big)$.

Now consider the generalisation of this algorithm to lattices of arbitrary rank. For simplicity, consider finding the shortest vector with respect to the Euclidean norm. For this, we require **QR** decomposition of matrices. Equipped with a procedure *QRdecompose* for **QR** decomposition and a procedure *sortnorm* for sorting (permuting) the columns of its matrix argument in descending order of Euclidean norm, we can now state an algorithm which is analogous to Algorithm 5.1 for lattices of arbitrary rank n in $\mathbb{R}^m$ given $\mathbf{B}$, an input basis. The variable $\mathbf{k}$ is an integer vector and the variable $\mathbf{s}$ is a real vector and both have dimension n.

ALGORITHM 5.2.

```
1  begin
2    sortnorm(B);
3    QRdecompose(B, Q, R);
4    k := 0;
5    j := 1;
6    while j ⩽ n do
7          k_j := k_j + 1;
8          s := Rk;
9          S := 0;
10         for l := j to n do S := S + s_l^2 od;
11         if S < ||b_n||_2^2 then
12           ρ := sqrt(||b_n||_2^2 − S);
13           if j > 1 then
14             j := j − 1;
15             k_j := ⌊−(ρ + s_j)/r_jj⌋;
16           else
17             if S < ||v||_2^2 then v := Bk fi;
18           fi;
19         else
20           k_j := 0;
21           j := j + 1;
22         fi;
```

23 **od**;
24 *output*(**v**);
25 **end**.

Inspection of this algorithm reveals that, for $n = 2$, Algorithm 5.2 is equivalent to Algorithm 5.1. It is also similar to an algorithm of Fincke & Pohst (1985), although that algorithm is stated in terms of minimising a quadratic form and, as a consequence which will be discussed in Section 6, it uses the Cholesky decomposition instead of the **QR** decomposition. We can state the following proposition.

Proposition 5.2. *If Algorithm 5.2 is executed on a basis matrix* $\mathbf{B}$ *of a lattice* Ω *of rank* n *in* $\mathbb{R}^m$ *then the algorithm terminates in a finite number of iterations and a non-zero vector* $\mathbf{v} \in \Omega$ *is output such that, for all non-zero* $\mathbf{w} \in \Omega$, $\|\mathbf{v}\|_2 \leqslant \|\mathbf{w}\|_2$.

Proof. To show that Algorithm 5.2 finds the shortest vector with respect to the Euclidean norm from a given lattice basis matrix, $\mathbf{B}$, we describe the operation of the algorithm. All references to $\mathbf{b}_j$ refer to the value of $\mathbf{B}$ after the *sortnorm* operation. The algorithm is essentially a series of nested **for** loops which cycles through a range of values for the k_j from $j = n$ (on the outermost level) down to $j = 1$ (on the innermost). The algorithm is not written out this way because the number of nested **for** loops is not known *a priori*.[1] In the innermost "loop," the algorithm tests lattice points constructed using the index vector $\mathbf{k}$ against the shortest lattice point yet found, $\mathbf{v}$, by testing on line 17 whether

$$\|\mathbf{Bk}\|_2^2 = \|\mathbf{QRk}\|_2^2 = \|\mathbf{Rk}\|_2^2 = S < \|\mathbf{v}\|_2^2 .$$

We will see that, for each test of this type, we have a value for $\mathbf{k}$ which ensures that, at the very least, $\|\mathbf{Bk}\|_2 < \|\mathbf{b}_n\|_2$. We will also see that every lattice point shorter than $\|\mathbf{b}_n\|_2$ is tested, up to symmetry. Specifically, we will see that the algorithm only tests those values of $\mathbf{k}$ for which $k_d > 0$ where d is the maximum index for which $k_d \neq 0$.

Let $\boldsymbol{\kappa}$ be the vector consisting of the last $n - j + 1$ elements of $\mathbf{k}$. That is,

$$\boldsymbol{\kappa} = (k_j, k_{j+1}, \ldots, k_n).$$

For a given j, consider the partitioning of $\mathbf{R}$ so that

$$\mathbf{R} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$$

where $\mathbf{X}$ is a $(j-1) \times (j-1)$ upper diagonal matrix with positive diagonal elements (and hence invertible), $\mathbf{Y}$ is a $(j-1) \times (n-j+1)$ matrix and $\mathbf{Z}$ is a $(n-j+1) \times$

[1] We could have used instead a recursive style to present the algorithm, but it is the author's opinion that, in this case, it would not have enhanced the clarity of the exposition.

$(n-j+1)$ upper diagonal matrix. Let $\mathbf{e} \in \mathbb{R}^n$ be given by

$$\mathbf{e} = \begin{pmatrix} \boldsymbol{\epsilon} \\ \mathbf{0} \end{pmatrix}.$$

where

$$\boldsymbol{\epsilon} = -\mathbf{X}^{-1}\mathbf{Y}\boldsymbol{\kappa}.$$

Then

$$\mathbf{R}(\mathbf{k}+\mathbf{e}) = \begin{pmatrix} \mathbf{X}\boldsymbol{\epsilon} + \mathbf{Y}\boldsymbol{\kappa} \\ \mathbf{Z}\boldsymbol{\kappa} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{Z}\boldsymbol{\kappa} \end{pmatrix}$$

It then becomes clear that

$$S = \|\mathbf{Z}\boldsymbol{\kappa}\|_2^2 = \|\mathbf{R}(\mathbf{k}+\mathbf{e})\|_2^2$$

and that, for any $\mathbf{e}' \in \mathbb{R}^n$ of the form

$$\mathbf{e}' = (\underbrace{e'_1, e'_2, \ldots, e'_{j-1}}_{\boldsymbol{\epsilon}'}, \underbrace{0, 0, \ldots, 0}_{n-j+1 \text{ times}}),$$

we have

$$\|\mathbf{B}(\mathbf{k}+\mathbf{e}+\mathbf{e}')\|_2^2 = \|\mathbf{R}(\mathbf{k}+\mathbf{e}+\mathbf{e}')\|_2^2 = \left\| \begin{pmatrix} \mathbf{X}\boldsymbol{\epsilon}' \\ \mathbf{Z}\boldsymbol{\kappa} \end{pmatrix} \right\|_2^2 = \|\mathbf{X}\boldsymbol{\epsilon}'\|_2^2 + \|\mathbf{Z}\boldsymbol{\kappa}\|_2^2 \geqslant S.$$

We conclude that if $S \geqslant \|\mathbf{b}_n\|_2^2$ (which is tested on line 11), then modifications to any of the k_l for $l < j$ will not produce a $\mathbf{k}$ for which $\|\mathbf{B}\mathbf{k}\|_2 < \|\mathbf{b}_n\|_2$. On the other hand, if $S \leqslant \|\mathbf{b}_n\|_2^2$ then modifications to the k_l with $l < j$ may produce an index vector $\mathbf{k}$ satisfying this criterion. The algorithm reflects this by incrementing j (thereby exiting the "loop" on that level) if the test on line 11 fails or by decrementing j (thereby entering the next nested "loop"), if it succeeds and if $j > 1$.

If the algorithm decrements j, the algorithm chooses a value for the new k_j such that if k_j were any less then the test on line 11 would fail on the next iteration. To see this, we observe that the value S on the next iteration, S', will be $S' = \|\mathbf{Z}'\boldsymbol{\kappa}'\|_2^2$ where

$$\mathbf{Z}' = \begin{pmatrix} r_{jj} & \boldsymbol{\eta} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\kappa}' = \begin{pmatrix} k_j \\ \boldsymbol{\kappa}. \end{pmatrix}$$

and $\boldsymbol{\eta} = (r_{j,j+1}, r_{j,j+2}, \ldots, r_{jn})$. Noting that $\boldsymbol{\eta}\boldsymbol{\kappa} = s_j$, we have

$$S' = \|\mathbf{Z}'\boldsymbol{\kappa}'\|_2^2 = \left\| \begin{pmatrix} r_{jj}k_j + \boldsymbol{\eta}\boldsymbol{\kappa} \\ \mathbf{Z}\boldsymbol{\kappa} \end{pmatrix} \right\|_2^2 = (r_{jj}k_j + s_j)^2 + S.$$

It is now clear that if the assignment for k_j, made on line 15, were to be replaced by an expression that made k_j smaller, then $S' \geqslant \|\mathbf{b}_n\|_2^2$. Since $k_j \geqslant (\rho - s_j)/r_{jj}$ also implies that $S' \geqslant \|\mathbf{b}_n\|_2^2$, we see that the number of iterations in each "loop" must be finite and so the algorithm must terminate after a finite number of iterations. $\square$

We could consider modifying Algorithm 5.2 to find the shortest vectors according to norms other than the Euclidean norm as we discussed for Algorithm 5.1. This would not pose a great problem because the underlying principle of the algorithm — to find all lattice points shorter than the smallest of the vectors in the input basis and output the shortest of these — is not very complex. However, for other norms, the **QR** decomposition which is employed extensively in Algorithm 5.2 may not be useful. Methods for numerical solutions may be required.

Another trivial modification to Algorithm 5.2 which could be considered is to have it output all non-zero lattice points with Euclidean norm less than some prescribed constant, c. This would merely involve replacing all occurrences of "$\|\mathbf{b}_n\|_2$" with "c" and modifying the location of the *output* statement.

Now, since Algorithm 5.2 tests every lattice point shorter than $\|\mathbf{b}_n\|_2$, up to symmetry, we expect that the running time of the algorithm will be at least proportional to $\|\mathbf{b}_n\|_2^n$ and this will be true also for generalisations of the algorithm in the ways described above. Thus the time required to find the shortest lattice point using this method is exponential in the rank of the lattice.

Can we perform this search substantially more quickly, say in an amount of time which is bounded above by a polynomial in the rank of the lattice? Unfortunately (depending on your point of view), the answer is probably no. VAN EMDE BOAS (1981) has shown that the problem of finding the shortest vector with respect to the sup-norm is an $\mathfrak{NP}$-complete problem. Thus, it is believed that the problem is probably computationally infeasible. We must resort to methods which find "almost shortest" or "sufficiently short" vectors in a reasonable amount of time. Such a method has been discovered by LENSTRA *et al.* (1982). Before discussing their algorithm, the so-called LLL algorithm, we introduce the topic of lattice reduction.

6. Lattice Reduction

We have already introduced the idea of reduction of quadratic forms in Section 1. We now consider reduction of lattices. In the most general sense, a reduced basis is a basis having some extremal property with respect to the set of all possible bases of that lattice. There are a number of definitions for a reduced basis. We will discuss reduction in the senses defined by GAUSS, MINKOWSKI, HERMITE, KORKIN & ZOLOTAREV and LOVÁSZ. In each case, the extremal property is related to the "shortness" of the vectors in the basis or to their orthogonality. Let us begin with reduction according to GAUSS.

6.1. Gaussian Reduction.

Definition and Algorithm for Point Lattices.

DEFINITION 6.1. A basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ of a lattice Ω of rank 2 is GAUSS-REDUCED if

$$\|\mathbf{b}_1\|_2 \leqslant \|\mathbf{b}_2\|_2 \qquad \text{and} \qquad |\mathbf{b}_1 \cdot \mathbf{b}_2| \leqslant \tfrac{1}{2}\|\mathbf{b}_1\|_2^2 . \tag{6.1}$$

As we will prove in Proposition 6.1, a Gauss-reduced basis can be obtained by application of the following algorithm.

ALGORITHM 6.1.

1 **begin**
2 **repeat**
3 *swap*$(\mathbf{b}_1, \mathbf{b}_2)$;
4 $\mathbf{b}_2 := \mathbf{b}_2 - \left\lfloor \dfrac{\mathbf{b}_1 \cdot \mathbf{b}_2}{\mathbf{b}_1 \cdot \mathbf{b}_1} \right\rceil \mathbf{b}_1$;
5 **while** $\|\mathbf{b}_1\|_2 > \|\mathbf{b}_2\|_2$;
6 *output*$(\mathbf{b}_1, \mathbf{b}_2)$;
7 **end**.

PROPOSITION 6.1. *If Algorithm 6.1 is executed on a basis* $\{\mathbf{b}_1, \mathbf{b}_2\}$ *of a lattice* Ω *of rank* 2 *then the algorithm will terminate after a finite number of iterations and will output a Gauss-reduced basis of* Ω.

PROOF. It is evident from inspection of Algorithm 6.1 that a basis of Ω is always maintained. After executing line 4, which we call a SIZE REDUCTION STEP, it is clear that the right-hand inequality of (6.1) will be satisfied. If the test on line 5 fails and the algorithm terminates then the left-hand inequality of (6.1) must also be satisfied and the algorithm outputs a Gauss-reduced basis of Ω. It remains to show that the algorithm terminates after a finite number of iterations.

Suppose the algorithm never terminates. Consider the state of the basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ just prior to the size reduction step on line 4 on some iteration through the main loop other than the initial iteration. Because the algorithm did not terminate on the previous iteration, we have

$$\|\mathbf{b}_1\|_2 < \|\mathbf{b}_2\|_2 \tag{6.2}$$

and

$$|\mathbf{b}_1 \cdot \mathbf{b}_2| < \tfrac{1}{2}\mathbf{b}_2 \cdot \mathbf{b}_2. \tag{6.3}$$

Let

$$r = \left\lfloor \frac{\mathbf{b}_1 \cdot \mathbf{b}_2}{\mathbf{b}_1 \cdot \mathbf{b}_1} \right\rceil.$$

Let $\mathbf{b}_2'$ be the value which is assigned to $\mathbf{b}_2$ after execution of the size reduction step. Thus,

$$\mathbf{b}_2' = \mathbf{b}_2 - r\mathbf{b}_1.$$

If $r = 0$ then $\mathbf{b}_2' = \mathbf{b}_2$ and the algorithm will terminate with a Gauss-reduced basis because of (6.2), contrary to our assumption. If $r = 1$ then

$$\mathbf{b}_1 \cdot \mathbf{b}_2 \geqslant \tfrac{1}{2}\mathbf{b}_1 \cdot \mathbf{b}_1$$

and so

$$\|\mathbf{b}_2'\|_2^2 = \mathbf{b}_2' \cdot \mathbf{b}_2' = \mathbf{b}_2 \cdot \mathbf{b}_2 - 2\mathbf{b}_1 \cdot \mathbf{b}_2 + \mathbf{b}_1 \cdot \mathbf{b}_1 > \mathbf{b}_1 \cdot \mathbf{b}_1 = \|\mathbf{b}_1\|_2^2$$

and the inequality arises because of (6.3). Therefore, the algorithm will terminate with a Gauss-reduced basis, contrary to our assumption. Through similar reasoning, we see that if $r = -1$ then the algorithm must terminate with a Gauss-reduced basis.

So, if the algorithm never terminates (as we have supposed for the sake of argument) then on each iteration apart from the first, we have $|r| \geqslant 2$ at the commencement of execution of the size reduction step. Now, this implies that

$$\frac{3}{2} \leqslant |r| - \tfrac{1}{2} \leqslant \frac{|\mathbf{b}_1 \cdot \mathbf{b}_2|}{\mathbf{b}_1 \cdot \mathbf{b}_1} \leqslant \frac{\|\mathbf{b}_2\|_2}{\|\mathbf{b}_1\|_2}$$

which implies that

$$\|\mathbf{b}_1\|_2 \leqslant \frac{2}{3} \|\mathbf{b}_2\|_2 .$$

Thus, unless $\|\mathbf{b}_2'\|_2 < \frac{2}{3} \|\mathbf{b}_2\|_2$, the algorithm will terminate. But this implies that the algorithm will produce lattice vectors of arbitrarily small length. Since the lattice contains a vector of smallest length, we have a contradiction. We conclude that the algorithm must terminate after a finite number of iterations. □

Proposition 6.1 leads directly to the following corollary.

COROLLARY 6.1. *For every lattice Ω of rank 2 there exists a Gauss-reduced basis.*

THEOREM 6.1. *If $\{\mathbf{b}_1, \mathbf{b}_2\}$ is a Gauss-reduced basis of a lattice Ω of rank 2 then*

$$\|\mathbf{v}\|_2 \geqslant \|\mathbf{b}_1\|_2 \tag{6.4}$$

for all $\mathbf{v} \in \Omega \setminus \{\mathbf{0}\}$ and

$$\|\mathbf{w}\|_2 \geqslant \|\mathbf{b}_2\|_2 \tag{6.5}$$

for all $\mathbf{w} \in \Omega \setminus \{a_1\mathbf{b}_1 \mid a_1 \in \mathbb{Z}\}$, that is, for all $\mathbf{w} \in \Omega$ which are linearly independent of $\mathbf{b}_1$.

PROOF. If $\mathbf{v} = k\mathbf{b}_1$, $k \in \mathbb{Z}$, $k \neq 0$, then $\|\mathbf{v}\|_2 \geqslant \|\mathbf{b}_1\|_2$. Suppose $\mathbf{w} \in \Omega$ is linearly independent of $\mathbf{b}_1$. Therefore, we can express $\mathbf{w}$ as

$$\mathbf{w} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2$$

with $a_1, a_2 \in \mathbb{Z}$ and $a_2 \neq 0$. Now,

$$\|\mathbf{w}\|_2^2 = \mathbf{w} \cdot \mathbf{w} = a_1^2\mathbf{b}_1 \cdot \mathbf{b}_1 + 2a_1a_2\mathbf{b}_1 \cdot \mathbf{b}_2 + a_2^2\mathbf{b}_2 \cdot \mathbf{b}_2.$$

If $|a_1| < |a_2|$ then

$$\|\mathbf{w}\|_2^2 \geqslant a_1^2\mathbf{b}_1 \cdot \mathbf{b}_1 - 2\big(a_2^2 - 1\big)|\mathbf{b}_1 \cdot \mathbf{b}_2| + a_2^2\mathbf{b}_2 \cdot \mathbf{b}_2 \geqslant a_1^2\mathbf{b}_1 \cdot \mathbf{b}_1 + \mathbf{b}_2 \cdot \mathbf{b}_2 \geqslant \|\mathbf{b}_2\|_2^2$$

through the application of (6.1). On the other hand, if $|a_1| \geqslant |a_2|$ then

$$\|\mathbf{w}\|_2^2 \geqslant a_1^2 \mathbf{b}_1 \cdot \mathbf{b}_1 - 2a_1^2 |\mathbf{b}_1 \cdot \mathbf{b}_2| + a_2^2 \mathbf{b}_2 \cdot \mathbf{b}_2 \geqslant a_2^2 \mathbf{b}_2 \cdot \mathbf{b}_2 \geqslant \|\mathbf{b}_2\|_2^2 .$$

□

Theorem 6.1 implies that a Gauss-reduced basis contains the shortest possible vectors in the sense defined by (6.4) and (6.5).

Relationship with Binary Quadratic Forms. Let us now briefly explore the connection between the Gaussian reduction of bases and the reduction of binary quadratic forms, the language in which the algorithm was originally expressed by Gauss (1801), art. 171. Let $\mathbf{B}$ be the $m \times 2$ matrix whose columns represent the basis vectors of a lattice Ω of rank 2 in $\mathbb{R}^m$. Then any lattice point $\mathbf{v} \in \Omega$ can be expressed

$$\mathbf{v} = \mathbf{B}\mathbf{k}$$

where $\mathbf{k} \in \mathbb{Z}^2$. The square Euclidean norm of $\mathbf{v}$ is

$$\|\mathbf{v}\|_2^2 = \mathbf{v}^T\mathbf{v} = \mathbf{k}^T\mathbf{B}^T\mathbf{B}\mathbf{k} = \mathbf{k}^T\mathbf{Q}\mathbf{k} \tag{6.6}$$

where $\mathbf{Q} = \mathbf{B}^T\mathbf{B}$ is a 2×2 symmetric, positive definite matrix. If we write

$$\mathbf{Q} = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \qquad \text{and} \qquad \mathbf{k} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{6.7}$$

then we can express that square Euclidean norm of $\mathbf{v}$ as a function of the indices x and y so that

$$\|\mathbf{v}\|_2^2 = Q(x, y) = ax^2 + 2bxy + cy^2.$$

If $a, b, c \in \mathbb{Z}$ also then $Q(x, y)$ is a positive binary quadratic form. Thus, finding a shortest vector in a lattice which is a subset of $\mathbb{Z}^2$ can be expressed as a problem of minimising a positive binary quadratic form. The opposite is also true since we can write any positive binary quadratic form such as (8) as a matrix equation such as (6.6) and (6.7). We can then use Cholesky decomposition (see Theorem 4.2) to decompose $\mathbf{Q}$ into $\mathbf{Q} = \mathbf{B}^T\mathbf{B}$ where $\mathbf{B}$ is a 2×2 upper diagonal matrix which can be regarded as consisting of the basis vectors of a lattice.

Furthermore, if we represent a binary quadratic form by a positive definite matrix in the way we have just described then it is easily confirmed that two such positive definite matrices $\mathbf{Q}$ and $\mathbf{Q}'$ represent equivalent (properly equivalent) positive binary quadratic forms if and only if

$$\mathbf{Q} = \mathbf{M}^T\mathbf{Q}'\mathbf{M}$$

where $\mathbf{M}$ is a unimodular matrix with $\det \mathbf{M} = \pm 1$ ($\det \mathbf{M} = 1$).

We will now see that reduction of binary quadratic forms is very nearly the same as Gaussian reduction of lattices. Suppose that we represent a positive binary quadratic form $Q'(x, y)$ by a positive definite matrix $\mathbf{Q}'$ and decompose $\mathbf{Q}'$ so that

$\mathbf{Q} = \mathbf{B'}^T\mathbf{B'}$ (the Cholesky decomposition of $\mathbf{Q'}$) and let $\mathbf{B'}$ represent the basis of a lattice. If we calculate a Gauss-reduced basis $\mathbf{B}$ of $\mathbf{B'}$ then

$$\mathbf{B} = \mathbf{B'M}$$

where $\mathbf{M}$ is a unimodular matrix with $\det \mathbf{M} = 1$ (which we can ensure by replacing one of the elements of the basis with its additive inverse, if necessary). Now let

$$\mathbf{Q} = \mathbf{B}^T\mathbf{B} = \mathbf{M}^T\mathbf{Q'M} = \begin{pmatrix} \mathbf{b}_1 \cdot \mathbf{b}_1 & \mathbf{b}_1 \cdot \mathbf{b}_2 \\ \mathbf{b}_1 \cdot \mathbf{b}_2 & \mathbf{b}_2 \cdot \mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

It is clear that a, b and c fulfill the conditions of (1.1) in Theorem 1.3, except the condition that $-a/2 < b$ where we have only $-a/2 \leqslant b$. Apart from this trivial difference (which can be easily overcome by the replacement of $\mathbf{b}_2$ by $\mathbf{b}_2 + \mathbf{b}_1$ if $-a/2 = b$), we see that the binary quadratic form represented by $\mathbf{Q}$ is a reduced form of $Q'(x, y)$.

Relationship with the Centred Continued Fraction. Let us now examine the relationship between Gaussian reduction and continued fractions. Consider some $\alpha \in \mathbb{C} \setminus \mathbb{Z}$ such that either $|\alpha^{-1}| < 1$ or $\Re\{\alpha^{-1}\} < 0$ or $\Re\{\alpha^{-1}\} > \frac{1}{2}$. We can write α as

$$\alpha = m_0 + \frac{\epsilon_0}{\xi_1}$$

where

$$m_0 = \lfloor \Re\{\alpha\} \rceil \qquad \text{and} \qquad \epsilon_0 = \operatorname{sgn}(\Re\{\alpha\} - \lfloor \Re\{\alpha\} \rceil).$$

The complex number ξ_1 then has the property that $0 \leqslant \Re\{\xi_1^{-1}\} \leqslant \frac{1}{2}$. Suppose $|\xi_1| > 1$, $\xi_1 \notin \mathbb{Z}$ and we repeat this expansion on ξ_1 so that we have

$$\alpha = m_0 + \cfrac{\epsilon_0}{m_1 + \cfrac{\epsilon_1}{\xi_2}}$$

where again $0 \leqslant \Re\{\xi_2^{-1}\} \leqslant \frac{1}{2}$. We can imagine continuing this fraction until, for some k, $|\xi_k| \leqslant 1$ or $\xi_k \in \mathbb{Z}$. Thus, we can expand a complex number α in this way to a (possibly infinitely) continued fraction of the form

$$\alpha = m_0 + \cfrac{\epsilon_0}{m_1 + \cfrac{\epsilon_1}{m_2 + \cfrac{\epsilon_2}{m_3 + \dots}}}$$

where the $\epsilon_i = \pm 1$, $m_i \in \mathbb{Z}$ and $m_i \geqslant 0$ when $i > 0$. We call an expansion of this type a CENTRED CONTINUED FRACTION expansion of α. The expansion so defined is clearly also unique according to the procedural definitions we have given, up to the choice of the nearest integer function $\lfloor \cdot \rceil$. We now formalise these procedures into an algorithm.

ALGORITHM 6.2.

1 **begin**
2 $\eta_{-1} := -1;\ \eta_{-2} = \alpha;$
3 $p_{-1} := 1;\ p_{-2} := 0;$
4 $q_{-1} := 0;\ q_{-2} := 1;$
5 $\xi_0 := \alpha;$
6 $n := 0;$
7 **while** $\xi_n \notin \mathbb{Z} \wedge \left(|\xi_n^{-1}| < 1 \vee \Re\{\xi_n^{-1}\} < 0 \vee \Re\{\xi_n^{-1}\} > \frac{1}{2}\right)$ **do**
8 $m_n := \lfloor \Re\{\xi_n\} \rceil;$
9 $\epsilon_n := \operatorname{sgn}(\Re\{\xi_n\} - m_n);$
10 $\eta_n := \epsilon_n(\eta_{n-2} + m_n\eta_{n-1});$
11 $n := n + 1;$
12 $\xi_n := \dfrac{-\eta_{n-2}}{\eta_{n-1}};$
13 **od**;
14 **end**.

It is not surprising that this algorithm bears a strong resemblance to Algorithm 3.2 of Chapter 2 for computing the s.c.f. expansion of a real number. To show that it calculates the centred continued fraction expansion it remains only to show that $\xi_n = -\eta_{n-2}/\eta_{n-1}$ corresponds to the definition implied in the preceding discussion. We show this by induction. If we label $\xi_0 = \alpha$ then it is true for $n = 0$. Suppose it is true for all $0 \leqslant n < N$. We have

$$\begin{aligned}\xi_N &= \frac{\epsilon_{N-1}}{\xi_{N-1} - m_{N-1}} \\ &= \frac{-\epsilon_{N-1}\eta_{N-2}}{\eta_{N-3} + m_{N-1}\eta_{N-2}} \\ &= \frac{-\eta_{N-2}}{\eta_{N-1}}\end{aligned}$$

and so it is true for $n = N$ also.

To see the similarity of the centred continued fraction algorithm of Algorithm 6.2 with the Gaussian reduction algorithm of Algorithm 6.1, consider the following characterisation of a basis $\{\mathbf{b}_1^*, \mathbf{b}_2^*\}$, represented by the matrix $\mathbf{B}^*$, of a lattice Ω^* of rank 2 in $\mathbb{R}^2$ by a complex number α. Let $\mathbf{T}$ be a linear transformation consisting of a rotation with scaling such that $\mathbf{T}\mathbf{b}_1^* = (-1, 0)^T$. Let $\mathbf{B} = \mathbf{T}\mathbf{B}^*$ be a basis of a new lattice Ω in $\mathbb{R}^2$. A Gauss-reduced basis of Ω will correspond to a Gauss-reduced basis of Ω^* through the application of the inverse transformation, $\mathbf{T}^{-1}$. Now, set $\eta_{-1} = -1 = b_{11} + ib_{21}$ and $\eta_{-2} = \alpha = b_{21} + ib_{22}$. The pair $\{1, \alpha\}$ defined in this way can be regarded as the basis of a lattice in $\mathbb{C}$. From inspection of Algorithm 6.2, we see that the pair $\{\eta_n, \eta_{n-1}\}$ always maintains a basis. Bearing in mind the obvious relationships of the magnitude of the ratio $|\eta_{n-2}/\eta_{n-1}|$ with the ratio of the Euclidean norms of their equivalent lattice points and of $\Re\{\eta_{n-2}/\eta_{n-1}\}$ with the

scalar product, we see that the two algorithms, Algorithm 6.1 and Algorithm 6.2, are essentially the same. A point of difference arises in that Algorithm 6.2 negates the "new lattice point" η_n if its "angle" with η_{n-1} would otherwise be acute through the use of the ϵ_n. However, this is hardly a substantial difference. Because of the similarity with Gaussian reduction, we can state the following theorem, which is really a corollary of Proposition 6.1.

THEOREM 6.2. *If $\alpha \in \mathbb{C}$ and $\Im\{\alpha\} \neq 0$ then the centred continued fraction expansion of α is finite.*

Complexity of Algorithm 6.1. We have already witnessed in the proof of Proposition 6.1 that, on each iteration of Algorithm 6.1 apart from the first on which the algorithm does not terminate, the Euclidean norm of the basis vector $\mathbf{b}_2$ is reduced by at least one third. To bound the number of iterations required by the algorithm, we introduce the inertia of a lattice basis.

DEFINITION 6.2. The INERTIA of a lattice basis is the sum of the squared Euclidean norms of the basis vectors.

THEOREM 6.3. *If I^* is the inertia of a Gauss-reduced basis of a lattice Ω and I is the inertia of any other basis of Ω then $I^* \leqslant I$.*

PROOF. The proof follows immediately from Theorem 6.1. □

Consider the number of iterations required in terms of inertia using only our arguments from the proof of Proposition 6.1. If I_n is the inertia of the basis on the n^{th} iteration after the size reduction step and I_0 is the inertia of the input basis then

$$I_n \leqslant \left(\frac{2}{1+\left(\frac{3}{2}\right)^2}\right)^{n-1} I_0 = \left(\frac{8}{13}\right)^{n-1} I_0.$$

If the inertia of the Gauss-reduced basis is I^* then the number of iterations, N, required by Algorithm 6.1 to produce a Gauss-reduced basis is bound above by

$$N \leqslant \log_{\frac{13}{8}} \frac{I_0}{I^*} + 1. \tag{6.8}$$

LAGARIAS (1980) was the first to discover a logarithmic upper bound of this type on the running time of algorithms for Gaussian reduction. VALLÉE (1991) considered the worst-case running time of a Gaussian reduction algorithm for lattices which are subsets of $\mathbb{Z}^2$. For these lattices, $I^* \geqslant 2$. The maximum number of iterations in this case can therefore be stated without reference to I^*. She found that

$$N \leqslant \tfrac{1}{2}\left[\log_{1+\sqrt{2}}\left(\frac{2\sqrt{2}}{3} I_0\right) + 1\right] \tag{6.9}$$

and that this bound is the best possible of its type.[2] Comparing the leading terms of (6.8) to (6.9) for lattices in $\mathbb{Z}^2$ we see that the former gives

$$N \leqslant 2.06 \log I_0 + O(1)$$

whereas the latter gives

$$N \leqslant 0.568 \log I_0 + O(1),$$

which is clearly a substantial improvement.

More recently, DAUDÉ *et al.* (1994) carried out, amongst other things, an analysis of the average-case complexity of Gauss' algorithm. They found that the number of iterations required by the algorithm obeys a geometric law and is thus $O(1)$ "on average" where this term is defined in an appropriate sense.

Generalisations. If we had instead defined a Gauss-reduced basis in terms of the properties stated in Theorem 6.1 then we can readily generalise the definition to other norms. To generalise Gauss' algorithm, such as we presented in Algorithm 6.1, we (naturally) replace all references to the Euclidean norm with the desired norm and replace the size reduction step on line 4 with

$$\mathbf{b}_2 := \mathbf{b}_2 - \mathbf{b}_1 \arg\min_{k \in \mathbb{Z}} \{\|\mathbf{b}_2 - k\mathbf{b}_1\|\}.$$

KAIB (1994) has considered such generalisations and has obtained sharp upper bounds on the number of iterations required.

GAUSS (1801), art. 272–275, also defines a reduced form for TERNARY QUADRATIC FORMS, that is, quadratic forms in three variables, and describes a procedure for reducing such forms (corresponding to lattices of rank 3). However, it does not suit our purposes to discuss it further, save to mention that LAGARIAS (1980) has analysed the computational complexity of the algorithm.

6.2. Minkowski Reduction. MINKOWSKI invented the notion of the successive minima of a lattice.

DEFINITION 6.3. The SUCCESSIVE MINIMA $\lambda_1, \lambda_2, \ldots, \lambda_n$ of a lattice Ω of rank n in $\mathbb{R}^m$ with respect to a norm $\|\cdot\|$ are the least real numbers for which it is true that, for each $1 \leqslant k \leqslant n$, there exists a set of k linearly independent lattice vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ such that

$$\max\{\|\mathbf{v}_1\|, \|\mathbf{v}_2\|, \ldots, \|\mathbf{v}_k\|\} = \lambda_k.$$

We now discuss a geometric interpretation of this definition. Consider the centrally symmetric convex body consisting of all points $\mathbf{x} \in \mathbb{R}^m$ with $\|\mathbf{x}\| < \lambda$. As λ is increased from 0, we know from Theorem 3.2 that there exists some $\lambda_1 \leqslant 2^r\Delta$ for

[2]The algorithm presented by VALLÉE (1991) differs from Algorithm 6.1 in that it only performs the *swap* operation on line 3 if $\|\mathbf{b}_1\| > \|\mathbf{b}_2\|$. Therefore, the *swap* operation might not be performed on the first iteration and so the upper bound of (6.9) should perhaps be increased by 1 in reference to Algorithm 6.1.

which a lattice point of Ω other than the origin, say $\mathbf{v}_1$, lies on the surface of the body. Obviously, $\mathbf{v}_1$ is the shortest vector in the lattice with respect to $\|\cdot\|$. This is the first successive minimum. As λ is further increased, a value λ_2 is reached at which point a non-zero lattice point, say $\mathbf{v}_2$, linearly independent of $\mathbf{v}_1$, is encountered. λ_2 is the second successive minima. We can continue to increase λ in this way until all the successive minima and associated MINIMAL VECTORS $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are found. The minimal vectors are not unique, since any of them can be replaced by their additive inverses.

The following theorem regarding successive minima is also due to MINKOWSKI, and is sometimes called his Second Theorem.

THEOREM 6.4. *Let* $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| < 1\}$ *and let* Ω *be a lattice of rank* n *in* $\mathbb{R}^n$. *If* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *denote the successive minima of* Ω *then*

$$\lambda_1 \lambda_2 \cdots \lambda_n \operatorname{vol} \mathcal{S} \leqslant 2^n.$$

Now, the minimal vectors of Ω, although linearly independent, may not form a basis. MINKOWSKI proposed a definition of a reduced basis which is analogous to that of the minimal vectors.

DEFINITION 6.4. A basis $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ of a lattice Ω of rank n in $\mathbb{R}^m$ is MINKOWSKI-REDUCED with respect to a norm $\|\cdot\|$ if

$$\|\mathbf{v}\| \geqslant \|\mathbf{b}_i\|$$

for all $\mathbf{v} \in \Omega$ which are linearly independent of $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{i-1}\}$ when $2 \leqslant i \leqslant n$ or for all non-zero $\mathbf{v} \in \Omega$ when $i = 1$.

This is the same property which we proved that a Gauss-reduced basis enjoys in Theorem 6.1. Thus, we can think of Minkowski reduction as a natural generalisation of Gaussian reduction to lattices of rank greater than 2.

In an example attributed to H. W. LENSTRA, JR., LAGARIAS (1994) describes a lattice of rank 14 with the property that, in constructing a Minkowski-reduced basis, a change in the ordering of a number of vectors in the basis which have equal length can make a difference to the choice of subsequent basis vectors and hence their length. Therefore, he proposes the idea of a LEXICOGRAPHICALLY MINKOWSKI-REDUCED BASIS which is that Minkowski-reduced basis which is lexicographically least according to the usual lexicographic ordering of the norms of the ordered Minkowski-reduced vectors $(\|\mathbf{b}_1\|, \|\mathbf{b}_2\|, \ldots, \|\mathbf{b}_n\|)$.

DEFINITION 6.5. The USUAL LEXICOGRAPHIC ORDERING of $\mathbb{R}^n$ is that for which

$$\mathbf{u} = (u_1, u_2, \ldots, u_n) < \mathbf{v} = (v_1, v_2, \ldots, v_n)$$

if there exists some k, $1 \leqslant k \leqslant n$, such that $u_k < v_k$ and $u_i = v_i$ for $i = 1, 2, \ldots, k$.

From an algorithmic point of view, it is probably computationally infeasible to calculate a (lexicographically) Minkowski-reduced basis for a given input basis, since it involves finding the shortest vector in the lattice.

6.3. Hermite Reduction. We briefly mention a notion of reduction due to HERMITE (1850). Although, it is not a very "strong" form of reduction, in that there are no bounds on the lengths of vectors in the reduced basis with respect to the shortest possible, it does ensure a degree of orthogonality. More importantly, it underlies the definitions of reduction which we will discuss afterwards.

DEFINITION 6.6. Let $\mathbf{B} = \mathbf{QR}$ be the $\mathbf{QR}$ decomposition of the $m \times n$ matrix $\mathbf{B}$, the columns of which represent the basis vectors of a lattice Ω of rank n in $\mathbb{R}^m$. The basis represented by $\mathbf{B}$ is HERMITE-REDUCED if

$$|r_{ij}| \leqslant \tfrac{1}{2} r_{ii} \tag{6.10}$$

for all $1 \leqslant i < j \leqslant n$.

The following simple algorithm effects Hermite reduction.

ALGORITHM 6.3.

```
1  begin
2    QRdecompose(B, Q, R);
3    for j := 1 to n do
4        for i := j − 1 to 1 step − 1 do
5            k := ⌊ r_ij / r_ii ⌋;
6            b_j := b_j − k b_i;
7            r_j := r_j − k r_i;
8        od;
9    od;
10   output(B);
11 end.
```

PROPOSITION 6.2. *If Algorithm 6.3 is executed on an $m \times n$ matrix $\mathbf{B}$, the columns of which represent the basis vectors of a lattice Ω of rank n in $\mathbb{R}^m$, then, after at most $n(n-1)/2$ iterations through the inner loop on lines 4–8, the algorithm terminates and outputs an Hermite-reduced basis.*

PROOF. The proof is by inspection of the algorithm. □

Observe that each iteration through the inner loop of Algorithm 6.3 on lines 4-8 involves operations on vectors of m elements. The total number of arithmetic operations is therefore $O(mn^2)$.

If, for Algorithm 6.3, we label the input $\mathbf{B}$ and the output $\mathbf{B}'$ then

$$\left\|\mathbf{b}'_j\right\|_2 \leqslant \|\mathbf{b}_j\|_2 \tag{6.11}$$

for all $j = 1, 2, \ldots, n$. This is because $\|\mathbf{b}_j\|_2 = \|\mathbf{r}_j\|_2$ and because r_{ij} is left unchanged by the algorithm if $i \geqslant j$ or if $|r_{ij}| < \frac{1}{2}|r_{ii}|$, but otherwise reduced to ensure (6.10). It is because of (6.11) that many authors refer to this operation as SIZE REDUCTION (as we did ourselves for the operation on line 4 of Algorithm 6.1).

6.4. Korkin-Zolotarev Reduction. The last of the classical (that is, pre-20$^{\text{th}}$ century) notions of reduction we will discuss is that of KORKIN and ZOLOTAREV. While a Minkowski-reduced basis consists of the shortest possible vectors and is easily generalised to all norms, the definition of a Korkin-Zolotarev-reduced basis, which we are about to give, is more concerned with the orthogonality of the constituent vectors and more closely tied to the Euclidean norm. The original definition is recursive and expressed in the language of quadratic forms. The definition we give here is an equivalent non-recursive definition, adapted from LAGARIAS *et al.* (1990).

DEFINITION 6.7. Let $\mathbf{B} = \mathbf{QR}$ be the $\mathbf{QR}$ decomposition of the $m \times n$ matrix $\mathbf{B}$, the columns of which represent the basis vectors of a lattice Ω of rank n in $\mathbb{R}^m$. The basis represented by $\mathbf{B}$ is KORKIN-ZOLOTAREV-REDUCED if it is Hermite-reduced and, for all $\mathbf{v} \in \Omega$ expressed as $\mathbf{v} = \mathbf{Bk}$, $\mathbf{k} \in \mathbb{Z}^n$, for which $k_j \neq 0$, $1 \leqslant j \leqslant n$, it is true that

$$\|\boldsymbol{\rho}\|_2 \geqslant r_{jj} \tag{6.12}$$

where $\boldsymbol{\rho}$ is the vector formed from the last $n - j + 1$ elements of $\mathbf{Rk}$.

We now give a geometric interpretation to this definition. The condition (6.12) means that a given vector $\mathbf{b}_j$ in the ordered Korkin-Zolotarev-reduced basis must belong to the set of shortest vectors which are linearly independent of the preceding vectors $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1})$ when that part of the vector is measured which lies in the orthogonal subspace to the span of those basis vectors. We say $\mathbf{b}_j$ "belongs to the set of shortest vectors" because the addition of any integer linear combination of the preceding basis vectors will not change the length of that part of it which lies in the orthogonal subspace. In particular, this means that $\mathbf{b}_1$ is the shortest lattice vector in $\mathbb{R}^n$, the part of $\mathbf{b}_2$ which is orthogonal to $\mathbf{b}_1$ is the shortest of all lattice points which are linearly independent of $\mathbf{b}_1$ and so on.

The condition that the reduced basis should also be Hermite-reduced is intended to ensure that the basis vector $\mathbf{b}_j$ should be as orthogonal as possible with respect to the preceding vectors $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1})$ in the ordered Korkin-Zolotarev-reduced basis.

From an algorithmic point of view, a Korkin-Zolotarev-reduced basis is probably computationally infeasible to compute from an arbitrary given basis, just as a Minkowski-reduced basis is, because it involves finding the shortest vector in the lattice.

6.5. Lovász Reduction. A major breakthrough in the computational complexity of finding sufficiently short vectors was achieved with the discovery by Lovász of a new notion of reduction and announced, along with an algorithm for constructing a reduced basis, in Lenstra *et al.* (1982).

Definition 6.8. Let $\mathbf{B} = \mathbf{QR}$ be the $\mathbf{QR}$ decomposition of the $m \times n$ matrix $\mathbf{B}$, the columns of which represent the ordered basis vectors of a lattice Ω of rank n in $\mathbb{R}^m$. The basis represented by $\mathbf{B}$ is Lovász-reduced if it is Hermite-reduced and

$$r_{j,j}^2 \leqslant 2r_{j+1,j+1}^2 \tag{6.13}$$

for all $1 \leqslant j < n$.

The condition (6.13) is the so-called Siegel reduction condition. It is not the condition that was used in the original paper of Lenstra *et al.* (1982). There, the condition used in place of (6.13) was that

$$r_{j,j}^2 \left(\frac{3}{4} - \frac{r_{j,j+1}^2}{r_{j+1,j+1}^2} \right) \leqslant r_{j+1,j+1}^2$$

for all $1 \leqslant j < n$. These conditions are equivalent in that all of the important properties (such as we prove in Theorem 6.5) of the reduced bases are shared. However, we prefer the Siegel reduction condition because of its simplicity.

At first glance, it is not clear why the reduction condition of (6.13) should ensure that the reduced basis will consist of short vectors. The following theorem and proof, adapted directly from Lenstra *et al.* (1982), shows that it does.

Theorem 6.5. *Let $\mathbf{B} = \mathbf{QR}$ be the $\mathbf{QR}$ decomposition of the $m \times n$ matrix $\mathbf{B}$, the columns of which represent the ordered basis vectors of a lattice Ω of rank n in $\mathbb{R}^m$. If the basis represented by $\mathbf{B}$ is Lovász-reduced then*

$$\|\mathbf{b}_i\|_2^2 \leqslant 2^{j-1} r_{jj}^2 \tag{6.14}$$

for all i, j such that $1 \leqslant i \leqslant j \leqslant n$ and

$$\|\mathbf{b}_1\|_2^2 \leqslant 2^{n-1} \|\mathbf{v}\|_2^2 \tag{6.15}$$

for all non-zero $\mathbf{v} \in \Omega$ and, more generally,

$$\|\mathbf{b}_j\|_2^2 \leqslant 2^{n-1} \max\left\{ \|\mathbf{v}_1\|_2^2, \|\mathbf{v}_2\|_2^2, \ldots, \|\mathbf{v}_j\|_2^2 \right\} \tag{6.16}$$

whenever $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_j$ are linearly independent lattice vectors and $1 \leqslant j \leqslant n$.

Proof. From (6.13), we see at once that

$$r_{ii}^2 \leqslant 2^{j-i} r_{jj}^2$$

for all i, j such that $1 \leqslant i \leqslant j \leqslant n$. Now, it follows that

$$\begin{aligned}\|\mathbf{b}_j\|_2^2 &= \sum_{i=1}^{j} r_{ij}^2 \\ &\leqslant r_{jj}^2 + \frac{1}{4}\sum_{i=1}^{j-1} r_{ii}^2 \\ &\leqslant r_{jj}^2 + \frac{1}{4} r_{jj}^2 \sum_{i=1}^{j-1} 2^{j-i} \\ &\leqslant 2^{j-1} r_{jj}^2\end{aligned}$$

and so

$$\|\mathbf{b}_i\|_2^2 \leqslant 2^{i-1} r_{ii}^2 \leqslant 2^{j-1} r_{jj}^2.$$

To prove the truth of (6.15) consider the expression of the lattice point $\mathbf{v}$ as $\mathbf{v} = \mathbf{Bk}$ where $\mathbf{k} \in \mathbb{Z}^n$. Let $\mathbf{s} = \mathbf{Rk}$. Now,

$$\|\mathbf{v}\|_2^2 = \sum_{i=1}^{n} s_i^2.$$

Also, if j is the largest index such that $k_j \neq 0$ then it follows from the upper triangular nature of $\mathbf{R}$ that $s_j = k_j r_{jj}$. Therefore

$$\|\mathbf{v}\|_2^2 \geqslant k_j^2 r_{jj}^2 \geqslant r_{jj}^2$$

and so, applying (6.14), we have

$$\|\mathbf{b}_1\|_2^2 \leqslant 2^{j-1} r_{jj}^2 \leqslant 2^{n-1} r_{jj}^2 \leqslant 2^{n-1} \|\mathbf{v}\|_2^2.$$

Now consider (6.16). Let $\mathbf{V}$ be an $m \times j$ matrix, the columns of which are linearly independent lattice vectors $\mathbf{v}_i$, $i = 1, 2, \ldots, j$. Therefore, we can write $\mathbf{V} = \mathbf{BK}$ where $\mathbf{K}$ is an $n \times j$ integer matrix with full column rank. Let t be the maximum index such that there exists some index u such that $k_{tu} \neq 0$. Clearly, $t \geqslant j$, otherwise $\mathbf{K}$ would not have full column rank. Let $\mathbf{S} = \mathbf{RK}$. Now,

$$\|\mathbf{v}_u\|_2^2 = \sum_{i=1}^{t} s_{iu}^2$$

and $s_{tu} = k_{tu} r_{tt}$ because of the upper triangular nature of $\mathbf{R}$. Therefore,

$$\|\mathbf{v}_u\|_2^2 \geqslant k_{tu}^2 r_{tt}^2 \geqslant r_{tt}^2$$

and so, applying (6.14), we have

$$\|\mathbf{b}_j\|_2^2 \leqslant 2^{t-1} r_{tt}^2 \leqslant 2^{n-1} r_{tt}^2 \leqslant 2^{n-1} \|\mathbf{v}_u\|_2^2$$

and (6.16) follows. □

We have shown in Theorem 6.5 that there is an upper bound to the ratio of the first basis vector in a Lovász-reduced basis and the smallest non-zero vector in the lattice. This ratio is unfortunately exponential in the rank of the lattice. However, given that the LLL algorithm produces a Lovász-reduced basis in a number of iterations which is bounded above by a polynomial in the rank of the lattice, we may be prepared to accept this gap between what is feasibly computable and the best possible.

7. The LLL Algorithm

We now present a version of the LLL algorithm, expressed in the "language" of **QR** decomposition whereas the original in LENSTRA *et al.* (1982) used a notation based on an unnormalised Gram-Schmidt orthogonalisation procedure. The algorithm presented here also makes use of the Siegel reduction condition and does not make use of intermediate size reduction steps (we will discuss the implications of this later). Nevertheless, the theoretical results are adapted directly from the original paper.

The algorithm below makes use of the procedures *swap* and *QRdecompose* for swapping values and for **QR** decomposition, which we have used previously, and a procedure *HermiteReduce*, which performs Hermite reduction (size reduction) on its argument in the way described by Algorithm 6.3.

ALGORITHM 7.1.

1 **begin**
2 $QRdecompose(\mathbf{B}, \mathbf{Q}, \mathbf{R})$;
3 $j := 1$;
4 **while** $j < n$ **do**
5 **if** $r_{j,j}^2 > 2r_{j+1,j+1}^2$ **then**
6 $k := \left\lfloor \dfrac{r_{j,j+1}}{r_{j,j}} \right\rceil$;
7 $\mathbf{b}_{j+1} := \mathbf{b}_{j+1} - k\mathbf{b}_j$;
8 $\mathbf{r}_{j+1} := \mathbf{r}_{j+1} - k\mathbf{r}_j$;
9 $swap(\mathbf{b}_j, \mathbf{b}_{j+1})$;
10 $QRdecompose(\mathbf{B}, \mathbf{Q}, \mathbf{R})$;
11 **if** $j > 1$ **then** $j := j - 1$; **fi**;
12 **else**
13 $j := j + 1$;
14 **fi**;
15 **od**;
16 $HermiteReduce(\mathbf{B})$;
17 $output(\mathbf{B})$;
18 **end**.

Without even fully understanding the properties of the algorithm, we can see that it appears somewhat inefficient to recalculate the whole **QR** decomposition of **B** just because we have swapped the order of two of the basis vectors, as we do on line 10. It is also obvious that we are making no use of the **Q** matrix. Therefore, we only need to update **R** and it is a simple matter to confirm that this can be achieved by applying the transformation

$$\mathbf{R}' = \mathbf{HGRH} \tag{7.1}$$

where $\mathbf{R}'$ represents the updated value of **R** and

$$\mathbf{G} = \begin{pmatrix} \mathbf{I}_{j-1} & & \mathbf{0} \\ & \mathbf{G}' & \\ \mathbf{0} & & \mathbf{I}_{r-j-1} \end{pmatrix} \qquad \text{and} \qquad \mathbf{H} = \begin{pmatrix} \mathbf{I}_{j-1} & & \mathbf{0} \\ & \mathbf{H}' & \\ \mathbf{0} & & \mathbf{I}_{r-j-1} \end{pmatrix}$$

and

$$\mathbf{G}' = \frac{1}{\sqrt{r_{j,j+1}^2 + r_{j+1,j+1}^2}} \begin{pmatrix} r_{j+1,j+1} & -r_{j,j+1} \\ r_{j,j+1} & r_{j+1,j+1} \end{pmatrix} \qquad \text{and} \qquad \mathbf{H}' = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The pre- and post-multiplication by **H** serves to swap the j^{th} and $(j+1)^{\text{th}}$ rows and columns of **GR**. The transformation **G** is a Givens rotation. A Givens rotation requires much less computational effort than a full **QR** decomposition (in this case, only the j^{th} and $(j+1)^{\text{th}}$ rows will be affected) and it can be implemented in a numerically stable way (Golub & van Loan, 1989). This discussion has prepared us for the following proposition.

Proposition 7.1. *If Algorithm 7.1 is executed on an $m \times n$ basis matrix $\mathbf{B}$ of a lattice Ω of rank n in $\mathbb{R}^m$, and, on some iteration, the test on line 5 succeeds then, at line 10 on the same iteration,*

$$r'_{j,j}{}^2 \leqslant \frac{3}{4} r_{j,j}^2, \qquad r'_{j+1,j+1}{}^2 \leqslant r_{j,j}^2 \qquad \textit{and} \qquad r'_{j,j} r'_{j+1,j+1} = r_{j,j} r_{j+1,j+1}$$

where $r'_{j,j}$ and $r'_{j+1,j+1}$ are the new values of $r_{j,j}$ and $r_{j+1,j+1}$ after the recalculation of the **QR** *decomposition (or Givens rotation).*

Proof. If the test on line 5 succeeds then $r_{j,j}^2 > 2r_{j+1,j+1}^2$. The execution of lines 6–8 ensures that $|r_{j,j+1}| \leqslant \frac{1}{2} r_{j,j}$. From consideration of (7.1) we find that

$$r'_{j,j}{}^2 = r_{j,j+1}^2 + r_{j+1,j+1}^2 \leqslant \frac{3}{4} r_{j,j}^2,$$

$$r'_{j+1,j+1}{}^2 = \frac{r_{j,j}^2 r_{j+1,j+1}^2}{r_{j,j+1}^2 + r_{j+1,j+1}^2} \leqslant r_{j,j}^2$$

and that

$$r'_{j,j} r'_{j+1,j+1} = \sqrt{r_{j,j+1}^2 + r_{j+1,j+1}^2}\, \frac{r_{j,j} r_{j+1,j+1}}{\sqrt{r_{j,j+1}^2 + r_{j+1,j+1}^2}} = r_{j,j} r_{j+1,j+1}.$$

□

Before proving the main result of this section, we require the following widely known facts from the geometry of numbers.

THEOREM 7.1. *If* $\mathbf{B} = \mathbf{QR}$ *is the* **QR** *decomposition of the* $m \times n$ *basis matrix* $\mathbf{B}$ *of a lattice* Ω *of rank* n *in* $\mathbb{R}^m$ *then* $\det \mathbf{R} = c$ *where* c *is constant which depends only on* Ω.

PROOF. Let $\mathbf{B}$ and $\mathbf{B}'$ be two matrices representing bases of Ω. Therefore $\mathbf{B}' = \mathbf{BM} = \mathbf{QRM}$, where $\mathbf{M}$ is a unimodular matrix, so $\det(\mathbf{RM}) = \det \mathbf{R}$. If the **QR** decomposition of $\mathbf{RM}$ is $\mathbf{PR}'$ then $\det \mathbf{R}' = \det \mathbf{R}$ since $\mathbf{P}$ is orthogonal. Now, $\mathbf{QP}$ is column orthogonal so $(\mathbf{QP})\mathbf{R}'$ is the **QR** decomposition of $\mathbf{B}'$. □

THEOREM 7.2. *If* $\mathbf{B} = \mathbf{QR}$ *is the* **QR** *decomposition of the* $m \times n$ *matrix* $\mathbf{B}$*, the columns of which represent the basis vectors of a lattice* Ω *of rank* n *in* $\mathbb{R}^m$ *then there exists a non-zero* $\mathbf{v} \in \Omega$ *such that*

$$\|\mathbf{v}\|_2 \leqslant \gamma_n \sqrt[n]{\det \mathbf{R}} \tag{7.2}$$

where $\gamma_n \in \mathbb{R}$ *is a positive constant which depends only on* n.

PROOF. Consider the parameterised set

$$\mathcal{S}(\mu) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 < \mu\}.$$

This is a hypersphere of radius μ centered about the origin. Recall that $\operatorname{vol} \mathcal{S}(\mu) = J_n \mu^n$ where

$$J_n = \frac{\pi^{\frac{1}{2}n}}{\Gamma\left(\frac{1}{2}n + 1\right)}$$

(see, for example SIEGEL, 1989, p. 26).

Consider the lattice Ω' of rank n in $\mathbb{R}^n$ constructed from the column vectors of $\mathbf{R}$. Every lattice point $\mathbf{v} \in \Omega$ can be expressed as $\mathbf{v} = \mathbf{Bk}$ where $\mathbf{k} \in \mathbb{Z}^n$ and there is an equivalent point $\mathbf{v}' = \mathbf{Rk} \in \Omega'$ such that $\|\mathbf{v}'\|_2 = \|\mathbf{v}\|_2$. Since $\mathcal{S}(\mu)$ is a convex body we can use Minkowski's First Theorem (Theorem 3.2) to show that if

$$\mu > 2 \sqrt[n]{\frac{\det \mathbf{R}}{J_n}}$$

then there is a non-zero lattice point $\mathbf{v}' \in \Omega'$ in $\mathcal{S}(\mu)$. Thus, setting

$$\gamma_n = \frac{2}{\sqrt[n]{J_n}}$$

we achieve the desired result. □

REMARK 7.1. The smallest possible values for γ_n in (7.2) are known as the HERMITE CONSTANTS. Their values are known only for $2 \leqslant r \leqslant 8$.

The following theorem places an upper bound on the Hermite constants.

THEOREM 7.3. *Theorem 7.2 is still true when γ_n is replaced by $(4/3)^{\frac{1}{2}(n-1)}$ in (7.2).*

PROOF. See CASSELS (1971). □

We can now state and prove the main result for this section.

PROPOSITION 7.2. *Let $\mathbf{B} = \mathbf{QR}$ be the $\mathbf{QR}$ decomposition of the $m \times n$ basis matrix $\mathbf{B}$ of a lattice Ω of rank n in $\mathbb{R}^m$. If Algorithm 7.1 is executed on $\mathbf{B}$ then, after a finite number of iterations through the main loop of lines 4–15, the algorithm terminates with a Lovász-reduced basis as its output.*

The number of iterations required is $O(n^2(n + \log(M/L)))$, where

$$M = \max\{\|\mathbf{b}_1\|_2, \|\mathbf{b}_2\|_2, \ldots, \|\mathbf{b}_n\|_2\},$$

and L is the Euclidean length of the shortest vector in Ω.

PROOF. By inspection of the algorithm, we can confirm that if the algorithm terminates, it produces a Lovász-reduced basis. It remains to show that the algorithm terminates within the number of iterations specified.

Let d_i represent the value at line 5 of the determinant of the $i \times i$ submatrix of $\mathbf{R}$ consisting of its first i rows and columns. Since this submatrix is upper triangular,

$$d_i = r_{11} r_{22} \cdots r_{ii}.$$

Consider the sublattice of Ω which is generated by the first i columns of $\mathbf{B}$. Let $\mathbf{v}_i^*$ be the shortest vector in this lattice. From Theorem 7.2 we know that

$$\|\mathbf{v}_i^*\|_2 \leqslant \gamma_i \sqrt[i]{d_i}$$

and so, from Theorem 7.3,

$$d_i \geqslant \frac{\|\mathbf{v}_i^*\|_2^i}{\gamma_i} \geqslant \left(\tfrac{3}{4}\right)^{\frac{1}{2}i(i-1)} \|\mathbf{v}_n^*\|_2^i,$$

where γ_i is a positive constant which depends only on i.

Now, let $D = d_1 d_2 \ldots d_n$. Then, at all times throughout the execution of the algorithm,

$$D \geqslant \frac{L^{\frac{1}{2}n(n+1)}}{\gamma_1 \gamma_2 \cdots \gamma_n} \geqslant \left(\tfrac{3}{4}\right)^{\frac{1}{6}\left(n^3-n\right)} L^{\frac{1}{2}n(n+1)} \tag{7.3}$$

Since $r_{jj} \leqslant \|\mathbf{b}_j\|_2$ for $1 \leqslant j \leqslant n$, when the algorithm commences we have

$$D \leqslant M^{\frac{1}{2}n(n+1)}. \tag{7.4}$$

If the algorithm does not terminate then it must pass through the EXCHANGE STEP on line 9 infinitely often. Let d_i' represent the value of d_i for each $1 \leqslant i \leqslant n$ and D' represent the value of D after the $\mathbf{QR}$ decomposition (or Givens rotation) on

the next line is calculated for an iteration in which the exchange step is undertaken. Proposition 7.1 implies that

$$d'_j \leqslant \frac{\sqrt{3}}{2} d_j$$

but $d'_i = d_i$ for $1 \leqslant i \leqslant n$, $i \neq j$. Therefore, we also have $D' \leqslant \sqrt{3}D/2$. Bearing in mind the lower and upper bounds of (7.3) and (7.4), we see that the number of exchange steps E which can be performed by the algorithm is bounded by

$$E \leqslant n(n+1)\left[\tfrac{1}{3}(n-1) + \tfrac{1}{2}\log_{2/\sqrt{3}} \frac{M}{L}\right].$$

We can therefore conclude that the algorithm must terminate after a finite number of iterations. Furthermore, because there can be no more than $n-1$ more iterations than the number of iterations which involve an exchange step (because j is incremented whenever an exchange does not take place), we can verify that the total number of iterations is indeed $O(n^2(n + \log(M/L)))$. □

The number of iterations required by the algorithm does not, in this case, amount to the number of arithmetic operations required. The subtraction of vectors of m elements on line 7 adds a further factor of m to the arithmetic complexity and so $O(mn^2(n + \log(M/L)))$ arithmetic operations are required. We mentioned at the beginning of this section that we have omitted from Algorithm 7.1 a size reduction procedure in which $\mathbf{b}_{j+1}$ is size reduced after each exchange step. This results in a saving of a factor of n in the complexity of the algorithm. However, in a practical implementation the size reduction is required to keep the sizes of matrix entries small.

It has been observed that the upper bound for the number of iterations required by Algorithm 7.1 and for the lengths of the vectors in the Lovász-reduced basis it produces as given by Theorem 6.5 are quite pessimistic in practice (see, for example Cohen, 1993; Schnorr & Euchner, 1994).

We conclude this section, and this chapter, by remarking that this algorithm has created a lot of interest because of its application to a wide range of problems such as cryptography, integer programming, algebra and simultaneous Diophantine approximation. New variants of the algorithm are appearing regularly. They aim to improve speed or tailor its performance in some other respect. We will not explore any of these applications or variants, except as regards simultaneous Diophantine approximation, which we will treat in the next chapter, and, as a consequence, problems in pulse train signal processing. Vallée (1991) and Schnorr & Euchner (1994) give many references to the aforementioned applications and variants.

CHAPTER 4

SIMULTANEOUS DIOPHANTINE APPROXIMATION

1. Introduction

Simultaneous Diophantine approximation has many possible definitions. One possible "definition" is that it is the theory or process by which one selects lattice points from a lattice of rank $n > 2$ which lie "close" to a linear form. Many authors restrict the linear form to being one-dimensional, *i.e.* a line. The sense in which the lattice point is close can be measured in either an ABSOLUTE SENSE or a RELATIVE SENSE. In the absolute sense, the closeness is measured by projecting the lattice point onto, for instance, the orthogonal complement of the linear form to be approximated and taking the norm of the projection. For example, given a projection matrix

$$\mathbf{P} = \mathbf{I} - \frac{\boldsymbol{\alpha}\boldsymbol{\alpha}^T}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}, \qquad \boldsymbol{\alpha} \in \mathbb{R}^n,$$

which projects along the line $\mathbb{R}\boldsymbol{\alpha}$, the problem of finding an integer vector $\mathbf{k} \in \mathbb{Z}^n$ which makes $\|\mathbf{Pk}\|$ small is simultaneous Diophantine approximation of the line $\mathbb{R}\boldsymbol{\alpha}$ by elements of $\mathbb{Z}^n$ in the absolute sense with respect to the norm $\|\cdot\|$. By approximation in the relative sense, we mean that the distance from the lattice point to the linear form is normalised by the distance of the lattice point from the origin. To use our previous example again, finding an integer vector $\mathbf{k} \in \mathbb{Z}^n$ which makes $\|\mathbf{Pk}\| \,/\, \|\mathbf{k}\|$ small is simultaneous Diophantine approximation of $\mathbb{R}\boldsymbol{\alpha}$ in the relative sense.

In particular, reference is frequently made in the literature to simultaneous Diophantine approximation problems that, given $n-1$ real numbers $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}$, involve finding n integers $p_1, p_2, \ldots, p_{n-1}, q$ such that

$$\max_{i=1,2,\ldots,n-1} \{|q\alpha_i - p_i|\}$$

is made small or that

$$\max_{i=1,2,\ldots,n-1} \left\{ \left| \alpha_i - \frac{p_i}{q} \right| \right\}$$

is made small. The former instance is simultaneous Diophantine approximation in the absolute sense and the latter in the relative sense.

As we mentioned above, some authors restrict the definition of simultaneous Diophantine approximation to encompass only the approximation of lines and not

higher-dimensional linear forms. Another form of approximation can then be distinguished that, given n real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$, involves finding n integers $k_1, k_2, \ldots, k_n$ in order to make

$$|k_1\alpha_1 + k_2\alpha_2 + \cdots + k_n\alpha|$$

as small as possible. If this expression can be made zero then an INTEGER RELATION for $\alpha_1, \alpha_2, \ldots, \alpha_n$ has been found.

One of the problems we address in this chapter is the problem of finding *best* simultaneous Diophantine approximations (or best approximate integer relations). This can be defined as the problem of finding lattice points that, of all lattice points which are of the same or smaller distance from the origin, most closely approximates the linear form of interest (a more general definition is forthcoming in Definition 2.7).

In this chapter, we consider two aspects of simultaneous Diophantine approximation. Firstly, we develop a theory regarding certain types of minimal sets of lattice points which are employed in algorithms for finding best simultaneous Diophantine approximations. This leads to realisable algorithms for lattices of ranks 2 and 3. Secondly, we review developments towards computationally feasible algorithms for finding good (but not necessarily best) simultaneous Diophantine approximations for lattices of arbitrary rank.

The definition of simultaneous Diophantine approximation we shall use is similar to that proposed by LAGARIAS (1983), which is a generalisation of the more classical definitions, of which those given in CASSELS (1957) and HARDY & WRIGHT (1979) are examples. From this definition, we produce a theory of (ρ, h)-*minimal* sets of lattice points. Through finding sequences of (ρ, h)-minimal sets, we show that best simultaneous Diophantine approximations can be found. We also introduce the idea of an *extended norm* which, while not contributing significantly to the complexity of the theory, extends the applicability of the algorithms we develop.

We apply this theory to construct algorithms for finding best simultaneous Diophantine approximations in lattices of ranks 2 and 3. Under appropriate conditions, we show that (ρ, h)-minimal sets always form bases of the lattice. In lattices of rank 2, we observe that the algorithm reduces to a simple additive continued fraction algorithm. We show that, for certain types of approximation problems, the algorithm will behave like Euclid's algorithm and produce outputs which can be interpreted as intermediate fractions of a simple continued fraction expansion of a real number. In other types of approximation problems, the algorithm will behave like Gauss' algorithm for lattice reduction.

For lattices of rank 3, we construct an algorithm which we show will find all best approximations (or their *equivalents*; see Definition 2.8) under appropriate conditions. The algorithm which we construct bears a strong resemblance to earlier algorithms for best simultaneous Diophantine approximation in lattices of rank 3, such

as the algorithms of Minkowski (1896*a*), Voronoĭ (1896)[1] and Furtwängler (1927). Our algorithm is not as fast the algorithms just mentioned, but it does have two advantages. Firstly, the new algorithm is capable of finding best approximations for a quite general class of approximation problems on lattices of rank 3 and, secondly, the algorithm has a fairly simple, additive nature (although it is not, strictly speaking, an additive continued fraction algorithm).

We demonstrate the operation of the algorithm for lattices of rank 3 with some numerical examples. We apply the algorithm to find best approximations of linear forms of a single variable (that is, simultaneous Diophantine approximation in the "traditional" sense) and of two variables (that is, approximate integer relations).

We also develop an accelerated version of the algorithm. This algorithm is able to find best simultaneous Diophantine approximations much more quickly. It does this at the expense of skipping a number of intermediate (ρ, h)-minimal sets. This algorithm is a generalisation of the algorithm of Furtwängler (1927).

Unfortunately, we do not have upper bounds on the running time required by either of the algorithms we present for lattices of rank 3. However, we present some numerical data that shows that both algorithms can find best approximations with an approximation error less than a prescribed bound in a "reasonable" amount of time.

Finally, we review developments towards computationally feasible algorithms for simultaneous Diophantine approximation with lattices of arbitrary rank. We briefly discuss the historical development of multi-dimensional continued fraction algorithms and explain one of the best-known examples: the algorithm of Brun (1919, 1920). We will see that, although it has a rather natural geometrical interpretation, it is known that it does not always find good approximations. Ideally, we would like an algorithm that produces best approximations for lattices of any rank. However, Lagarias (1982) has shown, following the result of van Emde Boas (1981) concerning lattice reduction, that certain (representative) best simultaneous Diophantine approximation problems are $\mathfrak{NP}$-hard. With the advent of the LLL algorithm, algorithms for finding good simultaneous Diophantine approximation have become computationally feasible. We conclude the chapter by discussing a close relative of the LLL algorithm, the *HJLS algorithm*, due to Hastad *et al.* (1989), and some similar algorithms. To conclude, we compare the bases produced by Brun's algorithm and the HJLS algorithm with our other algorithms through a numerical example for a lattice of rank 3.

2. Mathematical Preliminaries

2.1. Extended Norms. In this subsection we introduce extended norms and semi-norms and illustrate some of the properties we will subsequently require.

[1]This algorithm is known to the author only through the descriptions of Delone & Faddeev (1964) and Williams *et al.* (1980)

DEFINITION 2.1. An EXTENDED SEMI-NORM $\|\cdot\| = (\nu_1(\cdot), \nu_2(\cdot), \ldots, \nu_n(\cdot))$ is a map from a real vector space $\mathbb{R}^m$ to $\mathbb{R}^n$ such that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and for all $\lambda \in \mathbb{R}$,

(i) $\nu_1(\mathbf{x}) \geqslant \nu_2(\mathbf{x}) \geqslant \ldots \geqslant \nu_n(\mathbf{x}) \geqslant 0$,
(ii) $\|\lambda\mathbf{x}\| = |\lambda| \, \|\mathbf{x}\|$ and
(iii) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$

where $<$ is defined according to the usual lexicographic ordering of $\mathbb{R}^n$.

An extended semi-norm $\|\cdot\|$ is DEGENERATE if $\|\mathbf{x}\| = 0$ for all $\mathbf{x} \in \mathbb{R}^m$.

DEFINITION 2.2. An EXTENDED NORM $\|\cdot\|$ is an extended semi-norm for which it is true that

(i') $\|\mathbf{x}\| > 0$ if $\mathbf{x} \neq 0$.

REMARK 2.1. Obviously a norm is also an extended norm and a semi-norm is also an extended semi-norm.

REMARK 2.2. If $\|\cdot\| = (\nu_1(\cdot), \nu_2(\cdot), \ldots, \nu_n(\cdot))$ is an extended norm (extended semi-norm) then $\nu_1(\cdot)$ is a norm (semi-norm).

DEFINITION 2.3. An extended semi-norm (extended norm) is STRICTLY CONVEX if condition (iii) of Definition 2.1 is supplemented with

(iii') $\|\mathbf{x} + \mathbf{y}\| = \|\mathbf{x}\| + \|\mathbf{y}\| \;\Rightarrow\; \exists \lambda \in \mathbb{R}, \lambda \geqslant 0$ such that $\|\mathbf{x} - \lambda\mathbf{y}\| = 0$.

THEOREM 2.1. *Let $\|\cdot\|$ be an extended semi-norm on $\mathbb{R}^m$ and let $\Phi(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^m$, be the set*

$$\Phi(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^m \mid \|\mathbf{y}\| \leqslant \|\mathbf{x}\|\}.$$

The set $\Phi(\mathbf{x})$ is convex and centrally symmetric and if $\|\cdot\|$ is also an extended norm then $\operatorname{Int}\Phi(\mathbf{x})$ is a centrally symmetric convex body.

PROOF. That $\Phi(\mathbf{x})$ is convex follows from condition (iii) of Definition 2.1. That it is centrally symmetric follows from condition (ii). Now, $\operatorname{Int}\Phi(\mathbf{x})$ is an open convex set, so to show that $\operatorname{Int}\Phi(\mathbf{x})$ is a convex body if $\|\cdot\|$ is an extended norm, we need only show that $\operatorname{Int}\Phi(\mathbf{x})$ is bounded. To see this, consider any non-zero vector $\mathbf{y} \in \mathbb{R}^m$. Since $\|\mathbf{y}\| > 0$, we see that $\nu_1(\mathbf{y}) > 0$ as a consequence of condition (i). Thus, there exists some $\lambda \in \mathbb{R}$, $\lambda > 0$ such that $\|\lambda\mathbf{y}\| > \|\mathbf{x}\|$ and so $\Phi(\mathbf{x})$, and therefore its interior, is bounded. □

THEOREM 2.2. *A non-degenerate extended semi-norm $\|\cdot\| : \mathbb{R}^m \to \mathbb{R}^n$ is not an extended norm if and only if we can decompose $\|\cdot\|$ according to*

$$\|\mathbf{x}\| = \left\|\mathbf{P}^T\mathbf{x}\right\|', \tag{2.1}$$

where $\mathbf{P}^T \in \mathbb{R}^{m' \times m}$ is a matrix with $0 < m' < m$ and $\|\cdot\|' : \mathbb{R}^{m'} \to \mathbb{R}^n$ is an extended norm. Furthermore, if $\|\cdot\|$ is strictly convex then we can decompose it according to (2.1) with a strictly convex extended norm $\|\cdot\|'$.

PROOF. The sufficiency of the first part of the theorem statement is obvious. We will prove its necessity by construction. Let

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^m \mid \|\mathbf{x}\| = 0\}.$$

Now, $\mathcal{S}$ must be a vector subspace of $\mathbb{R}^m$ with $0 < \dim(\mathcal{S}) < m$. Consider an orthonormal basis of $\mathcal{S}$. Let $\mathbf{Q}$ be the matrix of vectors in the basis, arranged as columns. Let $\mathbf{P}$ be matrix of vectors of an orthonormal basis of the orthogonal complement of $\mathcal{S}$. Then $\mathbf{P} \in \mathbb{R}^{m \times m'}$ where $0 < m' = m - \dim(\mathcal{S}) < m$. It is readily deduced that

$$\mathbf{P}\mathbf{P}^T + \mathbf{Q}\mathbf{Q}^T = \mathbf{I}.$$

We define the extended norm $\|\mathbf{y}\|' = \left\|\mathbf{P}^T\mathbf{y}\right\|$ for all $\mathbf{y} \in \mathbb{R}^{m'}$. Then

$$\|\mathbf{x}\| = \left\|\left(\mathbf{P}\mathbf{P}^T + \mathbf{Q}\mathbf{Q}^T\right)\mathbf{x}\right\| = \left\|\mathbf{P}\mathbf{P}^T\mathbf{x}\right\| = \left\|\mathbf{P}^T\mathbf{x}\right\|'.$$

□

COROLLARY 2.1. *An extended semi-norm $\|\cdot\|$ on $\mathbb{R}^m$ can be expressed as the extended norm of a linear map from $\mathbb{R}^m$ to $\mathbb{R}^{m'}$ with $0 \leqslant m' \leqslant m$.*

DEFINITION 2.4. Two extended semi-norms $\|\cdot\|^*$ and $\|\cdot\|^\dagger$, both defined on $\mathbb{R}^m$, are TRANSVERSE with respect to each other if

$$\left\{\mathbf{x} \in \mathbb{R}^m \mid \|\mathbf{x}\|^* = \|\mathbf{x}\|^\dagger = \mathbf{0}\right\} = \{\mathbf{0}\}.$$

DEFINITION 2.5. An extended norm $\|\cdot\|^\dagger$ is an EXTENSION of another extended norm $\|\cdot\|^*$ if both are defined on $\mathbb{R}^m$ and, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$,

$$\|\mathbf{x}\|^* < \|\mathbf{y}\|^* \Rightarrow \|\mathbf{x}\|^\dagger < \|\mathbf{y}\|^\dagger.$$

THEOREM 2.3. *For every norm $\|\cdot\|^*$ defined on $\mathbb{R}^m$ there exists a strictly convex extended norm $\|\cdot\|^\dagger$ extended from it.*

PROOF. The proof is by construction. The Euclidean norm $\|\cdot\|_2$ is strictly convex. Because all norms on $\mathbb{R}^m$ are similar we know that there exists some $\mu \in \mathbb{R}$, $\mu > 0$ such that, for all $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\|^* \geqslant \mu \|\mathbf{x}\|_2$. Consider the extended norm $\|\mathbf{x}\|^\dagger = (\|\mathbf{x}\|^*, \mu \|\mathbf{x}\|_2)$ where $\|\cdot\|^\dagger : \mathbb{R}^m \to \mathbb{R}^2$. It can be easily verified that $\|\cdot\|^\dagger$ is strictly convex and extended from $\|\cdot\|$. □

In Theorem 2.3 we have the most important reason for introducing extended norms and semi-norms which is that norms which otherwise are not strictly convex (such as the sup-norm) can be extended to meet this criterion. The strict convexity is an important ingredient in what follows as it allows many of the theorem statements to be simplified and removes a number of special cases from consideration.

2.2. Simultaneous Diophantine Approximation.

DEFINITION 2.6. A SYSTEM for simultaneous Diophantine approximation consists of a lattice Ω together with two non-trivial, transverse extended semi-norms which we call the RADIUS FUNCTION and HEIGHT FUNCTION of the system.

Simultaneous Diophantine approximation involves finding elements of Ω that have "small" radius without having a height that is "too large."

Frequently, we will not refer to a system for simultaneous Diophantine approximation directly. Rather, its existence is implied from the context.

We can define best approximations in the following way.

DEFINITION 2.7. A non-zero lattice point $\mathbf{x}$ in a lattice Ω is a BEST APPROXIMATION IN THE ABSOLUTE SENSE with respect to a radius function, ρ, and a height function, h, if, for all non-zero $\mathbf{y} \in \Omega$,

$$\rho(\mathbf{y}) \leqslant \rho(\mathbf{x}) \;\Rightarrow\; h(\mathbf{y}) \geqslant h(\mathbf{x})$$

and

$$h(\mathbf{y}) \leqslant h(\mathbf{x}) \;\Rightarrow\; \rho(\mathbf{y}) \geqslant \rho(\mathbf{x}).$$

For the rest of this chapter, the only notion of best approximation we require is that of best approximation in the absolute sense. For this reason, we will omit the qualification "in the absolute sense."

DEFINITION 2.8. Two lattice points $\mathbf{x}$ and $\mathbf{y}$ in a simultaneous Diophantine approximation system are EQUIVALENT with respect to the radius function, ρ, and a height function, h, if $\rho(\mathbf{x}) = \rho(\mathbf{y})$ and $h(\mathbf{x}) = h(\mathbf{y})$.

DEFINITION 2.9. In a simultaneous Diophantine approximation system, we will say that the radius function ρ and the height function h are COMPLEMENTARY on the lattice Ω of rank n if ρ and h can be decomposed so that $\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|^*$ and $h(\mathbf{x}) = \left\|\mathbf{R}^T\mathbf{x}\right\|^\dagger$ where $\|\cdot\|^*$ and $\|\cdot\|^\dagger$ are extended norms and $\mathbf{P}^T$ and $\mathbf{R}^T$ are matrices such that there are n_1 columns of $\mathbf{P}$ and n_2 columns of $\mathbf{R}$ in the real span of Ω and $n_1 + n_2 = n$.

Suppose $\rho : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $h : \mathbb{R}^m \to \mathbb{R}^{n_2}$. We define $\rho h(\cdot) = (\rho(\cdot), h(\cdot))$, a map from $\mathbb{R}^m$ to $\mathbb{R}^{(n_1+n_2)}$. Similarly, we define $h\rho(\cdot) = (h(\cdot), \rho(\cdot))$. With the usual lexicographic ordering of $\mathbb{R}^{(n_1+n_2)}$, we notice that $\rho h(\cdot)$ and $h\rho(\cdot)$ are somewhat like extended semi-norms. In fact, they obey all the conditions of Definition 2.1 except condition (i). In its place, we can only say that $\rho h(\mathbf{x}) \geqslant 0$ and, because ρ and h are transverse, $\rho h(\mathbf{x}) > 0$ if $\mathbf{x} \neq 0$. Of course, the same is true for $h\rho(\cdot)$.

We now introduce the (ρ, h)-minimal set, upon which the theory of the following three sections are based.

DEFINITION 2.10. An n-tuple of linearly independent points $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ in a lattice Ω of rank n is (ρ, h)-MINIMAL for the radius function ρ and height function h if

$$\rho h(\mathbf{v}_i) \leqslant \rho h(\mathbf{v}_j) \tag{2.2}$$

when $i \leqslant j$ and if, for each $1 \leqslant j \leqslant n$ and for all non-zero $\mathbf{w} \in \Omega$ which are linearly independent of $\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\}$, it is true that either

$$\rho h(\mathbf{w}) \geqslant \max_{1 \leqslant i \leqslant j} \{\rho h(\mathbf{v}_i)\} \tag{2.3}$$

or

$$h\rho(\mathbf{w}) \geqslant \max_{1 \leqslant i \leqslant n} \{h\rho(\mathbf{v}_i)\} \tag{2.4}$$

REMARK 2.3. If $\rho = h$ is a norm (or extended norm) then a (ρ, h)-minimal set of the lattice is simply a set of minimal vectors with respect to that norm.

REMARK 2.4. If $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is (ρ, h)-minimal then $\mathbf{v}_1$ is a best approximation.

3. (ρ, h)-Minimal Sets

3.1. General Properties.

THEOREM 3.1. *If $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is an ordered set of linearly independent points in a lattice Ω of rank n in $\mathbb{R}^m$ such that $\rho h(\mathbf{x}_i) \leqslant \rho h(\mathbf{x}_j)$ whenever $i \leqslant j$ then there exists a (ρ, h)-minimal set $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ such that*

$$\rho h(\mathbf{v}_i) \leqslant \rho h(\mathbf{x}_i) \tag{3.1}$$

for all $i = 1, 2, \ldots, n$ and

$$\max_{1 \leqslant i \leqslant n} \{h\rho(\mathbf{v}_i)\} \leqslant \max_{1 \leqslant i \leqslant n} \{h\rho(\mathbf{x}_i)\}. \tag{3.2}$$

PROOF. We will describe a procedure through which a (ρ, h)-minimal set can be constructed which satisfies (3.1) and (3.2). The procedure iterates through n steps. At each step k, we construct an intermediate ordered set of linearly independent points $\mathcal{V}^{(k)} = \left(\mathbf{v}_1^{(k)}, \mathbf{v}_2^{(k)}, \ldots, \mathbf{v}_n^{(k)}\right)$ from the previous set, $\mathcal{V}^{(k-1)}$. To initialise, we set $\mathcal{V}^{(0)} = \mathcal{X}$.

Consider the sets

$$\Delta_j^{(k)} = \left\{ \mathbf{z} \in \mathbb{R}^m \;\middle|\; \rho h(\mathbf{z}) < \rho h\left(\mathbf{v}_j^{(k)}\right);\ h\rho(\mathbf{z}) < \max_{1 \leqslant i \leqslant n} \left\{ h\rho\left(\mathbf{v}_i^{(k)}\right) \right\} \right\}.$$

The $\Delta_j^{(k)}$ are bounded, centrally symmetric convex sets. The convexity and central symmetry of these sets are straightforward consequences of properties (ii) and (iii) of extended semi-norms in Definition 2.1. To see that the $\Delta_j^{(k)}$ are bounded, write $\rho(\cdot) = (\rho_1(\cdot), \rho_2(\cdot), \ldots, \rho_{n_1}(\cdot))$ and $h(\cdot) = (h_1(\cdot), h_2(\cdot), \ldots, h_{n_2}(\cdot))$. Consider $\mathbf{z} \in \Delta_j^{(k)}$

for some j and k. Because ρ and h are transverse, either $\rho_1(\mathbf{z}) > 0$ or $h_1(\mathbf{z}) > 0$. If $\rho_1(\mathbf{z}) > 0$ then there exists some $\mu \in \mathbb{R}$, $\mu > 0$ such that

$$\rho h(\mu\mathbf{z}) \geqslant \rho h\Big(\mathbf{v}_j^{(k)}\Big)$$

and if $h_1(\mathbf{z}) > 0$ then there exists some $\mu > 0$ such that

$$h\rho(\mu\mathbf{z}) \geqslant \max_{1\leqslant i\leqslant n}\Big\{h\rho\Big(\mathbf{v}_i^{(k)}\Big)\Big\}.$$

Thus, $\Delta_j^{(k)}$ is bounded.

Clearly,

$$\Delta_i^{(k)} \subseteq \Delta_j^{(k)}$$

whenever $i \leqslant j$. Let

$$\Phi_j^{(k)} = \Big\{\mathbf{x} \in \Omega \cap \Delta_j^{(k)} \;\Big|\; \mathbf{x} \text{ is linearly independent of } \Big\{\mathbf{v}_1^{(k)}, \mathbf{v}_2^{(k)}, \ldots, \mathbf{v}_{j-1}^{(k)}\Big\}\Big\}.$$

The cardinalities of these sets must be finite since the $\Delta_j^{(k)}$ are bounded.

Suppose that, at some step k, the elements of $\mathcal{V}^{(k-1)}$ obey

$$\rho h\Big(\mathbf{v}_i^{(k-1)}\Big) \leqslant \rho h\Big(\mathbf{v}_j^{(k-1)}\Big) \tag{3.3}$$

whenever $i \leqslant j$ and

$$\Phi_1^{(k-1)} = \Phi_2^{(k-1)} = \cdots = \Phi_{k-1}^{(k-1)} = \emptyset. \tag{3.4}$$

We observe that, for step 1, (3.3) holds by assumption and (3.4) holds trivially.

Assuming (3.3) and (3.4) hold at step k, we produce $\mathcal{V}^{(k)}$ as follows. If $\Phi_k^{(k-1)}$ is not empty then we choose a lattice point $\mathbf{s} \in \Phi_k^{(k-1)}$ such that

$$\rho h(\mathbf{s}) = \min_{\mathbf{t}\in\Phi_k^{(k-1)}}\{\rho h(\mathbf{t})\}.$$

If $\Phi_k^{(k-1)}$ is empty then we set $\mathbf{s} = \mathbf{v}_k^{(k-1)}$. Let r denote the minimum index such that $\mathbf{s}$ is linearly independent of $\Big\{\mathbf{v}_1^{(k-1)}, \mathbf{v}_2^{(k-1)}, \ldots, \mathbf{v}_{r-1}^{(k-1)}\Big\}$. Thus, $r \geqslant k$.

We now create a new set of linearly independent lattice points by replacing $\mathbf{v}_r^{(k-1)}$ with $\mathbf{s}$ and reordering so that (3.3) is satisfied for step $k+1$. Thus, we set

$$\mathcal{V}^{(k)} = \Big(\mathbf{v}_1^{(k-1)}, \ldots, \mathbf{v}_{k-1}^{(k-1)}, \mathbf{s}, \mathbf{v}_k^{(k-1)}, \ldots, \mathbf{v}_{r-1}^{(k-1)}, \mathbf{v}_{r+1}^{(k-1)}, \ldots, \mathbf{v}_n^{(k-1)}\Big),$$

which ensures that (3.3) and (3.4) again hold for step $k+1$. Furthermore,

$$\rho h\Big(\mathbf{v}_i^{(k)}\Big) \leqslant \rho h\Big(\mathbf{v}_i^{(k-1)}\Big) \tag{3.5}$$

for all $i = 1, 2, \ldots, n$ and

$$\max_{1\leqslant i\leqslant n}\Big\{h\rho\Big(\mathbf{v}_i^{(k)}\Big)\Big\} \leqslant \max_{1\leqslant i\leqslant n}\Big\{h\rho\Big(\mathbf{v}_i^{(k-1)}\Big)\Big\}. \tag{3.6}$$

Thus, at the completion of step n, the sets $\Phi_i^{(n)}$ are all empty and so $\mathcal{V}^{(n)}$ must be (ρ, h)-minimal. By induction on (3.5) and (3.6), we can see that (3.1) and (3.2) must also be satisfied, so the theorem is proved with $\mathcal{V} = \mathcal{V}^{(n)}$. □

LEMMA 3.1. *Suppose* $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ *and* $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ *are* (ρ, h)*-minimal sets of a lattice* Ω *of rank* n*. If*

$$\rho h(\mathbf{u}_i) \leqslant \rho h(\mathbf{v}_i) \tag{3.7}$$

for all $i = 1, 2, \ldots, n$ *then*

$$\max_{1\leqslant i\leqslant n} \{h\rho(\mathbf{u}_i)\} \geqslant \max_{1\leqslant i\leqslant n} \{h\rho(\mathbf{v}_i)\}. \tag{3.8}$$

PROOF. Suppose there exists a pair of (ρ, h)-minimal sets $\mathcal{U}$ and $\mathcal{V}$ which satisfies (3.7) but not (3.8). Now, for any $1 \leqslant q \leqslant n$, consider the minimum index r such that $\mathbf{u}_r$ is linearly independent of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{q-1}\}$. Thus, $r \leqslant q$. Furthermore, $\rho h(\mathbf{u}_r) \leqslant \rho h(\mathbf{v}_q)$ because $\rho h(\mathbf{u}_r) \leqslant \rho h(\mathbf{u}_q)$ and $\rho h(\mathbf{u}_q) \leqslant \rho h(\mathbf{v}_q)$. Since we assume that (3.8) is not satisfied, but $\mathcal{V}$ is (ρ, h)-minimal, we must conclude that $\rho h(\mathbf{u}_r) \geqslant \rho h(\mathbf{v}_q)$ which implies that

$$\rho h(\mathbf{u}_q) = \rho h(\mathbf{v}_q). \tag{3.9}$$

Since (3.9) is satisfied for all q, we conclude that (3.8) must be satisfied as an equality, and the lemma is proved. □

LEMMA 3.2. *Suppose* Ω *is a lattice of rank* n *in* $\mathbb{R}^m$*,* $\mathcal{E}$ *is its real span and* ρ *and* h *are radius and height functions. Consider the parameterised set*

$$\Psi(\lambda;\ \mathbf{x}, \mathbf{y}) = \{\mathbf{z} \in \mathcal{E} \ \mid\ \rho(\mathbf{z}) < \rho(\mathbf{x});\ h(\mathbf{z}) < \lambda h(\mathbf{y})\}$$

with $\lambda \in \mathbb{R}$ *and* $\mathbf{x}, \mathbf{y} \in \mathcal{E}$*. If* $\rho(\mathbf{x}) > \mathbf{0}$ *and* $h(\mathbf{y}) > \mathbf{0}$ *but* $\rho(\mathbf{y}) = \mathbf{0}$ *then there exists some* $\lambda_c > 0$ *such that* $\Psi(\lambda_c;\ \mathbf{x}, \mathbf{y})$ *contains a point in* Ω *other than the origin.*

PROOF. We write

$$\rho(\cdot) = (\rho_1(\cdot), \rho_2(\cdot), \ldots, \rho_{n_1}(\cdot)) \qquad \text{and} \qquad h(\cdot) = (h_1(\cdot), h_2(\cdot), \ldots, h_{n_2}(\cdot)).$$

Let $\mathbf{B}$ be an $m \times n$ matrix, the columns of which form a basis of Ω. Every vector $\mathbf{z} \in \mathcal{E}$ can be expressed as $\mathbf{z} = \mathbf{B}\mathbf{z}'$ with $\mathbf{z}' \in \mathbb{R}^n$. Consider the parameterised set

$$\Psi'(\lambda;\ \mathbf{x}, \mathbf{y}) = \{\mathbf{z}' \in \mathbb{R}^n \ \mid\ \rho_1(\mathbf{B}\mathbf{z}') < \rho_1(\mathbf{x});\ h_1(\mathbf{B}\mathbf{z}') < \lambda h_1(\mathbf{y})\}.$$

Now, if $\mathbf{z}' \in \Psi'(\lambda;\ \mathbf{x}, \mathbf{y})$ then $\mathbf{B}\mathbf{z}' \in \Psi(\lambda;\ \mathbf{x}, \mathbf{y})$ and $\Psi'(\lambda;\ \mathbf{x}, \mathbf{y})$ is a centrally symmetric convex body with non-zero volume when $\lambda > 0$.

If $\mathbf{z}' \in \Psi'\big(\frac{1}{2};\ \mathbf{x}, \mathbf{y}\big)$ then, with $\mathbf{z} = \mathbf{B}\mathbf{z}'$, $\mathbf{x} = \mathbf{B}\mathbf{x}'$ and $\mathbf{y} = \mathbf{B}\mathbf{y}'$, we have, for all $k \in \mathbb{N}$,

$$\begin{aligned} \rho_1\big(\mathbf{z} + \big(k + \tfrac{1}{2}\big)\mathbf{y}\big) &= \rho_1(\mathbf{z}) < \rho_1(\mathbf{x}), \\ h_1\big(\mathbf{z} + \big(k + \tfrac{1}{2}\big)\mathbf{y}\big) &\leqslant h_1(\mathbf{z}) + \big(k + \tfrac{1}{2}\big)h_1(\mathbf{y}) \\ &< (k+1)h_1(\mathbf{y}) \end{aligned}$$

and

$$\begin{aligned} h_1\big(\mathbf{z} + \big(k + \tfrac{1}{2}\big)\mathbf{y}\big) &\geqslant \big(k + \tfrac{1}{2}\big)h_1(\mathbf{y}) - h_1(\mathbf{z}) \\ &> kh_1(\mathbf{y}). \end{aligned}$$

Thus, $\mathbf{z}' + \big(k + \frac{1}{2}\big)\mathbf{y}' \in \Psi'(k+1;\ \mathbf{x}, \mathbf{y}) \setminus \Psi'(k;\ \mathbf{x}, \mathbf{y})$ which implies that

$$\operatorname{vol} \Psi'(k+1;\ \mathbf{x}, \mathbf{y}) \geqslant \operatorname{vol} \Psi'(k;\ \mathbf{x}, \mathbf{y}) + \operatorname{vol} \Psi'\big(\tfrac{1}{2};\ \mathbf{x}, \mathbf{y}\big).$$

This implies that there is some λ_c for which

$$\operatorname{vol} \Psi'(\lambda_c;\ \mathbf{x}, \mathbf{y}) > 2^n.$$

Using Minkowski's Fundamental Theorem (Theorem 3.2), we find that $\Psi'(\lambda_c;\ \mathbf{x}, \mathbf{y})$ must contain some non-zero point in $\mathbb{Z}^n$ and therefore $\Psi(\lambda_c;\ \mathbf{x}, \mathbf{y})$ must contain some non-zero point in Ω. □

Definition 3.1. An extended semi-norm $\|\cdot\|$ is NULL-SPANNED by a set of points $\mathcal{X}$ if there exists a subset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ of $\mathcal{X}$ such that $\|\mathbf{x}_i\| = 0$ for all $i = 1, 2, \ldots, m$, and $\|\mathbf{y}\| > 0$ for all non-zero $\mathbf{y}$ in the orthogonal complement of the subspace spanned by $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$.

Lemma 3.3. *If $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is (ρ, h)-minimal in a lattice Ω of rank n in $\mathbb{R}^m$ and the radius function ρ is not null-spanned by $\mathcal{V}$ then there exists a lattice point $\mathbf{s} \in \Omega$ such that, for some index r, $\rho h(\mathbf{s}) < \rho h(\mathbf{v}_r)$ and $\mathbf{s}$ is linearly independent of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{r-1}\}$.*

Proof. Let r be the smallest index such that

$$\rho(\mathbf{v}_r) > 0.$$

Let $\mathcal{E}'$ represent the vector space spanned by $\{\mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ and let Ω' be the lattice generated by this basis. Now, because ρ is not null-spanned by $\mathcal{V}$, we know that there is some non-zero $\mathbf{z} \in \mathcal{E}'$ such that $\rho(\mathbf{z}) = 0$.

Consider the set

$$\Psi(\lambda) = \{\mathbf{y} \in \mathcal{E}' \mid \rho(\mathbf{y}) < \rho(\mathbf{v}_r);\ h(\mathbf{y}) < \lambda h(\mathbf{z})\}.$$

From Lemma 3.2 there exists some positive value of λ for which this set contains a lattice point, say $\mathbf{s}$, in Ω'. Clearly, $\mathbf{s}$ is also an element of Ω and is linearly independent of $\{\mathbf{v}_1, \ldots, \mathbf{v}_{r-1}\}$. Thus, the lemma is proved. □

Definition 3.2. If $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ and $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ are (ρ, h)-minimal in a lattice Ω of rank n in $\mathbb{R}^m$ then $\mathcal{V}$ is a SUCCESSOR to $\mathcal{U}$ if

$$\rho h(\mathbf{v}_1) \leqslant \rho h(\mathbf{u}_1) \tag{3.10}$$

and

$$\max_{1 \leqslant i \leqslant n} \{h\rho(\mathbf{v}_i)\} \geqslant \max_{1 \leqslant i \leqslant n} \{h\rho(\mathbf{u}_i)\}. \tag{3.11}$$

Likewise, $\mathcal{U}$ is a PREDECESSOR of $\mathcal{V}$. $\mathcal{V}$ is a STRICT SUCCESSOR to $\mathcal{U}$ (and $\mathcal{U}$ is a STRICT PREDECESSOR of $\mathcal{V}$) if (3.10) is satisfied strictly. $\mathcal{V}$ is an IMMEDIATE SUCCESSOR to $\mathcal{U}$ (and $\mathcal{U}$ is an IMMEDIATE PREDECESSOR of $\mathcal{V}$) if there exists no (ρ, h)-minimal set which is both a strict successor to $\mathcal{U}$ and a strict predecessor of $\mathcal{V}$.

For a (ρ, h)-minimal set $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ in a lattice Ω, we will make use of the sets

$$\Xi_j(\mathcal{V}) = \{\mathbf{x} \in \Omega \mid \rho h(\mathbf{x}) < \rho h(\mathbf{v}_j); \quad \mathbf{x} \text{ is linearly independent of } \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}\}\}.$$

and

$$\Xi(\mathcal{V}) = \bigcup_{j=1}^{n} \Xi_j(\mathcal{V}).$$

THEOREM 3.2. *If $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ is (ρ, h)-minimal in a lattice Ω of rank n in $\mathbb{R}^m$ and $\Xi(\mathcal{U})$ is not empty then there exists $\mathbf{s} \in \Xi(\mathcal{U})$ such that, for all $\mathbf{t} \in \Xi(\mathcal{U})$,*

$$h\rho(\mathbf{s}) \leqslant h\rho(\mathbf{t}). \tag{3.12}$$

Furthermore, if r is the minimum index such that $\rho h(\mathbf{s}) \leqslant \rho h(\mathbf{u}_r)$ and q is the minimum index such that $\mathbf{s}$ is linearly dependent on $\{\mathbf{u}_1, \ldots, \mathbf{u}_q\}$ then $r \leqslant q$ and

$$\mathcal{V} = (\mathbf{u}_1, \ldots, \mathbf{u}_{r-1}, \mathbf{s}, \mathbf{u}_r, \ldots, \mathbf{u}_{q-1}, \mathbf{u}_{q+1}, \ldots, \mathbf{u}_n). \tag{3.13}$$

is (ρ, h)-minimal and an immediate successor to $\mathcal{U}$.

PROOF. The existence of $\mathbf{s}$ is guaranteed by the fact that, for any $\mathbf{r} \in \Xi_j(\mathcal{U})$, the set of points $\mathbf{x} \in \mathbb{R}^m$ which satisfy $\rho h(\mathbf{x}) < \rho h(\mathbf{u}_n)$ and $h\rho(\mathbf{x}) \leqslant h\rho(\mathbf{r})$ is bounded and thus contains only a finite number of lattice points from which $\mathbf{s}$ can be chosen. Clearly, $h\rho(\mathbf{s}) \geqslant h\rho(\mathbf{u}_i)$ for all $i = 1, 2, \ldots, n$, otherwise $\mathcal{U}$ would not be (ρ, h)-minimal.

Now $r \leqslant q$, for if $r > q$ then $\mathbf{s}$ cannot belong to any $\Xi_j(\mathcal{U})$. We now show that $\mathcal{V}$ is (ρ, h)-minimal. Clearly, (2.2) of Definition 2.10 is satisfied. Suppose (2.3) is not satisfied for some j. Thus we have some $\mathbf{w} \in \Omega$ which is linearly independent of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{j-1}\}$ such that $\rho h(\mathbf{w}) < \rho h(\mathbf{v}_j)$. If $1 \leqslant j \leqslant r$ or $q < j \leqslant n$ then, from (3.13), we see that $\mathbf{w}$ is linearly independent of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{j-1}\}$ and $\rho h(\mathbf{w}) < \rho h(\mathbf{u}_j)$. If, on the other hand, $r < j \leqslant q$ then $\mathbf{w}$ is linearly independent of $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{j-2}\}$ and $\rho h(\mathbf{w}) < \rho h(\mathbf{u}_{j-1})$. In either case, the implication is that $h\rho(\mathbf{w}) \geqslant h\rho(\mathbf{u}_i)$ for all $i = 1, 2, \ldots, n$. Furthermore, $h\rho(\mathbf{w}) \geqslant h\rho(\mathbf{s})$ because $\mathbf{w} \in \Xi_j(\mathcal{U})$. Thus, $h\rho(\mathbf{w}) \geqslant h\rho(\mathbf{v}_i)$ for all $i = 1, 2, \ldots, n$ and therefore (2.4) is satisfied. Thus, $\mathcal{V}$ is a successor to $\mathcal{U}$.

Finally, we show that $\mathcal{V}$ is an immediate successor to $\mathcal{U}$. That is, we show that there is no strict successor to $\mathcal{U}$ which is also a strict predecessor of $\mathcal{V}$. Suppose such a successor exists, say $\mathcal{U}' = (\mathbf{u}'_1, \mathbf{u}'_2, \ldots, \mathbf{u}'_n)$. Now, $\rho h(\mathbf{u}'_1) < \rho h(\mathbf{u}_1)$ which

implies that $\mathbf{u}'_1 \in \Xi_1(\mathcal{U})$. Hence, $h\rho(\mathbf{u}'_1) \geqslant h\rho(\mathbf{s})$ but, because $\mathcal{U}'$ is a predecessor of $\mathcal{V}$, $h\rho(\mathbf{u}'_1) \leqslant h\rho(\mathbf{s})$ which implies that $h\rho(\mathbf{u}'_1) = h\rho(\mathbf{s})$ which in turn implies that $\rho h(\mathbf{u}'_1) \geqslant \rho h(\mathbf{v}_1)$, contradicting the assumption that $\mathcal{U}'$ is a strict predecessor of $\mathcal{V}$. Hence, the theorem is proved. □

DEFINITION 3.3. We call an immediate successor $\mathcal{V}$ to $\mathcal{U}$ of the type described in Theorem 3.2 an INCREMENTAL SUCCESSOR. Furthermore, we call $\mathbf{s}$ the INNOVATION into $\mathcal{V}$ and $\mathbf{u}_q$ the INVETERATION from $\mathcal{U}$.

COROLLARY 3.1. *If ρ and h are radius and height functions and ρ is not null-spanned by a lattice Ω of rank n in $\mathbb{R}^m$ then there exists a sequence of (ρ, h)-minimal sets in Ω which can be ordered so that each element of the sequence is a successor of the previous element and the sequence is non-terminating.*

THEOREM 3.3. *Consider a procedure which, given a (ρ, h)-minimal set of a lattice Ω, produces an incremental successor and consider an algorithm which consists of iterating this procedure and outputting the innovations in sequence. Suppose such an algorithm is initialised with a (ρ, h)-minimal set $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. If there exists a best approximation $\mathbf{p} \in \Omega$ with respect to ρ and h such that $\rho(\mathbf{p}) \leqslant \rho(\mathbf{v}_1)$ then, after a finite number of iterations, an equivalent best approximation will appear as an output of the algorithm.*

PROOF. If $\rho h(\mathbf{p}) = \rho h(\mathbf{v}_1)$ then $\mathbf{p}$ is equivalent to $\mathbf{v}_1$. Suppose $\rho h(\mathbf{p}) < \rho h(\mathbf{v}_1)$, in which case

$$h\rho(\mathbf{p}) > \max_{i=1,2,\ldots,n} \{h\rho(\mathbf{v}_i)\}.$$

Now, $\mathcal{V}$ has an incremental successor because $\Xi(\mathcal{V})$ is not empty since $\rho(\mathbf{v}_1) > \rho(\mathbf{p})$. Suppose that, initially, the innovation is chosen from the set of lattice points $\mathbf{x} \in \Omega$ which satisfy $\rho h(\mathbf{x}) < \rho h(\mathbf{v}_n)$ and $h\rho(\mathbf{x}) < h\rho(\mathbf{p})$. The number of possibilities is finite and must strictly decrease from one iteration to the next. Therefore, within a finite number of iterations, an incremental successor $\mathcal{V}^* = (\mathbf{v}_1^*, \mathbf{v}_2^*, \ldots, \mathbf{v}_n^*)$ is found in which $h\rho\big(\mathbf{v}_j^*\big) \geqslant h\rho(\mathbf{p})$ is satisfied for that index j which corresponds to the innovation and for that index only. Because $\mathbf{p}$ is a best approximation, this implies that $\rho h(\mathbf{v}_i^*) < \rho h(\mathbf{p})$ for all $i \neq j$. If $h\rho\big(\mathbf{v}_j^*\big) > h\rho(\mathbf{p})$ then we have a contradiction: either $\mathbf{p}$ cannot be a best approximation or $\mathcal{V}^*$ is not (ρ, h)-minimal. Therefore $\mathbf{v}_j^*$ is a best approximation that is equivalent to $\mathbf{p}$. □

DEFINITION 3.4. We call an algorithm of the type described in Theorem 3.3 an INCREMENTAL SUCCESSOR ALGORITHM.

THEOREM 3.4. *Consider a system for simultaneous Diophantine approximation consisting of a lattice Ω of rank n in $\mathbb{R}^m$, a radius function ρ and a height function h. Let us write $\rho(\mathbf{x})$ as*

$$\rho(\mathbf{x}) = (\rho_1(\mathbf{x}), \rho_2(\mathbf{x}), \ldots, \rho_{n_1}(\mathbf{x})).$$

Suppose an incremental successor algorithm is executed on a (ρ, h)-minimal set $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ with $\rho(\mathbf{u}_1) > \mathbf{0}$ and let K be the number of incremental successors which are found by the algorithm before a best approximation $\mathbf{p}$ is output with $\rho h(\mathbf{p}) < \rho h(\mathbf{u}_1)$. Then

$$K \leqslant \left[1 + 2\frac{\rho_1(\mathbf{u}_n)}{\rho_1(\mathbf{u}_1)}\right]^n. \tag{3.14}$$

PROOF. The proof uses the pigeonhole principle. Let $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(K)}$ denote the sequence of innovations into the incremental successors $\mathcal{V}^{(1)}, \mathcal{V}^{(2)}, \ldots, \mathcal{V}^{(K)}$ found by the incremental successor algorithm for the input $\mathcal{U}$. Suppose none of the innovations is a best approximation of the type referred to in the theorem statement.

Let us write

$$h(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_{n_2}(\mathbf{x})).$$

Consider the norm $\|\cdot\|$ that is defined on $\mathbb{R}^n$ as

$$\|\mathbf{y}\| = \max\left\{\frac{\eta_1(\mathbf{B}\mathbf{y})}{\eta_1(\mathbf{u}_1)}, \frac{\mu_1(\mathbf{B}\mathbf{y})}{\mu_1(\mathbf{s}^{(K)})}\right\}$$

where $\mathbf{B}$ is a basis matrix of Ω. Now, consider the parameterised set

$$\mathcal{S}(\mathbf{y}, \lambda) = \{\mathbf{z} \in \mathbb{R}^n \mid \|\mathbf{z} - \mathbf{y}\| < \lambda\}.$$

The set $\mathcal{S}(\mathbf{0}, \lambda)$ is a convex body with a non-zero volume when $\lambda > 0$ and $\mathcal{S}(\mathbf{y}, \lambda)$ is its translation in $\mathbb{R}^n$ by $\mathbf{y}$ (a metric ball of radius λ centred on $\mathbf{y}$). Furthermore,

$$\operatorname{vol}\mathcal{S}(\mathbf{y}, \lambda) = V(\lambda) = \lambda^n V(1).$$

Let us express each $\mathbf{s}^{(j)}$ as $\mathbf{s}^{(j)} = \mathbf{B}\mathbf{y}^{(j)}$ where $\mathbf{y}^{(j)} \in \mathbb{Z}^n$. If

$$\mathcal{S}\big(\mathbf{y}^{(i)}, \tfrac{1}{2}\big) \cap \mathcal{S}\big(\mathbf{y}^{(j)}, \tfrac{1}{2}\big) \neq \emptyset$$

for some $1 \leqslant i < j \leqslant K$ then there exists some $\mathbf{z} \in \mathbb{R}^n$ such that

$$\|\mathbf{y}^{(i)} - \mathbf{z}\| < \tfrac{1}{2} \qquad \text{and} \qquad \|\mathbf{y} - \mathbf{s}^{(j)}\| < \tfrac{1}{2}$$

which implies that

$$\|\mathbf{y}^{(i)} - \mathbf{y}^{(j)}\| < 1.$$

In turn, this implies that

$$\rho(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}) < \rho(\mathbf{u}_1) \qquad \text{and} \qquad h(\mathbf{s}^{(i)} - \mathbf{s}^{(j)}) < h(\mathbf{s}^{(K)}).$$

Accordingly, $\mathcal{V}^{(K)}$ cannot be (ρ, h)-minimal. Therefore, the sets $\mathcal{S}\big(\mathbf{y}^{(j)}, \frac{1}{2}\big)$ must be disjoint. However, because

$$\rho_1(\mathbf{s}^{(j)}) \leqslant \rho_1(\mathbf{u}_n) \qquad \text{and} \qquad h_1(\mathbf{s}^{(j)}) \leqslant h_1(\mathbf{s}^{(K)})$$

we see that

$$\mathcal{S}\big(\mathbf{y}^{(j)}, \tfrac{1}{2}\big) \subset \mathcal{S}\left(\mathbf{0}, \frac{\rho_1(\mathbf{u}_n)}{\rho_1(\mathbf{u}_1)} + \tfrac{1}{2}\right). \tag{3.15}$$

Therefore, the number K is bounded above by the ratio of the volumes of the sets in (3.15). That is,

$$K \leqslant \frac{V\left(\frac{1}{2}+\rho_1(\mathbf{u}_n)/\rho_1(\mathbf{u}_1)\right)}{V\left(\frac{1}{2}\right)} = \left[1+2\frac{\rho_1(\mathbf{u}_n)}{\rho_1(\mathbf{u}_1)}\right]^n.$$

□

LEMMA 3.4. *Suppose $\mathcal{V}=\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ are linearly independent points in a lattice Ω of rank n in $\mathbb{R}^m$ which satisfy (2.2) and suppose ρ and h are strictly convex radius and height functions. Let*

$$\Theta_j = \left\{ \frac{1}{d}\sum_{i=1}^{j} a_i\mathbf{v}_i \;\Big|\; \sum_{i=1}^{j}|a_i| \leqslant |d|;\ 0<|a_j|<|d| \right\} \cap \Omega.$$

If, for all $j=1,2,\ldots,n$ and for all $\mathbf{w}\in\Theta_j$, either (2.3) or (2.4) is satisfied then the hyperoctahedral $\mathcal{O}(\mathcal{V})$ is perfect in Ω.

PROOF. Note that Θ_j represents the set of all lattice points which lie within or on the boundary of the hyperoctahedral of $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_j\}$, apart from the origin and the vertices themselves and apart from those points contained in the hyperoctahedral of $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{j-1}\}$. We will show that all the Θ_j are empty and therefore $\mathcal{O}(\mathcal{V})$ is perfect in Ω.

Suppose that Θ_j is not empty for some j. Therefore, there exists a non-zero lattice point $\mathbf{w}$ which can be expressed as

$$\mathbf{w} = \frac{a_1\mathbf{v}_1+a_2\mathbf{v}_2+\cdots+a_j\mathbf{v}_j}{d}. \tag{3.16}$$

Let $\{p_k\}$ be the set of indices such that $a_{p_k}\neq 0$ in ascending order and let s be the cardinality of this set. Hence, $p_s=j$. The case $s=1$ is trivial, so we will assume $s>1$. Applying the triangle inequality, we find that

$$\rho(\mathbf{w}) \leqslant \max_{1\leqslant i\leqslant s}\{\rho(\mathbf{v}_{p_i})\} \tag{3.17}$$

and

$$h(\mathbf{w}) \leqslant \max_{1\leqslant i\leqslant s}\{h(\mathbf{v}_{p_i})\}. \tag{3.18}$$

Suppose (3.17) is satisfied as an equality. In order for this to occur, we require that

$$\rho(\mathbf{v}_{p_1}) = \rho(\mathbf{v}_{p_2}) = \cdots = \rho(\mathbf{v}_{p_s}).$$

The strict convexity of ρ implies that we must also have

$$\rho(\mathbf{v}_{p_t}-\mathbf{v}_{p_u}) = 0 \tag{3.19}$$

for all t and u. Similarly, if (3.18) is to hold as an equality then

$$h(\mathbf{v}_{p_1}) = h(\mathbf{v}_{p_2}) = \cdots = h(\mathbf{v}_{p_s})$$

and

$$h(\mathbf{v}_{p_t} - \mathbf{v}_{p_u}) = 0 \tag{3.20}$$

for all t and u in $\{1, 2, \ldots, s\}$. Now, (3.19) and (3.20) cannot both be satisfied simultaneously because ρ and h are transverse and so at least one of the inequalities (3.17) and (3.18) must be satisfied strictly.

If (3.17) is satisfied strictly then (2.3) is not satisfied and so (2.4) must be satisfied instead. From (2.2), we conclude that $\rho(\mathbf{w}) < \rho(\mathbf{v}_{p_s})$. In order to satisfy (2.4), we must therefore have $h(\mathbf{w}) > h(\mathbf{v}_{p_s})$. However, upon applying the triangle inequality to (3.16), we find that there must be some $r < s$ such that $h(\mathbf{w}) < h(\mathbf{v}_{p_r})$. Hence, (2.4) cannot be satisfied.

If (3.18) is satisfied strictly then (2.4) is not satisfied and so (2.3) must be satisfied instead. There must be some $r \leqslant s$ such that $h(\mathbf{w}) < h(\mathbf{v}_{p_r})$. In order to satisfy (2.3), we must therefore have $\rho(\mathbf{w}) > \rho(\mathbf{v}_{p_r})$, but this implies that $\rho(\mathbf{w}) < \rho(\mathbf{v}_{p_s})$. Hence, (2.3) cannot be satisfied.

The hyperoctahedral of $\mathcal{V}$ must therefore be perfect in Ω and the lemma is proved. □

THEOREM 3.5. *If $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is (ρ, h)-minimal in a lattice Ω of rank n in $\mathbb{R}^m$ and ρ and h are strictly convex radius and height functions then the hyperoctahedral of $\mathcal{V}$ is perfect in Ω.*

PROOF. The proof follows directly from the application of Lemma 3.4. □

3.2. Properties in Lattices of Rank 2. In this subsection we explore the nature of (ρ, h)-minimal sets in lattices of rank 2. We show that the theory enables us to derive the algorithms of EUCLID and GAUSS for Diophantine approximation and lattice reduction, respectively, from a more general algorithm for producing (ρ, h)-minimal sets.

THEOREM 3.6. *If $(\mathbf{v}_1, \mathbf{v}_2)$ is (ρ, h)-minimal in a lattice Ω of rank 2 and ρ and h are strictly convex radius and height functions then $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis of Ω.*

PROOF. The proof is a direct consequence of Theorem 2.5 and Theorem 3.5. □

LEMMA 3.5. *Suppose $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k\}$, $k \geqslant 2$, are points in a lattice Ω and ρ and h are radius and height functions. If $\mathbf{w} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ with $a_1, a_2 \in \mathbb{Z}$ then, with $\mathbf{v} = \operatorname{sgn}(a_1)\mathbf{b}_1 + \operatorname{sgn}(a_2)\mathbf{b}_2$, we find that*

$$\rho h(\mathbf{w}) < \max_{1 \leqslant i \leqslant k} \{\rho h(\mathbf{b}_i)\} \;\Rightarrow\; \rho h(\mathbf{v}) < \max_{1 \leqslant i \leqslant k} \{\rho h(\mathbf{b}_i)\} \tag{3.21}$$

and

$$h\rho(\mathbf{w}) < \max_{1 \leqslant i \leqslant k} \{h\rho(\mathbf{b}_i)\} \;\Rightarrow\; h\rho(\mathbf{v}) < \max_{1 \leqslant i \leqslant k} \{h\rho(\mathbf{b}_i)\}. \tag{3.22}$$

PROOF. We will only show that (3.21) must be satisfied, since (3.22) follows by symmetry. If $a_1 = 0$ or $a_2 = 0$ then (3.21) is satisfied trivially since $\rho h(\mathbf{v}) \leqslant \rho h(\mathbf{w})$. Otherwise, we can write the identity

$$\mathbf{v} = \frac{\operatorname{sgn}(a_1)(|a_2| - 1)\mathbf{b}_1 + \operatorname{sgn}(a_2)(|a_1| - 1)\mathbf{b}_2 + \mathbf{w}}{|a_1| + |a_2| - 1}.$$

Using the triangle inequality, we find that

$$\rho h(\mathbf{v}) \leqslant \frac{(|a_2| - 1)\rho h(\mathbf{b}_1) + (|a_1| - 1)\rho h(\mathbf{b}_2) + \rho h(\mathbf{w})}{|a_1| + |a_2| - 1}.$$

Thus $\rho h(\mathbf{v}) \leqslant \max\{\rho h(\mathbf{b}_1), \rho h(\mathbf{b}_2), \rho h(\mathbf{w})\}$. If $|a_1| = |a_2| = 1$ or $\rho h(\mathbf{v}) \leqslant \rho h(\mathbf{w})$ then (3.21) is obviously satisfied. Otherwise, we have $\rho h(\mathbf{v}) > \rho h(\mathbf{w})$ and either $|a_1| > 1$ or $|a_2| > 1$, which implies that $\rho h(\mathbf{v}) < \max\{\rho h(\mathbf{b}_1), \rho h(\mathbf{b}_2)\}$. Again (3.21) is satisfied and so the lemma is proved. □

We will make use of the set Λ_2 defined in terms of an ordered set of lattice points $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)$ so that

$$\Lambda_2(\mathcal{B}) = \{\pm\mathbf{b}_1 \pm \mathbf{b}_2\}. \tag{3.23}$$

Once again, the use of $\pm$ is used as a shorthand to (recursively) indicate that the expression following, and its additive inverse, are elements of the set.

THEOREM 3.7. *Suppose $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2)$ is an ordered basis of a lattice Ω of rank 2 with $\rho h(\mathbf{b}_1) \leqslant \rho h(\mathbf{b}_2)$. If, for each $\mathbf{w} \in \Lambda_2(\mathcal{B})$, it is true that either*

$$\rho h(\mathbf{w}) \geqslant \max\{\rho h(\mathbf{b}_1), \rho h(\mathbf{b}_2)\} \tag{3.24}$$

or

$$h\rho(\mathbf{w}) \geqslant \max\{h\rho(\mathbf{b}_1), h\rho(\mathbf{b}_2)\} \tag{3.25}$$

then $(\mathbf{b}_1, \mathbf{b}_2)$ is (ρ, h)-minimal.

PROOF. Suppose $(\mathbf{b}_1, \mathbf{b}_2)$ is not (ρ, h)-minimal. Then either there exists some non-zero integer multiple $\mathbf{w}$ of $\mathbf{b}_1$ which satisfies $\rho h(\mathbf{w}) < \rho h(\mathbf{b}_1)$, which is obviously impossible, or there exists some $\mathbf{s} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ with $a_1, a_2 \in \mathbb{Z}$ and $a_2 \neq 0$ such that $\mathbf{w} = \mathbf{s}$ fails to satisfy both (3.24) and (3.25). Consider the implications of Lemma 3.5. If $a_1 = 0$ then $\mathbf{w} = \mathbf{b}_2$ fails to satisfy both (3.24) and (3.25), which is impossible, and if $a_1 \neq 0$ then there exists $\mathbf{w} \in \Lambda_2(\mathcal{B})$ which fails similarly, contrary to our assumption. Therefore the theorem is proved. □

THEOREM 3.8. *Suppose $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2)$ is (ρ, h)-minimal in a lattice Ω of rank 2 and $\mathcal{U}$ is also a basis of Ω. If $\Xi(\mathcal{U})$ is not empty then $\mathcal{U}$ has an incremental successor in which the innovation is an element of $\Lambda_2(\mathcal{U})$.*

PROOF. If $\Xi(\mathcal{U})$ is not empty then there must be a point $\mathbf{w} \in \Omega$, which we can write as $\mathbf{w} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2$ with $a_1, a_2 \in \mathbb{Z}$ and neither equal to zero, such that $\rho h(\mathbf{w}) < \rho h(\mathbf{u}_2)$. Lemma 3.5 implies that there exists $\mathbf{t} \in \Lambda_2(\mathcal{U})$ such that $\rho h(\mathbf{t}) < \rho h(\mathbf{u}_2)$ and so $\Lambda_2(\mathcal{U}) \cap \Xi(\mathcal{U})$ is not empty. Choose $\mathbf{s} \in \Lambda_2(\mathcal{U}) \cap \Xi(\mathcal{U})$ such that $h\rho(\mathbf{s})$ is minimised on this set. We will show that $\mathbf{s}$ is an innovation in an incremental successor to $\mathcal{U}$.

Suppose $\mathbf{s}$ cannot be an innovation. Then there exists a point $\mathbf{x} \in \Xi(\mathcal{U})$ with $h\rho(\mathbf{x}) < h\rho(\mathbf{s})$. This implies that

$$\rho h(\mathbf{x}) < \max\{\rho h(\mathbf{u}_1), \rho h(\mathbf{u}_2), \rho h(\mathbf{s})\} = \rho h(\mathbf{u}_2)$$

and

$$h\rho(\mathbf{x}) < \max\{h\rho(\mathbf{u}_1), h\rho(\mathbf{u}_2), h\rho(\mathbf{s})\} = h\rho(\mathbf{s}).$$

Hence, Lemma 3.5 implies that there must exist an element of $\Lambda_2(\mathcal{U})$ which satisfies these conditions also since $\pm\mathbf{u}_1$ and $\pm\mathbf{u}_2$ do not belong to $\Xi(\mathcal{U})$, and this furnishes a contradiction, completing the proof. □

We are now able to formulate a simple algorithm for producing sequences of (ρ, h)-minimal sets in a lattice of rank 2. For this algorithm, and also for Algorithm 4.1 and Algorithm 5.1, we define the Boolean function (or parameterised relation) $L(\cdot)$ as

$$L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) = (\rho h(\mathbf{x}) < \rho h(\mathbf{z}) \ \wedge\ \rho h(\mathbf{y}) < \rho h(\mathbf{z})) \ \wedge\ h\rho(\mathbf{x}) < h\rho(\mathbf{y}) \\ \vee\ \neg[\rho h(\mathbf{x}) < \rho h(\mathbf{z}) \ \wedge\ \rho h(\mathbf{y}) < \rho h(\mathbf{z})] \ \wedge\ \rho h(\mathbf{x}) < \rho h(\mathbf{y}) \tag{3.26}$$

with $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^m$. To give a geometric interpretation to $L(\cdot)$, we can think of $L(\cdot)$ being **true** if $\mathbf{x}$ is "better" than $\mathbf{y}$ with respect to $\mathbf{z}$. The set of points $\mathbf{w} \in \mathbb{R}^m$ with $\rho h(\mathbf{w}) < \rho h(\mathbf{z})$ is an open cylinder (if ρ is an extended semi-norm, otherwise it is a convex body) in $\mathbb{R}^m$ along with a part of its surface (for instance, where $h(\mathbf{w}) < h(\mathbf{z})$). If both $\mathbf{x}$ and $\mathbf{y}$ lie inside the cylinder, the function $L(\cdot)$ determines that $\mathbf{x}$ is "better" than $\mathbf{y}$ if $\mathbf{x}$ has the lesser height, or if the heights are equal, if $\mathbf{x}$ has the lesser radius. If either $\mathbf{x}$ or $\mathbf{y}$ lie outside the cylinder then $\mathbf{x}$ is "better" than $\mathbf{y}$ if $\mathbf{x}$ has the lesser radius, or if they are equal, the lesser height.

ALGORITHM 3.1.

```
1 begin
2   if L(b2, b1; b2) then swap(b1, b2) fi;
3   while ρ(b1) > ε ∧ [L(b2 + b1, b2; b2) ∨ L(b2 − b1, b2; b2)] do
4         if L(b2 − b1, b2 + b1; b2) then
5           b2 := b2 − b1
6         else
7           b2 := b2 + b1
```

```
8          fi;
9          output(b_2);
10         if L(b_2, b_1; b_2) then swap(b_1, b_2) fi;
11     od
12 end;
```

We denote by $\mathcal{B}$ the ordered set of values of the variables $(\mathbf{b}_1, \mathbf{b}_2)$ at a specified line and iteration in Algorithm 3.1. We will use $\mathcal{B}'$ to denote the new value of $\mathcal{B}$ at line 3 on the subsequent iteration of this algorithm.

If Algorithm 3.1 is executed with $\mathcal{B}$ initially set as a basis of a lattice Ω then $\mathcal{B}$ will continue to be a basis of this lattice throughout every iteration of the algorithm.

THEOREM 3.9. *Suppose that $\mathcal{B}^{(0)}$ is a basis of a lattice Ω of rank 2 and ρ and h are radius and height functions. If Algorithm 3.1 is executed with $\mathcal{B}^{(0)}$ as its initial basis then, after a finite number of steps, $\mathcal{B}$ will be a (ρ, h)-minimal set of Ω at line 3 or the algorithm will terminate with $\rho(\mathbf{b}_1) \leqslant \epsilon$.*

PROOF. If, at line 3, $\mathcal{B}$ is not (ρ, h)-minimal then, by definition, there must exist an element $\mathbf{w} \in \Omega$, linearly independent of $\{\mathbf{b}_1\}$, such that

$$\rho h(\mathbf{w}) < \max\{\rho h(\mathbf{b}_1), \rho h(\mathbf{b}_2)\}$$

and

$$h\rho(\mathbf{w}) < \max\{h\rho(\mathbf{b}_1), h\rho(\mathbf{b}_2)\}.$$

Theorem 3.7 implies that if there is an element $\mathbf{w}$ which satisfies these two inequalities then there exists a point in $\Lambda_2(\mathcal{B})$ which does also. There are only a finite number of lattice points which can satisfy these criteria and the number must decrease from one iteration to the next. Therefore, a (ρ, h)-minimal set must be produced after a finite number of steps unless the algorithm terminates in the meantime with $\rho(\mathbf{b}_1) \leqslant \epsilon$. □

THEOREM 3.10. *If, at line 3 of Algorithm 3.1, $\mathcal{B}$ is (ρ, h)-minimal in a lattice Ω of rank 2 and $\mathcal{B}$ has a strict successor then, on the subsequent iteration, $\mathcal{B}'$ will be an incremental successor to $\mathcal{B}$ or the algorithm terminates with $\rho(\mathbf{b}_1) \leqslant \epsilon$.*

PROOF. If $\mathcal{B}$ has a strict successor then $\Xi(\mathcal{B})$ must be non-empty. Theorem 3.8 can then be applied to show that the algorithm must indeed produce an incremental successor to $\mathcal{B}$. □

REMARK 3.1. We note that if, at line 3 of Algorithm 3.1, $\mathcal{B}$ is (ρ, h)-minimal, then Algorithm 3.1 will behave like the algorithm described in Theorem 3.3 to the extent that if $\mathbf{p}$ is a best approximation such that

$$\epsilon \leqslant \rho(\mathbf{p}) \leqslant \rho(\mathbf{b}_1)$$

then there exists an equivalent best approximation $\mathbf{q}$ which will be output by the algorithm after a finite number of steps.

REMARK 3.2. Algorithm 3.1 reverts to an additive version of the Euclid's algorithm (Algorithm 3.2 of Chapter 2) if we set $\Omega = \mathbb{Z}^2$ and, with $\mathbf{x} = (p, q) \in \mathbb{Z}^2$, we set $\rho(\mathbf{x}) = |p\alpha - q|$ for some $\alpha \in \mathbb{R}$, $\alpha > 0$, and $h(\mathbf{x}) = |q|$ and we execute the algorithm with an initial basis consisting of $(1, 0)$ and $(0, 1)$. By "additive," we mean that, rather than computing partial quotients by division, we use repeated addition (subtraction), as in our interpretation of an "authentic" version of Euclid's algorithm in Algorithm 3.1 of Chapter 2. The outputs of Algorithm 3.1 can be interpreted as the intermediate fractions, including convergents, of the simple continued fraction expansion of α.

REMARK 3.3. Algorithm 3.1 behaves like an additive form of Gauss' algorithm for lattice reduction (see Algorithm 6.1 of Chapter 3) if we set $\rho = h$ to be a norm (or extended norm).

3.3. Properties in Lattices of Rank 3. We now examine the properties of (ρ, h)-minimal sets in lattices of rank 3. This will allow us to formulate an algorithm in the next section which is able to find best simultaneous Diophantine approximations for such lattices under appropriate conditions on the radius and height functions.

One of these conditions is that the radius and height functions are complementary on the real span of the lattice. Recalling Definition 2.9, the implication in this instance is that the radius and height functions can be expressed as

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|^* \qquad \text{and} \qquad h(\mathbf{x}) = \left\|\mathbf{R}^T\mathbf{x}\right\|^\dagger$$

where $\|\cdot\|^*$ and $\|\cdot\|^\dagger$ extended norms and $\mathbf{P}^T$ and $\mathbf{R}^T$ are matrices. Furthermore, the implication is that there are n_1 columns of $\mathbf{P}$ and n_2 columns of $\mathbf{R}$ which lie in $\mathcal{E}$, where we use $\mathcal{E}$ to represent the real span of the lattice, and $n_1 + n_2 = 3$. This means that either $n_1 = 1$ and $n_2 = 2$ or $n_1 = 2$ and $n_2 = 1$. The only norm in one dimension is the absolute value of its argument, up to scaling, and, similarly, this is the only extended norm in one dimension, up to scaling and multiplicity. By this, we mean that that extended norms in one dimension all have the form $\|x\| = (\lambda_1|x|, \lambda_2|x|, \ldots, \lambda_q|x|)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_q > 0$. Therefore, we assume that either $\rho(\mathbf{x})$ or $h(\mathbf{x})$ has the form $|\boldsymbol{\alpha} \cdot \mathbf{x}|$ for some vector $\boldsymbol{\alpha} \in \mathcal{E}$ whenever $\mathbf{x} \in \mathcal{E}$.

LEMMA 3.6. *Suppose* $x_1, x_2, x_3 \in \mathbb{R}$ *with*

$$|x_1| \leqslant |x_2| \leqslant |x_3|$$

and $|x_3| > 0$. *Let*

$$\mathcal{H} = \left\{x = \tfrac{1}{2}(x_1 \pm x_2 \pm x_3) \;\middle|\; |x| \leqslant |x_3|\right\}$$

and

$$\mathcal{H}' = \left\{x = \tfrac{1}{2}(x_1 \pm x_2 \pm x_3) \;\mid\; |x| < |x_3|\right\}$$

Now, $|\mathcal{H}| \geqslant 3$ *and* $|\mathcal{H}'| \geqslant 2$. *Furthermore, if* $|\mathcal{H}| = 3$ *then* $|\mathcal{H}'| = 3$ *and if* $|\mathcal{H}| = 4$ *and* $|\mathcal{H}'| = 2$ *then* $x_1 = 0$ *and* $|x_2| = |x_3|$.

Proof. Without loss of generality, suppose that x_1, x_2 and x_3 are all non-negative. We can quickly confirm that

$$\tfrac{1}{2}\max\left\{|x_1 + x_2 - x_3|, |x_1 - x_2 + x_3|, |x_1 - x_2 - x_3|\right\} \leqslant |x_3| \tag{3.27}$$

and so $|\mathcal{H}| \geqslant 3$. Furthermore,

$$\tfrac{1}{2}\max\left\{|x_1 + x_2 - x_3|, |x_1 - x_2 + x_3|\right\} < |x_3|, \tag{3.28}$$

which implies that $|\mathcal{H}'| \geqslant 2$.

Now, $|\mathcal{H}'| \leqslant |\mathcal{H}|$. Suppose $|\mathcal{H}| \geqslant 3$ and $|\mathcal{H}'| = 2$. In this case, (3.27) and (3.28) imply that $\frac{1}{2}|x_1 - x_2 - x_3| = |x_3|$, which in turn implies that $x_1 = 0$ and $|x_2| = |x_3|$. Hence $\frac{1}{2}|x_1 + x_2 + x_3| = |x_3|$, which implies that $|\mathcal{H}| = 4$. □

Lemma 3.7. *Suppose* $\|\cdot\|$ *is a strictly convex extended norm on* $\mathbb{R}^2$ *and suppose* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^2$ *satisfy*

$$\|\mathbf{x}_1\| \leqslant \|\mathbf{x}_2\| \leqslant \|\mathbf{x}_3\|$$

with $\|\mathbf{x}_2\| > 0$ *and* $\mathbf{x}_1$, $\mathbf{x}_2$ *and* $\mathbf{x}_3$ *are not colinear with origin. Let*

$$\mathcal{R} = \left\{\mathbf{x} = \tfrac{1}{2}(\mathbf{x}_1 \pm \mathbf{x}_2 \pm \mathbf{x}_3) \;\mid\; \|\mathbf{x}\| < \|\mathbf{x}_3\|\right\}$$

and

$$\mathcal{R}' = \left\{\mathbf{x} = \tfrac{1}{2}(\mathbf{x}_1 \pm \mathbf{x}_2 \pm \mathbf{x}_3) \;\mid\; \|\mathbf{x}\| < \|\mathbf{x}_2\|\right\}$$

Now, $|\mathcal{R}| \geqslant 2$. *If* $|\mathcal{R}'| = 0$ *then* $|\mathcal{R}| \geqslant 3$.

Proof. It is always possible to reorder and relabel $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ as $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$ so that

$$\mathbf{y}_3 = \xi\mathbf{y}_1 + \eta\mathbf{y}_2$$

with

$$\max\left\{|\xi|, |\eta|\right\} \leqslant 1$$

and

$$\|\mathbf{y}_1\| \leqslant \|\mathbf{y}_2\|.$$

Without loss of generality, we assume that $0 \leqslant \xi \leqslant \eta \leqslant 1$ (the other cases can be proved by symmetry). Now,

$$\begin{aligned} \tfrac{1}{2}\left\|\mathbf{y}_3 - \mathbf{y}_1 - \mathbf{y}_2\right\| &= \tfrac{1}{2}\left\|(\xi-1)\mathbf{y}_1 + (\eta-1)\mathbf{y}_2\right\| \\ &\leqslant \tfrac{1}{2}(1-\xi)\left\|\mathbf{y}_1\right\| + \tfrac{1}{2}(1-\eta)\left\|\mathbf{y}_2\right\| \\ &\leqslant \left\|\mathbf{y}_2\right\|. \end{aligned} \tag{3.29}$$

At least one of the last two inequalities in (3.29) must be strict. If $\|\mathbf{y}_1\| > 0$ then it will be the former inequality which is satisfied strictly because of the strict convexity of $\|\cdot\|$, otherwise the latter. Also,

$$\begin{aligned} \tfrac{1}{2}\left\|\mathbf{y}_3 + \mathbf{y}_1 - \mathbf{y}_2\right\| &= \tfrac{1}{2}\left\|(\xi+1)\mathbf{y}_1 + (\eta-1)\mathbf{y}_2\right\| \\ &\leqslant \tfrac{1}{2}(\xi+1)\left\|\mathbf{y}_1\right\| + \tfrac{1}{2}(1-\eta)\left\|\mathbf{y}_2\right\| \\ &\leqslant \left\|\mathbf{y}_2\right\|. \end{aligned} \tag{3.30}$$

Again, one of the last two inequalities in (3.30) must be strict. Thus, $|\mathcal{R}| \geqslant 2$. If

$$(1-\xi)\left\|\mathbf{y}_1\right\| \geqslant (1-\eta)\left\|\mathbf{y}_2\right\|$$

then, from (3.29),

$$\tfrac{1}{2}\left\|\mathbf{y}_3 - \mathbf{y}_1 - \mathbf{y}_2\right\| < \left\|\mathbf{y}_1\right\|$$

and so $|\mathcal{R}'| > 0$. Otherwise,

$$\begin{aligned} \tfrac{1}{2}\left\|\mathbf{y}_3 + \mathbf{y}_1 - \mathbf{y}_2\right\| &< \tfrac{1}{2}(\xi+1)\left\|\mathbf{y}_1\right\| + \tfrac{1}{2}(1-\eta)\left\|\mathbf{y}_2\right\| \\ &= \tfrac{1}{2}(1-\xi)\left\|\mathbf{y}_1\right\| + \xi\left\|\mathbf{y}_1\right\| + \tfrac{1}{2}(1-\eta)\left\|\mathbf{y}_2\right\| \\ &\leqslant (1-\eta)\left\|\mathbf{y}_2\right\| + \xi\left\|\mathbf{y}_1\right\| \\ &\leqslant \left\|\mathbf{y}_2\right\| \end{aligned}$$

and, again, one of the last two inequalities must be strict. So, $|\mathcal{R}| \geqslant 3$. □

LEMMA 3.8. *Suppose* $\mathbf{v}_1$, $\mathbf{v}_2$ *and* $\mathbf{v}_3$ *are points in a lattice* Ω *of rank* 3 *in* $\mathbb{R}^m$ *with*

$$\rho h(\mathbf{v}_1) \leqslant \rho h(\mathbf{v}_2) \leqslant \rho h(\mathbf{v}_3)$$

and the octahedral of $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ *is perfect in* Ω. *Suppose* ρ *and* h *are strictly convex, complementary radius and height functions. Let*

$$\mathcal{S} = \left\{\tfrac{1}{2}(\mathbf{v}_1 \pm \mathbf{v}_2 \pm \mathbf{v}_3)\right\} \cap \Omega.$$

If, for each $\mathbf{w} \in \mathcal{S}$, *the inequalities* (2.3) *and* (2.4) *are satisfied then* $\mathcal{V}$ *forms a basis of* Ω.

PROOF. If $\mathcal{S}$ is empty then we can apply Theorem 2.6 of Chapter 3 to show that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ forms a basis of Ω. Therefore, we will suppose that $\mathcal{S}$ is non-empty and demonstrate a contradiction. If $\mathcal{S}$ is non-empty then it must contain all four possible elements.

Let p, q and r be distinct indices such that

$$h\rho(\mathbf{v}_p) \leqslant h\rho(\mathbf{v}_q) \leqslant h\rho(\mathbf{v}_r).$$

As we discussed at the beginning of this subsection, for all $\mathbf{x}$ which lie in the real span of Ω, we can express ρ and h so that $\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|^*$ and $h(\mathbf{x}) = \left\|\mathbf{R}^T\mathbf{x}\right\|^\dagger$ where $\|\cdot\|^*$ and $\|\cdot\|^\dagger$ are extended norms and $\mathbf{P}^T$ is an $n_1 \times m$ matrix and $\mathbf{R}^T$ is an $n_2 \times m$ matrix, both having full column rank, where either

$$n_1 = 2 \qquad \text{and} \qquad n_2 = 1$$

or

$$n_1 = 1 \qquad \text{and} \qquad n_2 = 2.$$

Suppose we have $n_1 = 2$ and $n_2 = 1$. Now, $\rho(\mathbf{v}_2) > 0$ and $h(\mathbf{v}_r) > 0$, otherwise $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are not linearly independent. Applying Lemma 3.6, we can see that there exist at least three elements of $\mathcal{S}$ such that

$$h(\mathbf{w}) \leqslant h(\mathbf{v}_r). \tag{3.31}$$

Applying Lemma 3.7, we can see that there are at least two elements of $\mathcal{S}$ such that

$$\rho(\mathbf{w}) < \rho(\mathbf{v}_3). \tag{3.32}$$

Therefore, there must be at least one element of $\mathcal{S}$, say $\mathbf{w}_1^*$, for which both (3.31) and (3.32) are simultaneously satisfied. Thus, $\rho h(\mathbf{w}_1^*) < \rho h(\mathbf{v}_3)$, so failing to satisfy (2.3). If $\mathbf{w}_1^*$ satisfies (3.31) strictly then we fail to satisfy (2.4) also. Hence, suppose that $\mathbf{w}_1^*$ satisfies (3.31) as an equality while also satisfying (3.32). Lemma 3.6 implies that all elements of $\mathcal{S}$ obey (3.31). Therefore, there must exist another element of $\mathcal{S}$, say $\mathbf{w}_2^*$, which obeys (3.31) as an equality while also satisfying (3.32). Furthermore, Lemma 3.6 implies that $h(\mathbf{v}_p) = 0$ and $h(\mathbf{v}_q) = h(\mathbf{v}_r)$. Lemma 3.7 implies that either there is a third element of $\mathcal{S}$ which satisfies (3.32) or there is an element $\mathbf{w} \in \mathcal{S}$ such that

$$\rho(\mathbf{w}) < \rho(\mathbf{v}_2). \tag{3.33}$$

In the former case, we immediately have a contradiction since there must be at least one element of $\mathcal{S}$ for which (3.31) is satisfied strictly and (3.32) is satisfied also. In the latter case, that element which satisfies (3.33) must be $\mathbf{w}_1^*$ or $\mathbf{w}_2^*$, say $\mathbf{w}_1^*$. However, this contradicts (2.4) for $\mathbf{w} = \mathbf{w}_1^*$, because

$$\max_{i=2,3} \{h(\mathbf{v}_i)\} = h(\mathbf{v}_r)$$

and so either $h\rho(\mathbf{w}_1^*) < h\rho(\mathbf{v}_2)$ or $h\rho(\mathbf{w}_1^*) < h\rho(\mathbf{v}_3)$. Hence, the lemma is satisfied when $n_1 = 2$ and $n_2 = 1$.

A symmetric argument can be employed to show that the lemma also holds if $n_1 = 1$ and $n_2 = 2$ instead. The symmetry occurs because the conditions (2.3)

and (2.4) for (ρ, h)-minimal sets are symmetric in this case, since all elements of $\mathcal{S}$ are linearly independent of $\{\mathbf{v}_1, \mathbf{v}_2\}$. □

THEOREM 3.11. *If $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ is (ρ, h)-minimal in a lattice Ω of rank 3 in $\mathbb{R}^m$ and ρ and h are strictly convex, complementary radius and height functions then $\mathcal{V}$ forms a basis of Ω.*

PROOF. We use Theorem 3.5 to show that the octahedral of $\mathcal{V}$ is perfect in Ω. We then use Lemma 3.8 to show that $\mathcal{V}$ must be a basis of Ω. □

In addition to the set Λ_2 which we defined in (3.23), we will make use of the set Λ_3, defined in terms of an ordered set of lattice points $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n)$ so that

$$\Lambda_3(\mathcal{B}) = \{a_1\mathbf{b}_1 + a_2\mathbf{b}_2 \pm \mathbf{b}_3 \mid a_1, a_2 \in \mathbb{Z}\}.$$

THEOREM 3.12. *Suppose $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ is a basis of a lattice Ω of rank 3 with $\rho h(\mathbf{b}_1) \leqslant \rho h(\mathbf{b}_2) \leqslant \rho h(\mathbf{b}_3)$. If ρ and h are strictly convex, complementary radius and height functions and, for each $\mathbf{w} \in \Lambda_j(\mathcal{B})$, $j = 2, 3$, it is true that either*

$$\rho h(\mathbf{w}) \geqslant \max_{1 \leqslant i \leqslant j} \{\rho h(\mathbf{b}_i)\} \tag{3.34}$$

or

$$h\rho(\mathbf{w}) \geqslant \max_{1 \leqslant i \leqslant 3} \{h\rho(\mathbf{b}_i)\} \tag{3.35}$$

then $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ is (ρ, h)-minimal.

PROOF. Clearly, (3.34) and (3.35) are simply (2.3) and (2.4) from Definition 2.10 with $\mathbf{b}_i$ substituted for $\mathbf{v}_i$.

Suppose $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ is not (ρ, h)-minimal. Then there exists some non-zero $\mathbf{w} \in \Omega \setminus \Lambda_2 \setminus \Lambda_3$ which is linearly independent of $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1}\}$ for some $j = 1, 2, 3$ which satisfies neither (3.34) nor (3.35). Now, for any such $\mathbf{w}$ there exists a maximum index k such that $\mathbf{w}$ is linearly independent of $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{k-1}\}$.

The case $k = 1$ is trivial since $\mathbf{w}$ must have the form $\mathbf{w} = a_1\mathbf{b}_1$ for some non-zero $a_1 \in \mathbb{Z}$. Clearly, (3.34) and (3.35) must both be satisfied for $j = 1$.

Suppose $k = 2$. In this case, there must exist some $S = a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ with $a_1, a_2 \in \mathbb{Z}$ and $a_2 \neq 0$ such that neither of the inequalities (3.34) nor (3.35) are satisfied with $\mathbf{w} = \mathbf{s}$ and $j = 2$. As in the proof of Theorem 3.7, we can apply Lemma 3.5 to show that either $\mathbf{w} = \mathbf{b}_2$ fails to satisfy these inequalities, which is impossible, or there exists some $\mathbf{w} \in \Lambda_2(\mathcal{B})$ which fails to satisfy them, contrary to our assumption.

Suppose $k = 3$. Let $\mathbf{s} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2 + a_3\mathbf{b}_3$ be an element of $\Omega \setminus \Lambda_3$ with smallest absolute value for a_3 which does not satisfy either (3.34) or (3.35). By assumption, $|a_3| > 1$. We will now show that the octahedral of $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{s}\}$ is perfect in Ω. Suppose it isn't so that there exists some non-zero lattice point $\mathbf{t}$, distinct from the vertices, within or on the boundary of the octahedral. We know that $\mathbf{t}$ must be linearly

independent of $\mathbf{b}_1$ and $\mathbf{b}_2$ because $\{\mathbf{b}_1, \mathbf{b}_2\}$ is a primitive basis of Ω. Furthermore, if $\mathbf{t}$ is expressed as a linear combination of $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ then the coefficient of $\mathbf{b}_3$ has smaller absolute value than a_3. Thus, by assumption, $\mathbf{t}$ must satisfy (3.34) or (3.35). If $\mathbf{t}$ satisfies (3.34) but $\rho h(\mathbf{t}) < \rho h(\mathbf{s})$ then $\mathbf{s}$ also satisfies (3.34), contrary to our assumption. Similarly, if $\mathbf{t}$ satisfies (3.35) but

$$h\rho(\mathbf{t}) < h\rho(\mathbf{s})$$

then $\mathbf{s}$ also satisfies (3.35). Thus, either

$$\rho h(\mathbf{t}) \leqslant \rho h(\mathbf{b}_1) \text{ and } \rho h(\mathbf{t}) \leqslant \rho h(\mathbf{b}_2) \text{ and } \rho h(\mathbf{t}) \leqslant \rho h(\mathbf{s})$$

or

$$h\rho(\mathbf{t}) \leqslant h\rho(\mathbf{b}_1) \text{ and } h\rho(\mathbf{t}) \leqslant h\rho(\mathbf{b}_2) \text{ and } h\rho(\mathbf{t}) \leqslant h\rho(\mathbf{s}).$$

Relabelling and reordering $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{s}$ as $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ to ensure that (2.2) is satisfied, we can see that Lemma 3.4 can be applied to show that the octahedral of $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{s}\}$ is perfect. Because ρ and h are strictly convex, complementary radius and height functions, we can use Lemma 3.8 to show that $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{s}\}$ forms a basis of Ω. However, this implies that $|a_3| = 1$, contrary to our assumption. □

THEOREM 3.13. *If $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ is (ρ, h)-minimal in a lattice Ω of rank 3, ρ and h are strictly convex, complementary radius and height functions and ρ is not null-spanned by $\mathcal{U}$ then there exists an incremental successor to $\mathcal{U}$ in which the innovation is an element of $\Lambda_2(\mathcal{U}) \cup \Lambda_3(\mathcal{U})$.*

PROOF. If ρ is not null-spanned by $\mathcal{U}$ then $\Xi(\mathcal{U})$ is not empty, by Lemma 3.3. Thus, Theorem 3.2 can be applied to show that an incremental successor, $\mathcal{V}$, to $\mathcal{U}$ exists. It remains to show that the innovation, $\mathbf{s}$, is an element of $\Lambda_2(\mathcal{U}) \cup \Lambda_3(\mathcal{U})$.

Consider the index q which is defined, as in Theorem 3.2, as the minimum index for which $\mathbf{s}$ is linearly dependent on $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q\}$. That is, q is the index of the inveteration from $\mathcal{U}$.

It is impossible that $q = 1$ since this implies that $\mathbf{s} = a_1\mathbf{b}_1$, $a_1 \neq 0$ and $\mathbf{s} \in \Xi_1(\mathcal{U})$.

Suppose $q = 2$, in which case $\mathcal{V}$ consists of some arrangement of $\mathbf{u}_1$, $\mathbf{u}_3$ and $\mathbf{s}$. From the definition of q, we conclude that there exists $\mathbf{w} \in \Xi_2(\mathcal{U})$ which can be expressed as $\mathbf{w} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2$ with $a_1, a_2 \in \mathbb{Z}$ and neither equal to zero. Lemma 3.5 implies that that there exists an element of $\Lambda_2(\mathcal{U})$ in $\Xi_2(\mathcal{U})$ because $\mathbf{w} \in \Xi_2(\mathcal{U})$. We choose $\mathbf{s} \in \Lambda_2(\mathcal{U}) \cap \Xi_2(\mathcal{U})$ so that $h\rho(\mathbf{s})$ is minimised in this set. As in the proof of Theorem 3.8, we find that $\mathbf{s}$ satisfies the conditions of Theorem 3.2, since we can show that $h\rho(\mathbf{w}) \geqslant h\rho(\mathbf{s})$ using Lemma 3.5 for any choice of $\mathbf{w}$ as defined above and so $\mathbf{s}$ must be an innovation of an incremental successor to $\mathcal{U}$.

If $q = 3$ then $\mathcal{V}$ consists of some arrangement of $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{s}$. Theorem 3.11 implies that $\mathcal{V}$ must form a basis of Ω. Thus, $\mathbf{s}$ must complete a basis with $\mathbf{u}_1$ and $\mathbf{u}_2$, and so $\mathbf{s} \in \Lambda_3(\mathcal{U})$. □

4. An Additive Algorithm for Lattices of Rank 3

4.1. The Additive Algorithm. Before describing the algorithm of this section, we observe some properties of $L(\cdot)$ and $\neg L(\cdot)$, as defined in (3.26), in a system for simultaneous Diophantine approximation. For all $\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^m$, $L(\cdot)$ has properties which we call

REFLEXIVE:	$\neg L(\mathbf{x}, \mathbf{x};\ \mathbf{z})$,
ANTISYMMETRIC:	$L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) \ \Rightarrow\ \neg L(\mathbf{y}, \mathbf{x};\ \mathbf{z})$,
EQUIVALENCE:	$\neg L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) \ \wedge\ \neg L(\mathbf{y}, \mathbf{x};\ \mathbf{z})$
	$\quad \Leftrightarrow\ \rho(\mathbf{x}) = \rho(\mathbf{y}) \ \wedge\ h(\mathbf{x}) = h(\mathbf{y})$,
TRANSITIVE:	$L(\mathbf{w}, \mathbf{x};\ \mathbf{z}) \ \wedge\ L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) \ \Rightarrow\ L(\mathbf{w}, \mathbf{y};\ \mathbf{z})$,
	$\neg L(\mathbf{x}, \mathbf{w};\ \mathbf{z}) \ \wedge\ L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) \ \Rightarrow\ L(\mathbf{w}, \mathbf{y};\ \mathbf{z})$,
	$L(\mathbf{w}, \mathbf{x};\ \mathbf{z}) \ \wedge\ \neg L(\mathbf{y}, \mathbf{x};\ \mathbf{z}) \ \Rightarrow\ L(\mathbf{w}, \mathbf{y};\ \mathbf{z})$,
	$\neg L(\mathbf{w}, \mathbf{x};\ \mathbf{z}) \ \wedge\ \neg L(\mathbf{x}, \mathbf{y};\ \mathbf{z}) \ \Rightarrow\ \neg L(\mathbf{w}, \mathbf{y};\ \mathbf{z})$,
UNIMODAL:	$L(\mathbf{x} - \mathbf{y}, \mathbf{x};\ \mathbf{z}) \ \Rightarrow\ L(\mathbf{x}, \mathbf{x} + \mathbf{y};\ \mathbf{z})$,
	$\neg L(\mathbf{x}, \mathbf{x} + \mathbf{y};\ \mathbf{z}) \ \Rightarrow\ \neg L(\mathbf{x} - \mathbf{y}, \mathbf{x};\ \mathbf{z})$.

PROOF. All but the last property can be easily proved in sequence. We will prove the unimodal property by contradiction. Observing the trivial identity

$$\mathbf{x} = \tfrac{1}{2}(\mathbf{x} + \mathbf{y}) + \tfrac{1}{2}(\mathbf{x} - \mathbf{y})$$

we conclude that

$$\rho h(\mathbf{x}) \leqslant \tfrac{1}{2}\rho h(\mathbf{x} + \mathbf{y}) + \tfrac{1}{2}\rho h(\mathbf{x} - \mathbf{y}) \tag{4.1}$$

and

$$h\rho(\mathbf{x}) \leqslant \tfrac{1}{2}h\rho(\mathbf{x} + \mathbf{y}) + \tfrac{1}{2}h\rho(\mathbf{x} - \mathbf{y}). \tag{4.2}$$

Suppose that the unimodal property does not hold and

$$\neg L(\mathbf{x}, \mathbf{x} + \mathbf{y};\ \mathbf{z}) \ \wedge\ L(\mathbf{x} - \mathbf{y}, \mathbf{x};\ \mathbf{z}).$$

If

$$\rho h(\mathbf{x} - \mathbf{y}) < \rho h(\mathbf{z}) \ \wedge\ \rho h(\mathbf{x}) < \rho h(\mathbf{z}) \ \wedge\ h\rho(\mathbf{x} - \mathbf{y}) < h\rho(\mathbf{x})$$

then, from (4.2), we find that $h\rho(\mathbf{x}) < h\rho(\mathbf{x} + \mathbf{y})$. If also $\rho h(\mathbf{x} + \mathbf{y}) < \rho h(\mathbf{z})$ then $L(\mathbf{x}, \mathbf{x} + \mathbf{y};\ \mathbf{z})$, but this relation also holds if $\rho h(\mathbf{x} + \mathbf{y}) \geqslant \rho h(\mathbf{z})$. Therefore, suppose instead that

$$\neg[\rho h(\mathbf{x} - \mathbf{y}) < \rho h(\mathbf{z}) \ \wedge\ \rho h(\mathbf{x}) < \rho h(\mathbf{z})] \ \wedge\ h\rho(\mathbf{x} - \mathbf{y}) < h\rho(\mathbf{x})$$

in which case it is clear that $\rho h(\mathbf{x}) \geqslant \rho h(\mathbf{z})$ and $\rho h(\mathbf{x}) < \rho h(\mathbf{x} + \mathbf{y})$. Again, we see that the relation $L(\mathbf{x}, \mathbf{x} + \mathbf{y};\ \mathbf{z})$ holds. This completes the proof. □

REMARK 4.1. The reflexive and transitive properties of $\neg L(\cdot)$ mean that $\neg L(\cdot)$ defines a TOTAL PREORDERING of $\mathbb{R}^m$, parameterised by its third argument.

We also define the set $\Upsilon(\mathcal{B})$, where $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$, as

$$\Upsilon(\mathcal{B}) = \{\xi, \eta \in \mathbb{R} \mid \rho(\xi\mathbf{b}_1 + \eta\mathbf{b}_2 + \mathbf{b}_3) < \rho(\mathbf{b}_3)\}.$$

We can now set out the following algorithm which, as we shall prove in the next subsection, will produce a sequence of (ρ, h)-minimal sets, each an incremental successor of the previous one, after a finite number of initialisation steps, subject to certain conditions.

ALGORITHM 4.1.

```
1  begin
2    if L(b2, b1; b2) then swap(b1, b2) fi;
3    if L(b3, b2; b3) then swap(b2, b3) fi;
4    if L(b2, b1; b2) then swap(b1, b2) fi;
5    while ρ(b1) > ε ∧ Υ(B) ≠ ∅ do
6          if L(b3 + b1, b3 − b1; b3) then c1 := b1 else c1 := −b1 fi;
7          if L(b3 + b2, b3 − b2; b3) then c2 := b2 else c2 := −b2 fi;
8          do
9              t1 := 0; t2 := 0;
10             if L(b3 + c2 + c1, b3 + c2 − c1; b3)
11               then d1 := c1 else d1 := −c1 fi;
12             if L(b3 − c2 + c1, b3 − c2 − c1; b3)
13               then d2 := c1 else d2 := −c1 fi;
14             while L(b3 + c2 + t1 + d1, b3 + c2 + t1; b3)
15                   do t1 := t1 + d1 od;
16             while L(b3 − c2 + t2 + d2, b3 − c2 + t2; b3)
17                   do t2 := t2 + d2 od;
18             if L(b3 + c2 + t1, b3 − c2 + t2; b3)
19               then c2 := c2 + t1 else c2 := −c2 + t2 fi;
20             swap(c1, c2);
21         while L(b3 + c1, b3 + c2; b3);
22         if L(b2 + b1, b2 − b1; b2) then e1 := b1 else e1 := −b1 fi;
23         if L(b2 + e1, b3 + c2; b3) ∧ L(b2 + e1, b2; b2) then
24           b2 := b2 + e1; output(b2)
25         else b3 := b3 + c2; output(b3) fi;
26         if L(b3, b2; b3) then swap(b2, b3) fi;
27         if L(b2, b1; b2) then swap(b1, b2) fi;
28   od
29 end;
```

4.2. Analysis of the Additive Algorithm. We denote by $\mathcal{B}$ the ordered set consisting of the values of the variables $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ at a specified line and iteration and by $\mathcal{B}'$ the set of the new values of the variables at line 5 on the subsequent

iteration, assuming this point is ever reached, which we will show must happen after a finite number of steps.

We denote by $\mathcal{E}'(\mathcal{B})$ (or just $\mathcal{E}'$) the vector space spanned by the vectors $\{\mathbf{b}_1, \mathbf{b}_2\}$ and by $\Omega'(\mathcal{B})$ (or simply Ω') the lattice spanned by these vectors over $\mathcal{E}'$. We note that at line 21 the set $\{\mathbf{c}_1, \mathbf{c}_2\}$ forms a basis of Ω'. We can then make the important observation that if $\mathcal{B}$ forms a basis of a lattice Ω of rank 3 then $\mathcal{B}'$ must also. We also note that, at line 5, it is always true that

$$\rho h(\mathbf{b}_1) \leqslant \rho h(\mathbf{b}_2) \leqslant \rho h(\mathbf{b}_3).$$

We now state and prove a number of propositions regarding the algorithm, before arriving at our main results. All the propositions assume that $\mathcal{B}$ is a basis of Ω and that Ω, ρ and h form a system for simultaneous Diophantine approximation.

PROPOSITION 4.1. *The loops on lines 14–17 terminate after a finite number of steps and, at their completion,*

$$\neg L(\mathbf{b}_3 + \mathbf{c}_2 + \mathbf{t}_1 + k\mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_2 + \mathbf{t}_1;\ \mathbf{b}_3) \\ \wedge\ \neg L(\mathbf{b}_3 - \mathbf{c}_2 + \mathbf{t}_2 + k\mathbf{c}_1, \mathbf{b}_3 - \mathbf{c}_2 + \mathbf{t}_2;\ \mathbf{b}_3)$$

for all $k \in \mathbb{Z}$.

PROOF. Without loss of generality, consider the execution of the loop on lines 14–15. Recalling that $\neg L(\cdot)$ acts as a total preordering of the lattice points, we will show that there exists a "minimum of $\neg L(\cdot)$" on the line $\mathbf{b}_3 + \mathbf{c}_2 + i\mathbf{c}_1$, $i \in \mathbb{Z}$, which is to say that there exists a point $\mathbf{b}_3 + \mathbf{c}_2 + j\mathbf{c}_1$, such that, for all $k \in \mathbb{Z}$,

$$\neg L(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_2 + j\mathbf{c}_1;\ \mathbf{b}_3) \tag{4.3}$$

and, moreover, $\mathbf{t}_1 = j\mathbf{c}_1$.

We define λ as

$$\lambda = \begin{cases} \{2\rho(\mathbf{b}_3 + \mathbf{c}_2) + \rho(\mathbf{b}_3)\}/\rho(\mathbf{c}_1) & \text{if } \rho(\mathbf{c}_1) > 0, \\ 2h(\mathbf{b}_3 + \mathbf{c}_2)/h(\mathbf{c}_1) & \text{otherwise.} \end{cases}$$

The value of λ is well-defined because $h(\mathbf{c}_1) > 0$ if $\rho(\mathbf{c}_1) = 0$ as ρ and h are transverse. Suppose $\rho(\mathbf{c}_1) > 0$. If $|k| > \lambda$, $k \in \mathbb{Z}$, then

$$\rho(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) \geqslant \rho(k\mathbf{c}_1) - \rho(\mathbf{b}_3 + \mathbf{c}_2) > \rho(\mathbf{b}_3 + \mathbf{c}_2) + \rho(\mathbf{b}_3).$$

This implies that $\rho h(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) > \rho h(\mathbf{b}_3 + \mathbf{c}_2)$ and that $\rho h(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) > \rho h(\mathbf{b}_3)$ and hence $\neg L(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$. Suppose $\rho(\mathbf{c}_1) = 0$. If $|k| > \lambda$, $k \in \mathbb{Z}$, then

$$h(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) \geqslant h(k\mathbf{c}_1) - h(\mathbf{b}_3 + \mathbf{c}_2) > h(\mathbf{b}_3 + \mathbf{c}_2)$$

and

$$\rho(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) = \rho(\mathbf{b}_3 + \mathbf{c}_2).$$

This implies that $\rho h(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) > \rho h(\mathbf{b}_3 + \mathbf{c}_2)$ and that $h\rho(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1) > h\rho(\mathbf{b}_3 + \mathbf{c}_2)$ and hence $\neg L(\mathbf{b}_3 + \mathbf{c}_2 + k\mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$. Since $\neg L(\cdot)$ is transitive in its first two arguments, there must be some point with $|j| \leqslant \lambda$ such that (4.3) must hold.

We can then use the unimodal property of $\neg L(\cdot)$ and the antisymmetric property of $L(\cdot)$ to assure ourselves that the loop will find this minimum (with $\mathbf{t}_1 = j\mathbf{c}_1$) and terminate after a finite number of steps. □

PROPOSITION 4.2. *At line 20,*

$$\neg L(\mathbf{b}_3 - \mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3) \wedge \neg L(\mathbf{b}_3 - \mathbf{c}_2, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3). \tag{4.4}$$

PROOF. Clearly, (4.4) holds upon entering the **do** ... **while** loop at line 8–21. That this condition is maintained is a straightforward consequence of Proposition 4.1. □

PROPOSITION 4.3. *If, at line 8, there exists a non-zero point* $\mathbf{s} \in \Omega'$ *such that* $L(\mathbf{b}_3 + \mathbf{s}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$ *then either the* **do** ... **while** *loop on lines 8–21 will not terminate on the current iteration or*

$$\rho h(\mathbf{b}_3 + \mathbf{c}_1) < \rho h(\mathbf{b}_3) \wedge h\rho(\mathbf{b}_3 + \mathbf{c}_1) < h\rho(\mathbf{b}_3). \tag{4.5}$$

PROOF. Suppose (4.5) does not hold. Because $\{\mathbf{c}_1, \mathbf{c}_2\}$ forms a basis of Ω', we can express $\mathbf{s}$ as

$$\mathbf{s} = a_1\mathbf{c}_1 + a_2\mathbf{c}_2$$

with $a_1, a_2 \in \mathbb{Z}$ and not both zero.

Consider the point $\mathbf{r} \in \Omega'$ where

$$\mathbf{r} = \begin{cases} \operatorname{sgn}(a_1)\mathbf{c}_1 & \text{if } a_2 = 0, \\ q\mathbf{c}_1 + \operatorname{sgn}(a_2)\mathbf{c}_2 & \text{otherwise} \end{cases}$$

and

$$q = \left\lceil \frac{a_1}{|a_2|} \right\rceil.$$

We will show that $L(\mathbf{b}_3 + \mathbf{r}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$.

If $a_2 = 0$ then $\mathbf{r} = \mathbf{s}/|a|$ and therefore

$$\mathbf{b}_3 + \mathbf{r} = \frac{(|a_1| - 1)\mathbf{b}_3 + (\mathbf{b}_3 + \mathbf{s})}{|a_1|}.$$

If $a_2 \neq 0$, then with $p = q|a_2| - a_1$ (which implies that $0 \leqslant p < |a_2|$), it is easy to show that

$$\mathbf{r} = \frac{p\mathbf{c}_1 + \mathbf{s}}{|a_2|}$$

and therefore

$$\mathbf{b}_3 + \mathbf{r} = \frac{(|a_2| - p - 1)\mathbf{b}_3 + p(\mathbf{b}_3 + \mathbf{c}_1) + \mathbf{b}_3 + \mathbf{s}}{|a_2|}.$$

Regardless of the value of a_2, we can apply the triangle inequality to find that

$$\rho h(\mathbf{b}_3 + \mathbf{r}) \leqslant \max\{\rho h(\mathbf{b}_3), \rho h(\mathbf{b}_3 + \mathbf{c}_1), \rho h(\mathbf{b}_3 + \mathbf{s})\}$$

and

$$h\rho(\mathbf{b}_3 + \mathbf{r}) \leqslant \max\{h\rho(\mathbf{b}_3), h\rho(\mathbf{b}_3 + \mathbf{c}_1), h\rho(\mathbf{b}_3 + \mathbf{s})\}.$$

We are assuming that (4.5) does not hold, so either $\rho h(\mathbf{b}_3 + \mathbf{c}_1) \geqslant \rho h(\mathbf{b}_3)$ or $h\rho(\mathbf{b}_3 + \mathbf{c}_1) \geqslant h\rho(\mathbf{b}_3)$. If $\rho h(\mathbf{b}_3 + \mathbf{c}_1) \geqslant \rho h(\mathbf{b}_3)$ then $\rho h(\mathbf{b}_3 + \mathbf{s}) < \rho h(\mathbf{b}_3 + \mathbf{c}_1)$ because $L(\mathbf{b}_3 + \mathbf{s}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$. This implies that $\rho h(\mathbf{b}_3 + \mathbf{r}) < \rho h(\mathbf{b}_3 + \mathbf{c}_1)$ and so $L(\mathbf{b}_3 + \mathbf{r}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$. If $\rho h(\mathbf{b}_3 + \mathbf{c}_1) < \rho h(\mathbf{b}_3)$ and $h\rho(\mathbf{b}_3 + \mathbf{c}_1) \geqslant h\rho(\mathbf{b}_3)$ then $\rho h(\mathbf{b}_3 + \mathbf{s}) < \rho h(\mathbf{b}_3)$ and $h\rho(\mathbf{b}_3 + \mathbf{s}) < h\rho(\mathbf{b}_3 + \mathbf{c}_1)$ because $L(\mathbf{b}_3 + \mathbf{s}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$. This implies that $\rho h(\mathbf{b}_3 + \mathbf{r}) < \rho h(\mathbf{b}_3)$ and $h\rho(\mathbf{b}_3 + \mathbf{r}) < h\rho(\mathbf{b}_3 + \mathbf{c}_1)$ and so again $L(\mathbf{b}_3 + \mathbf{r}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$.

Now, from Proposition 4.1, we know that either $\neg L(\mathbf{b}_3 + \mathbf{c}_2 + \mathbf{t}_1, \mathbf{b}_3 + \mathbf{r};\ \mathbf{b}_3)$ or $\neg L(\mathbf{b}_3 - \mathbf{c}_2 + \mathbf{t}_2, \mathbf{b}_3 + \mathbf{r};\ \mathbf{b}_3)$ and so $L(\mathbf{b}_3 + \mathbf{c}_1, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$ at the end of the **do** ... **while** loop on lines 8–21. Thus, the loop will not terminate on the current iteration and the proposition is proved. □

PROPOSITION 4.4. *The* **do** ... **while** *loop on lines 8–21 must terminate after a finite number of iterations.*

PROOF. We divide the proof into two cases according to whether or not there exists a non-zero point $\mathbf{v}_1 \in \mathcal{E}'$ such that $\rho(\mathbf{v}_1) = 0$.

Case I: There does not exist a non-zero point $\mathbf{v}_1 \in \mathcal{E}'$ *such that* $\rho(\mathbf{v}_1) = 0$.

Consider the state of the algorithm at line 8. Let

$$\Theta = \{\mathbf{z} \in \mathcal{E}' \mid \rho h(\mathbf{z}) < \rho h(\mathbf{c}_1) + 2\rho h(\mathbf{b}_3)\}.$$

Now, Θ is centrally symmetric, convex and bounded and so it must contain a finite number of elements of Ω'. Furthermore, it must contain at least two such elements, the origin and $\mathbf{c}_1$. Notice that, for all $\mathbf{p} \in \Omega' \setminus \Theta$,

$$\rho h(\mathbf{p} + \mathbf{b}_3) \geqslant \rho h(\mathbf{p}) - \rho h(\mathbf{b}_3) \geqslant \rho h(\mathbf{c}_1) + \rho h(\mathbf{b}_3) \geqslant \rho h(\mathbf{c}_1 + \mathbf{b}_3).$$

Additionally, $\rho h(\mathbf{p} + \mathbf{b}_3) \geqslant \rho h(\mathbf{b}_3)$. Therefore, $\neg L(\mathbf{b}_3 + \mathbf{p}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$. If we denote by $\mathbf{c}_1'$ the value of $\mathbf{c}_1$ on line 8 on the subsequent iteration then we have $L(\mathbf{b}_3 + \mathbf{c}_1', \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$ and so $\mathbf{c}_1' \in \Theta$ also. Since Θ contains only a finite number of elements of Ω', the loop must terminate.

Case II: There exists a non-zero point $\mathbf{v}_1 \in \mathcal{E}'$ *such that* $\rho(\mathbf{v}_1) = 0$.

Suppose $\rho(\mathbf{b}_3 + \mathbf{c}_1) \leqslant \rho(\mathbf{b}_3)$ at line 8. Let

$$\Theta = \{\mathbf{z} \in \mathcal{E}' \mid \rho h(\mathbf{z}) < \rho h(\mathbf{c}_1) + 2\rho h(\mathbf{b}_3);\ h\rho(\mathbf{z}) < h\rho(\mathbf{c}_1) + 2h\rho(\mathbf{b}_3)\}.$$

Again, Θ is centrally symmetric, convex and bounded and so it must contain a finite number of elements of Ω', including the origin and $\mathbf{c}_1$. We observe that, for all $\mathbf{p} \in \Omega' \setminus \Theta$,

$$\rho h(\mathbf{p} + \mathbf{b}_3) \geqslant \rho h(\mathbf{p}) - \rho h(\mathbf{b}_3) \geqslant \rho h(\mathbf{b}_3) + \rho h(\mathbf{c}_1) \geqslant \max\{\rho h(\mathbf{b}_3), \rho h(\mathbf{b}_3 + \mathbf{c}_1)\}$$

or

$$h\rho(\mathbf{p} + \mathbf{b}_3) \geqslant h\rho(\mathbf{p}) - h\rho(\mathbf{b}_3) \geqslant h\rho(\mathbf{b}_3) + h\rho(\mathbf{c}_1) \geqslant h\rho(\mathbf{b}_3 + \mathbf{c}_1).$$

If $\rho h(\mathbf{b}_3 + \mathbf{c}_1) \leqslant \rho h(\mathbf{b}_3)$ then $\neg L(\mathbf{b}_3 + \mathbf{p}, \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$. As we observed previously, $L(\mathbf{b}_3 + \mathbf{c}_1', \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$ so $\mathbf{c}_1' \in \Theta$ also and therefore the loop must terminate in a finite number of iterations. Suppose, on the other hand, that $\rho h(\mathbf{b}_3 + \mathbf{c}_1) > \rho h(\mathbf{b}_3)$. This implies that $\rho(\mathbf{b}_3 + \mathbf{c}_1) = \rho(\mathbf{b}_3)$ and $h(\mathbf{b}_3 + \mathbf{c}_1) > h(\mathbf{b}_3)$. If $\mathbf{c}_1' \notin \Theta$ then we find that $\rho h(\mathbf{b}_3 + \mathbf{c}_1') \leqslant \rho h(\mathbf{b}_3)$, since

$$\rho h(\mathbf{b}_3 + \mathbf{c}_1') \geqslant \rho h(\mathbf{b}_3 + \mathbf{c}_1) \ \Rightarrow\ \neg L(\mathbf{b}_3 + \mathbf{c}_1', \mathbf{b}_3 + \mathbf{c}_1;\ \mathbf{b}_3)$$

and

$$\rho h(\mathbf{b}_3) < \rho h(\mathbf{b}_3 + \mathbf{c}_1') < \rho h(\mathbf{b}_3 + \mathbf{c}_1) \ \Rightarrow\ \mathbf{c}_1' \in \Theta.$$

Thus, we can use our arguments above on the next iteration to show that the number of remaining iterations is finite.

Suppose instead that $\rho(\mathbf{b}_3 + \mathbf{c}_1) > \rho(\mathbf{b}_3)$ at line 8. Therefore, $\rho(\mathbf{c}_1) > 0$. Consider a basis of $\mathcal{E}'$ consisting of $\mathbf{v}_1$ and another vector $\mathbf{v}_2$. Now, $\rho(\mathbf{v}_2) > 0$ since $\rho(\mathbf{v}_1) = 0$, otherwise $\Upsilon(\mathcal{B}) = \emptyset$. Thus, we can express $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2)$ as $\mathbf{C} = \mathbf{V}\mathbf{X}$ where where $\mathbf{X} \in \mathbb{R}^{2\times 2}$. We observe that $x_{21} \neq 0$ since $\rho(\mathbf{c}_1) > 0$. Also, we can quickly confirm that there exists some non-zero $\alpha \in \mathbb{R}$ such that, for all $\beta \in \mathbb{R}$,

$$\beta(\beta - \alpha) \leqslant 0 \ \Leftrightarrow\ \rho(\mathbf{b}_3 + \beta\mathbf{v}_2) \leqslant \rho(\mathbf{b}_3) \tag{4.6}$$

and that if $|\alpha| \leqslant |\beta| < |\gamma|$, for some $\gamma \in \mathbb{R}$, and $\beta\gamma > 0$ then

$$\rho(\mathbf{b}_3 + \gamma\mathbf{v}_2) > \rho(\mathbf{b}_3 + \beta\mathbf{v}_2). \tag{4.7}$$

With

$$\kappa = \left\lfloor \frac{x_{22}}{x_{21}} \right\rceil$$

we have

$$|x_{22} - \kappa x_{21}| \leqslant \tfrac{1}{2}|x_{21}| \qquad \text{and} \qquad |x_{22} - \kappa x_{21}| \leqslant |x_{22} - j x_{21}| \tag{4.8}$$

for all $j \in \mathbb{Z}$. If

$$|x_{22} - \kappa x_{21}| \leqslant |\alpha| \tag{4.9}$$

then either

$$\rho(\mathbf{b}_3 + \mathbf{c}_2 - \kappa\mathbf{c}_1) \leqslant \rho(\mathbf{b}_3) \qquad \text{or} \qquad \rho(\mathbf{b}_3 - \mathbf{c}_2 + \kappa\mathbf{c}_1) \leqslant \rho(\mathbf{b}_3)$$

and so $\rho(\mathbf{b}_3 + \mathbf{c}'_1) \leqslant \rho(\mathbf{b}_3)$. On the next iteration, we can then use the arguments above where $\rho(\mathbf{b}_3 + \mathbf{c}_1) \leqslant \rho(\mathbf{b}_3)$ to show that the number of iterations is finite. Otherwise, because of (4.6)–(4.8), we find that

$$\mathbf{c}'_1 = \pm(\mathbf{c}_2 - \kappa\mathbf{c}_1)$$

and that, after a finite number of iterations, (4.9) must become true. Therefore, the total number of iterations must be finite. □

We are now able to state and prove the main results of this section.

PROPOSITION 4.5. *Suppose that $\mathcal{B}^{(0)}$ is a basis of a lattice Ω of rank 3 and ρ and h are strictly convex, complementary radius and height functions. If Algorithm 4.1 is executed with $\mathcal{B}^{(0)}$ as its initial basis then, after a finite number of steps, $\mathcal{B}$ will be (ρ, h)-minimal at line 5 or the algorithm will terminate.*

PROOF. Proposition 4.1 and Proposition 4.4 imply that the **do** ... **while** loop on lines 8–21 must terminate after a finite number of steps. Suppose that, when this loop terminates, (3.34) and (3.35) in Theorem 3.12 are satisfied for $\mathbf{w} = \mathbf{b}_3 + \mathbf{c}_2$, $j = 3$, and $\mathbf{w} = \mathbf{b}_2 + \mathbf{e}_1$, $j = 2$. That is,

$$\rho h(\mathbf{b}_3 + \mathbf{c}_2) \geqslant \max_{i=1,2,3} \{\rho h(\mathbf{b}_i)\} \ \vee \ h\rho(\mathbf{b}_3 + \mathbf{c}_2) \geqslant \max_{i=1,2,3} \{h\rho(\mathbf{b}_i)\} \tag{4.10}$$

and

$$\rho h(\mathbf{b}_2 + \mathbf{e}_1) \geqslant \max_{i=1,2} \{\rho h(\mathbf{b}_i)\} \ \vee \ h\rho(\mathbf{b}_2 + \mathbf{e}_1) \geqslant \max_{i=1,2,3} \{h\rho(\mathbf{b}_i)\}. \tag{4.11}$$

Now, (4.10) and Proposition 4.3 imply that $\neg L(\mathbf{b}_3 + \mathbf{s}, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$ for all non-zero $\mathbf{s} \in \Omega'$. This, in turn, implies that conditions (3.34) and (3.35) are satisfied for all $\mathbf{w} \in \Lambda_3(\mathcal{B})$. Similarly, (4.11) and the tests on line 22 imply that these same conditions are satisfied for all $\mathbf{w} \in \Lambda_2(\mathcal{B})$. Hence, Theorem 3.12 can be applied to show that $\mathcal{B}$ is (ρ, h)-minimal at this point.

The number of points $\mathbf{w} \in \Omega$ which cannot satisfy either (3.34) or (3.35) is finite and must decrease strictly for each iteration through the outer loop of the algorithm until (4.10) and (4.11) are satisfied or the algorithm terminates. We have shown that each iteration through this loop is completed in a finite number of iterations through the inner loops, so the proposition is proved. □

PROPOSITION 4.6. *If, at line 5 of Algorithm 4.1, $\mathcal{B}$ is (ρ, h)-minimal in a lattice Ω and ρ and h are strictly convex, complementary radius and height functions and $\epsilon > 0$ then, on the subsequent iteration, $\mathcal{B}'$ will be an incremental successor to $\mathcal{B}$.*

PROOF. Now, ρ is not null-spanned by $\mathcal{B}$ since $\rho(\mathbf{b}_1) > \epsilon > 0$. From Theorem 3.13 and Theorem 3.2 there either exists an innovation $\mathbf{s} \in \Lambda_3(\mathcal{B})$ such that $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{s}$ can be arranged to form an incremental successor to $\mathcal{B}$ or there exists an innovation $\mathbf{s} \in \Lambda_2(\mathcal{B})$ such that $\mathbf{b}_1$, $\mathbf{s}$ and $\mathbf{b}_3$ can be arranged to form an incremental successor.

Suppose the former. Since $\mathcal{B}$ is (ρ, h)-minimal, Proposition 4.3 implies that $\neg L(\mathbf{r}, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$ for all $\mathbf{r} \in \Lambda_3(\mathcal{B})$ when the **do** ... **while** loop terminates. Thus $h\rho(\mathbf{b}_3 + \mathbf{c}_2) \leqslant h\rho(\mathbf{r})$. But (3.12) of Theorem 3.2 implies that $h\rho(\mathbf{b}_3 + \mathbf{c}_2) = h\rho(\mathbf{s})$ and therefore $\neg L(\mathbf{b}_2 + \mathbf{e}_1, \mathbf{b}_3 + \mathbf{c}_2;\ \mathbf{b}_3)$ at line 23 and $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3 + \mathbf{c}_2$ can be, and will be, arranged to form an incremental successor to $\mathcal{B}$.

Suppose there is no choice for $\mathbf{s} \in \Lambda_3(\mathcal{B})$ such that $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{s}$ can be arranged to form an incremental successor. Thus, there exists $\mathbf{s} \in \Lambda_2(\mathcal{B})$ such that $\mathbf{b}_1$, $\mathbf{s}$ and $\mathbf{b}_3$ can be arranged to form an incremental successor. It is clear that, at line 23, $\mathbf{e}_1$ has been chosen such that $\neg L(\mathbf{r}, \mathbf{b}_2 + \mathbf{e}_1;\ \mathbf{b}_2)$ for all $\mathbf{r} \in \Lambda_2(\mathcal{B}) \cup \Lambda_3(\mathcal{B})$, which implies that $h\rho(\mathbf{b}_2 + \mathbf{e}_1) = h\rho(\mathbf{s})$. Hence, $\mathbf{b}_1$, $\mathbf{b}_2 + \mathbf{e}_1$ and $\mathbf{b}_3$ can be, and will be, arranged to form an incremental successor to $\mathcal{B}$. □

REMARK 4.2. We note that if, at line 8 of Algorithm 4.1, $\mathcal{B}$ is (ρ, h)-minimal and ρ and h are strictly convex, complementary radius and height functions, then Algorithm 4.1 is an incremental successor algorithm to the extent that if $\mathbf{p}$ is a best approximation such that

$$\epsilon \leqslant \rho(\mathbf{p}) \leqslant \rho(\mathbf{b}_1)$$

then there exists an equivalent best approximation which will be output by the algorithm after a finite number of steps, unless the algorithm terminates with $\Upsilon(\mathcal{B}) = \emptyset$. If this termination condition occurs, we can then use Algorithm 3.1 to find further best approximations in Ω'.

4.3. Numerical Examples. We illustrate the operation of Algorithm 4.1 with some numerical examples. All numerical results which represent real numbers are given to three significant figures.

EXAMPLE 4.1. This is the example which was used by BRENTJES (1981) to demonstrate the operation of his multi-dimensional continued fraction algorithm. In this example, and indeed all examples in this subsection, we shall use the lattice $\mathbb{Z}^3$. We define $\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_2$ where

$$\mathbf{P}^T = \begin{pmatrix} -\sqrt[3]{5} & 1 & 0 \\ -\sqrt[3]{25} & 0 & 1 \end{pmatrix}. \tag{4.12}$$

We define $h(\mathbf{x}) = |\boldsymbol{\alpha} \cdot \mathbf{x}|$ where $\boldsymbol{\alpha} = (1, 0, 0)$. We note that ρ and h are strictly convex, complementary radius and height functions. With the identity matrix used as the initial basis, Algorithm 4.1 was executed with $\epsilon = 0.1$. The results are listed in Table 1. This table lists the state of $\mathcal{B}$ at line 5 on each iteration of the algorithm.

TABLE 1. Outputs and important variables of Algorithm 4.1 in BRENTJES' example.

It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{P}^T\mathbf{v}$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$			
2	$(0,1,0)$	$(0,0,1)$	$(\mathbf{1},\mathbf{3},\mathbf{2})$	$(1.29,-0.924)$	1.59	1
3	$(\mathbf{1},\mathbf{2},\mathbf{3})$	$(0,1,0)$	$(0,0,1)$	$(0.290,0.0760)$	0.300	1
4	$(1,2,3)$	$(\mathbf{1},\mathbf{1},\mathbf{3})$	$(0,0,1)$	$(-0.710,0.0760)$	0.714	1
5	$(1,2,3)$	$(1,1,3)$	$(\mathbf{1},\mathbf{2},\mathbf{2})$	$(0.290,-0.924)$	0.968	1
6	$(1,2,3)$	$(\mathbf{2},\mathbf{3},\mathbf{6})$	$(1,2,2)$	$(-0.420,0.152)$	0.447	2
7	$(1,2,3)$	$(2,3,6)$	$(\mathbf{2},\mathbf{3},\mathbf{5})$	$(-0.420,-0.848)$	0.946	2
8	$(\mathbf{3},\mathbf{5},\mathbf{9})$	$(1,2,3)$	$(2,3,5)$	$(-0.130,0.228)$	0.262	3
9	$(3,5,9)$	$(1,2,3)$	$(\mathbf{3},\mathbf{5},\mathbf{8})$	$(-0.130,-0.772)$	0.783	3
10	$(3,5,9)$	$(1,2,3)$	$(\mathbf{4},\mathbf{7},\mathbf{11})$	$(0.160,-0.696)$	0.714	4
11	$(3,5,9)$	$(1,2,3)$	$(\mathbf{6},\mathbf{10},\mathbf{17})$	$(-0.260,-0.544)$	0.603	6
12	$(3,5,9)$	$(1,2,3)$	$(\mathbf{7},\mathbf{12},\mathbf{20})$	$(0.0302,-0.468)$	0.469	7
13	$(\mathbf{10},\mathbf{17},\mathbf{29})$	$(3,5,9)$	$(1,2,3)$	$(-0.100,-0.240)$	0.260	10
14	$(\mathbf{11},\mathbf{19},\mathbf{32})$	$(10,17,29)$	$(3,5,9)$	$(-0.190,0.164)$	0.251	11
15	$(\mathbf{13},\mathbf{22},\mathbf{38})$	$(11,19,32)$	$(10,17,29)$	$(-0.230,-0.0122)$	0.230	13
16	$(\mathbf{14},\mathbf{24},\mathbf{41})$	$(13,22,38)$	$(11,19,32)$	$(0.0603,0.0638)$	0.0878	14

The number of the iteration is listed in the first column. The vector in bold face is the innovation into $\mathcal{B}$ on that iteration, which we denote by $\mathbf{v}$. We also list $\mathbf{P}^T\mathbf{v}$ and the radius and height of $\mathbf{v}$. In many cases the symmetric lattice point is listed to that actually found by the algorithm, to ensure that all lattice coordinates are positive. Observe that the algorithm has produced a (ρ, h)-minimal set by the third iteration. We know this because the innovation $\mathbf{v}$ produced on the third iteration (and listed in the table in bold face for iteration 4) satisfies $h\rho(\mathbf{v}) \geqslant \max_{i=1,2,3}\{h\rho(\mathbf{b}_i)\}$.

Figure 1 illustrates the "itineraries" of innovations produced by Algorithm 4.1. Three itineraries are plotted. An ITINERARY is a record of the successive innovations which replace a given initial element in the basis, regardless of subsequent reordering. The iteration on which a given basis element becomes an inveteration is displayed near the line joining the projection of that element with the projection of the innovation which replaces it. Best approximations are circled.

EXAMPLE 4.2. We now consider the use of Algorithm 4.1 for finding best approximate integer relations to $(e^2, e, 1)$. Implicitly, the lattice to be used is $\mathbb{Z}^3$. We define "best" in terms of the radius function $\rho(\mathbf{x}) = |\boldsymbol{\alpha} \cdot \mathbf{x}|$ where $\boldsymbol{\alpha} = (e^2, e, 1)$ and a height function $h(X) = \left\|\mathbf{P}^T\mathbf{x}\right\|_2$ where we set $\mathbf{P}$ to be an orthonormal basis of the complementary subspace of $\boldsymbol{\alpha}$ in $\mathbb{R}^3$. Since ρ and h are strictly convex, complementary radius and height functions, we can use Algorithm 4.1 to find a list of successive best approximations. Using the identity matrix as the initial basis, we

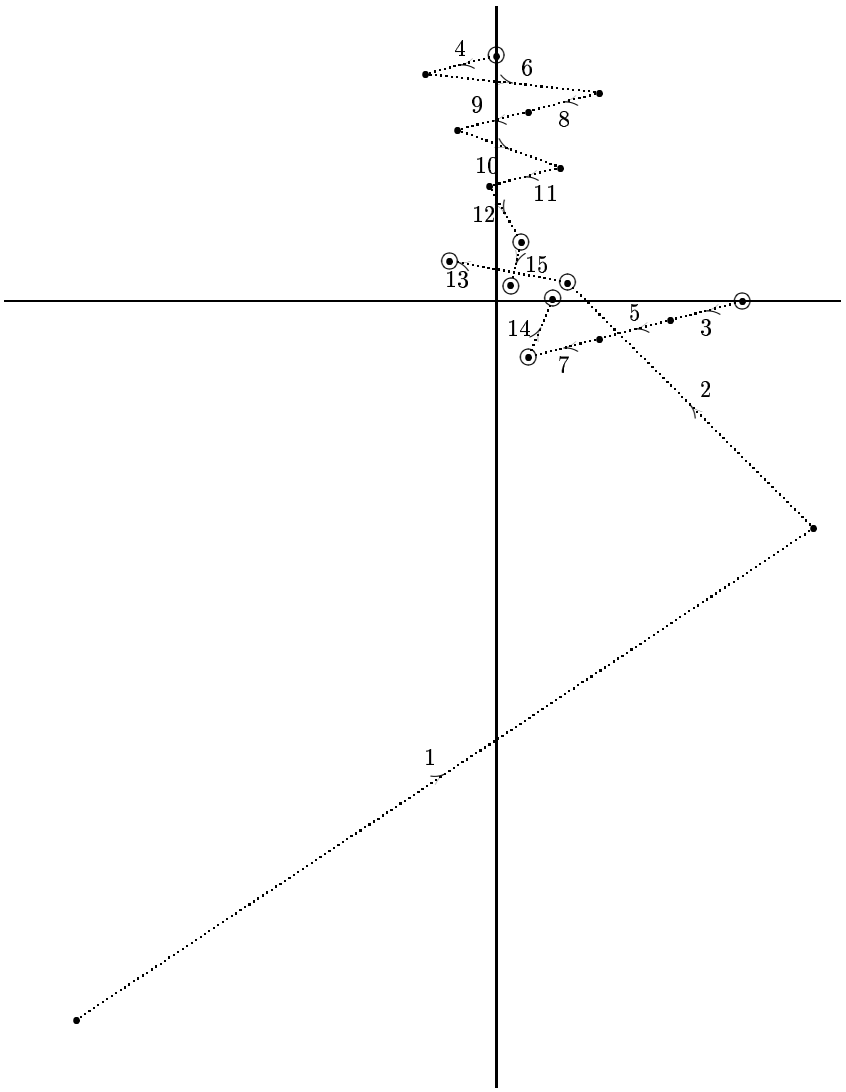


FIGURE 1. Itineraries of innovations in BRENTJES' example.

find, first of all, that the identity matrix is (ρ, h)-minimal since at the beginning of the second iteration the new element in the basis, $\mathbf{v}$, has

$$h\rho(\mathbf{v}) > \max_{i=1,2,3} \{h\rho(\mathbf{e}_i)\}$$

where the $\mathbf{e}_i$ are the columns of the identity matrix. With $\epsilon = 10^{-4}$, we can then be assured that the algorithm will not "miss" any best approximations, in the sense of Theorem 3.3, with a radius between ϵ and 1. Table 2 lists the best approximate integer relations found by the algorithm.

EXAMPLE 4.3. We now present a slightly novel application of Diophantine approximation, which is to calculate the simultaneous overlap of periodic pulse trains. We will return to this subject in the next chapter. A pulse train can be thought of

TABLE 2. Best approximate integer relations for $(e^2, e, 1)$.

$\mathbf{p}$	$\rho(\mathbf{p})$	$h(\mathbf{p})$
$(0, 0, 1)$	1.00	0.992
$(0, 1, -2)$	0.718	2.23
$(-1, 2, 2)$	4.75×10^{-2}	3.00
$(2, -8, 7)$	3.19×10^{-2}	10.8
$(-3, 10, -5)$	1.56×10^{-2}	11.6
$(3, -3, -14)$	1.23×10^{-2}	14.6
$(-6, 13, 9)$	3.33×10^{-3}	16.9
$(8, -28, 17)$	5.58×10^{-4}	33.7
$(-35, 83, 33)$	4.28×10^{-4}	95.9
$(43, -111, -16)$	1.29×10^{-4}	120
$(-11, 70, -109)$	1.11×10^{-4}	130
$(54, -181, 93)$	1.84×10^{-5}	211

as a binary function $f(t;\ T, \tau)$ defined so that

$$f(t;\ T, \tau) = \begin{cases} 1 & \text{if } |t - \lfloor t/T \rceil T| \leqslant \frac{1}{2}\tau, \\ 0 & \text{otherwise.} \end{cases}$$

The variable t represents time. The parameter $T > 0$ is the PULSE REPETITION INTERVAL (PRI) of the pulse train and τ is the PULSE WIDTH which satisfies $0 < \tau < T$. The i^{th} PULSE of the pulse train occurs on the interval $\left[iT - \frac{1}{2}\tau, iT + \frac{1}{2}\tau\right]$. Two pulse trains, with PRIs of T_1 and T_2 and pulse widths of τ_1 and τ_2, OVERLAP whenever $f(t;\ T_1, \tau_1) = f(t;\ T_2, \tau_2) = 1$. Thus, they overlap whenever there exists $i, j \in \mathbb{Z}$ such that

$$|iT_1 - jT_2| \leqslant \tfrac{1}{2}(\tau_1 + \tau_2).$$

Clearly, overlap always occurs at $t = i = j = 0$.

Now, consider the simultaneous overlap of three pulse trains with PRIs T_1, T_2 and T_3 and pulse widths τ_1, τ_2 and τ_3. Simultaneous overlap occurs whenever

$$f(t;\ T_1, \tau_1) = f(t;\ T_2, \tau_2) = f(t;\ T_3, \tau_3) = 1$$

or, equivalently, whenever there exists $i, j, k \in \mathbb{Z}$ such that

$$|iT_1 - jT_2| \leqslant \tfrac{1}{2}(\tau_1 + \tau_2),$$
$$|jT_2 - kT_3| \leqslant \tfrac{1}{2}(\tau_2 + \tau_3)$$

and

$$|iT_1 - kT_3| \leqslant \tfrac{1}{2}(\tau_1 + \tau_3).$$

We will now pose a specific problem which we can use Algorithm 4.1 to solve. Suppose we are interested in three pulse trains. We know that the PRIs are $T_1 = e^2$,

$T_2 = e$ and $T_3 = 1$ and that the pulse widths are proportional to their PRIs with some constant of proportionality γ, which we call the DUTY CYCLE, so that $\tau_1 = \gamma T_1$, $\tau_2 = \gamma T_2$ and $\tau_3 = \gamma T_3$. Furthermore, suppose we want to find the first overlap in terms of the pulse index of the first pulse train $i > 0$ other than the obvious overlap at $t = i = j = k = 0$ for a range of γ. This is a simultaneous Diophantine approximation problem. With the lattice $\mathbb{Z}^3$, we could define ρ so that

$$\rho(\mathbf{x}) = 2 \max \left\{ \frac{|x_1 e - x_2|}{e+1}, \frac{|x_2 e - x_3|}{e+1}, \frac{|x_1 e^2 - x_3|}{e^2+1} \right\}$$

and

$$h(\mathbf{x}) = |x_1|.$$

Substituting $\mathbf{v} = (i, j, k) \in \mathbb{Z}^3$ for $\mathbf{v}$ we see that $\rho(\mathbf{p}) \leqslant \gamma$ if and only if the three pulse trains simultaneously overlap for that duty cycle and pulse indices i, j and k. We observe that ρ and h are transverse and complementary, but ρ is not strictly convex. We can overcome this by redefining ρ as an extended semi-norm so that

$$\rho(\mathbf{x}) = 2 \operatorname{sort} \left\{ \frac{|x_1 e - x_2|}{e+1}, \frac{|x_2 e - x_3|}{e+1}, \frac{|x_1 e^2 - x_3|}{e^2+1} \right\}$$

where sort $\{\cdot\}$ returns its arguments in descending order. It can be readily checked that this extended semi-norm is now strictly convex. Algorithm 4.1 can now be executed with the initial basis set to the identity to find the list of all first simultaneous overlaps (which are best approximations with respect to ρ and h). After the third iteration the algorithm produces the (ρ, h)-minimal set

$$\mathcal{B} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 8 & 2 & 1 \end{pmatrix}$$

where the lattice points are the columns of this matrix. Hence, every subsequent iteration produces a new (ρ, h)-minimal set, the first element of which represents a first simultaneous overlap. The results obtained from the algorithm with $\epsilon = 0.01$ are presented in Table 3. For any listed vector in the table, the corresponding value

TABLE 3. Pulse indices for first simultaneous overlap of pulse trains.

(i, j, k)	γ
$(1, 3, 8)$	0.152
$(3, 8, 22)$	0.136
$(4, 11, 30)$	0.106
$(18, 49, 133)$	0.105
$(21, 57, 155)$	0.0451
$(64, 174, 473)$	0.0239
$(252, 685, 1862)$	0.0124
$(1537, 4178, 11357)$	9.96×10^{-3}

of γ is the minimum duty cycle for which the vector represents a first simultaneous overlap and the previous value of γ is the maximum duty cycle, but the duty cycle must be strictly less than this maximum.

In the example above, it was necessary to extend our radius function in order to satisfy the condition of strict convexity. If we had not done this, it would not have made any difference in this case. However, this cannot be guaranteed in general. We must also exercise a degree of care in interpreting best approximations for a radius function which has been extended from a radius function that is not strictly convex but is our true interest. It is necessary to check the preceding best approximation to ensure that the radii differ in their "true" radius, otherwise the best approximation in the extended radius function is not best with respect to the true radius function.

5. An Accelerated Algorithm for Lattices of Rank 3

5.1. Furtwängler's Algorithm. In the next subsection we will discuss an accelerated version of the additive algorithm (Algorithm 4.1) we presented in Section 4.1. The accelerated algorithm can be regarded as a generalisation of an algorithm described by FURTWÄNGLER (1927). We briefly review his algorithm in this subsection.

FURTWÄNGLER proposed an algorithm for finding all best approximations to a line in three dimensions when the radius function is the sup-norm. He considers only lattices Ω of rank 3 in $\mathbb{R}^3$. According to our definitions of a system for simultaneous Diophantine approximation, he specifies the algorithm only for systems in which the radius function and height functions have the form

$$\rho(\mathbf{x}) = \max\{|x_1|, |x_2|\} \qquad \text{and} \qquad h(\mathbf{x}) = |x_3|$$

He also requires that, for all $\mathbf{x} \in \Omega$ with $h(\mathbf{x}) \neq 0$, neither $x_1 = 0$ nor $x_2 = 0$ and nor is there any other lattice point $\mathbf{y}$ with $h(\mathbf{y}) \neq 0$ for which $\rho(\mathbf{x}) = \rho(\mathbf{y})$.

At each iteration of Furtwängler's algorithm, a new basis, $\mathcal{B}'$, is generated from the given or old one, $\mathcal{B}$. The matrix $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ of basis vectors is assumed to have certain properties. Firstly, it is assumed that $b_{3,i} > 0$ for $i = 1, 2, 3$. Secondly, $\mathbf{b}_1$ is a best approximation (which FURTWÄNGLER calls an APPROXIMATION POINT). Lastly, $\mathbf{b}_2$ is an AUXILIARY APPROXIMATION POINT. By this, it is meant that $\rho(\mathbf{b}_2) \geqslant \rho(\mathbf{b}_1)$ and, for all $\mathbf{w} \in \Omega$ such that $\mathbf{w}$ is not a multiple of $\mathbf{b}_1$, either

$$\rho(\mathbf{w}) \geqslant \rho(\mathbf{b}_2) \tag{5.1}$$

or

$$h(\mathbf{w}) \geqslant \max\{h(\mathbf{b}_1), h\rho(\mathbf{b}_2)\}. \tag{5.2}$$

The two points $\mathbf{b}_1$ and $\mathbf{b}_2$ can be thought of as defining an APPROXIMATION PRISM consisting only of the origin, $\pm\mathbf{b}_2$ and integer multiples of $\mathbf{b}_1$. Observe that (5.1)

and (5.2) bear a strong resemblance to the conditions (2.3) and (2.4) for (ρ, h)-minimal sets in Definition 2.10. Also observe that the auxiliary approximation point may itself be a best approximation, but only if $h(\mathbf{b}_2) \leqslant h(\mathbf{b}_1)$.

Consider extending the "roof" and "floor," that is, the height, of the approximation prism until a lattice point is encountered which is not a multiple of $\mathbf{b}_1$. Call such a point $\mathbf{s}$, the NEXT APPROXIMATION POINT. Either $\mathbf{s}$ will be a new best approximation point or an auxiliary approximation.

FURTWÄNGLER shows that the next approximation point (in order of increasing height), $\mathbf{s}$, has either the form

$$\mathbf{s} = \mathbf{b}_1 + \mathbf{b}_2 \tag{5.3}$$

or

$$\mathbf{s} = c_{1k}\mathbf{b}_1 + c_{2k}\mathbf{b}_2 + \mathbf{b}_3 \tag{5.4}$$

where a maximum of four candidate pairs (c_{1k}, c_{2k}), $k = 1, 2, 3, 4$, need to be considered and they are easily calculated and decided between. Furthermore, the c_{2k} are consecutive integers. Note that $-\mathbf{s}$ may be used instead if $s_3 < 0$.

If $\mathbf{s}$ takes the form (5.3) then it replaces $\mathbf{b}_2$ in the new basis, otherwise it replaces $\mathbf{b}_3$. Clearly, $\mathbf{s}$ and $\mathbf{b}_1$ will form a new approximation prism and $\mathbf{s}$ may be a new best approximation. If this is so, then the basis is arranged so that $\mathbf{b}'_1 = \mathbf{s}$ and $\mathbf{b}'_2 = \mathbf{b}_1$, otherwise the opposite assignments are made.

For the lattices considered by FURTWÄNGLER, it can be shown that all best approximations will be found by this algorithm, provided it can be initialised. FURTWÄNGLER claims that it is always possible to find a suitable initial lattice base but does not show how this can be done in general.

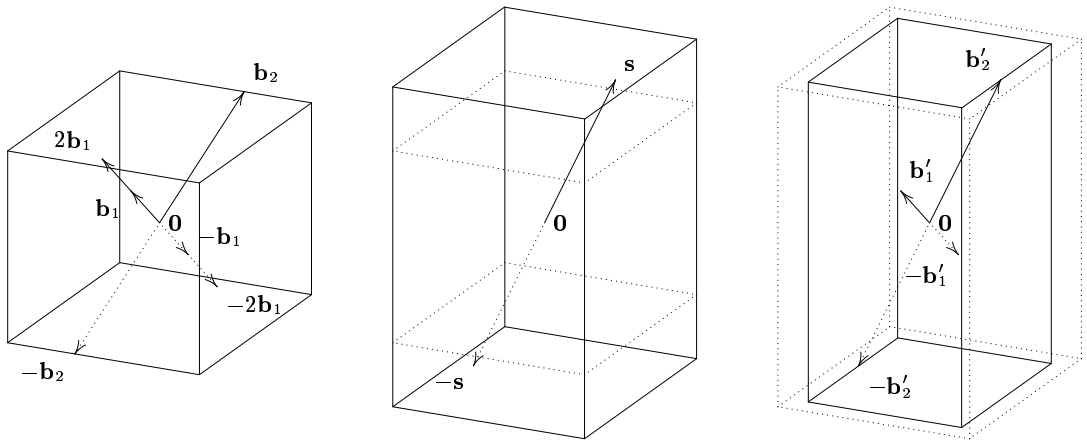


FIGURE 2. An iteration of Furtwängler's algorithm.

Figure 2 illustrates a single iteration of Furtwängler's algorithm. At left, an approximation prism is depicted at the beginning of an iteration. It can be seen that the prism contains $\pm\mathbf{b}_2$ and $k\mathbf{b}_1$, $k = -2, -1, 0, 1, 2$. The height of the prism is

then extended until a lattice point, $\mathbf{s}$, is encountered which is linearly independent of $\mathbf{b}_1$. This is depicted in the central diagram. At right, the iteration concludes by making the assignments $\mathbf{b}'_1 = \mathbf{b}_1$ and $\mathbf{b}'_2 = \mathbf{s}$. The new approximation prism has a greater height but smaller radius.

5.2. The Principles of the Accelerated Algorithm. We now investigate accelerating our additive algorithm (Algorithm 4.1). We will find that this leads to a generalisation of Furtwängler's algorithm.

From Theorems 3.11 to 3.13, we know that, for a simultaneous Diophantine approximation system consisting of a lattice Ω of rank 3 in $\mathbb{R}^m$ and strictly convex, complementary radius and height functions, ρ and h, the (ρ, h)-minimal sets are always bases of the lattice. If $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ is a (ρ, h)-minimal basis and ρ is not null-spanned by the basis, an incremental successor, can be found in which the innovation, $\mathbf{s}$, has the form

$$\mathbf{s} = \mathbf{b}_1 \pm \mathbf{b}_2 \qquad \text{or} \qquad \mathbf{s} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2 + \mathbf{b}_3 \tag{5.5}$$

with $a_1, a_2 \in \mathbb{Z}$. That is, $\mathbf{s} \in \Lambda_2(\mathcal{B})$ or $\mathbf{s} \in \Lambda_3(\mathcal{B})$.

As we discussed at the beginning of Section 3.3, because ρ and h are complementary, we can express ρ and h for any $\mathbf{x}$ in the real span of Ω as

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\| \qquad \text{and} \qquad h(\mathbf{x}) = |\boldsymbol{\alpha} \cdot \mathbf{x}| \tag{5.6}$$

or

$$\rho(\mathbf{x}) = |\boldsymbol{\alpha} \cdot \mathbf{x}| \qquad \text{and} \qquad h(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\| \tag{5.7}$$

where $\|\cdot\|$ is a strictly convex extended norm, $\mathbf{P}^T$ is a $2 \times m$ matrix and $\{\mathbf{p}_1, \mathbf{p}_2, \boldsymbol{\alpha}\}$ forms a basis of the real span of Ω. To avoid making specific reference to either (5.6) or (5.7), let us make use of an underline and an overline to represent the appropriate mappings so that

$$\rho(\mathbf{x}) = \|\underline{\mathbf{x}}\|^* \qquad \text{and} \qquad h(\mathbf{x}) = \|\overline{\mathbf{x}}\|^\dagger$$

where $\|\cdot\|^*$ and $\|\cdot\|^\dagger$ are strictly convex extended norms acting on $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Either $n_1 = 1$ and $n_2 = 1$ or else $n_1 = 2$ and $n_2 = 1$. Therefore, either $\underline{\mathbf{x}}$ or $\overline{\mathbf{x}}$ is a scalar for all $\mathbf{x} \in \mathbb{R}^m$.

To expound the principles on which our accelerated algorithm is built, we require the notion of a primitively (ρ, h)-minimal set.

Definition 5.1. A set $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ is PRIMITIVELY (ρ, h)-MINIMAL in a lattice Ω of rank $n > k$ if there exist $n - k$ lattice points $\mathbf{v}_{k+1}, \mathbf{v}_{k+2}, \ldots, \mathbf{v}_n$ such that $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is (ρ, h)-minimal.

Corollary 5.1. *If $\{\mathbf{v}_1, \mathbf{v}_2\}$ is primitively (ρ, h)-minimal in a lattice Ω of rank r in $\mathbb{R}^m$ and ρ and h are strictly convex, complementary radius and height functions then $\{\mathbf{v}_1, \mathbf{v}_2\}$ forms a primitive basis of Ω.*

PROOF. This is a direct consequence of Theorem 3.11. □

Consider a primitively (ρ, h)-minimal set $(\mathbf{b}_1, \mathbf{b}_2)$ with $\rho(\mathbf{b}_1) > 0$ in our simultaneous Diophantine approximation system. We know that any (ρ, h)-minimal set containing $(\mathbf{b}_1, \mathbf{b}_2)$ must be a basis of the lattice. Consider the (ρ, h)-minimal set $\mathcal{W} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{w})$ for which the innovation $\mathbf{s}^*$ into the incremental successor of $\mathcal{W}$ satisfies $\rho h(\mathbf{s}^*) < \rho h(\mathbf{b}_2)$. That is, $\mathbf{s}^* \in \Xi_2(\mathcal{B})$. We know that such a (ρ, h)-minimal set can be found for the primitively (ρ, h)-minimal set $(\mathbf{b}_1, \mathbf{b}_2)$ as a consequence of Theorem 3.4. Furthermore, from Theorem 3.13, we know that the $\mathbf{s}^* \in \Lambda_2(\mathcal{W})$ or $\mathbf{s}^* \in \Lambda_3(\mathcal{W})$.

The key idea of the accelerated algorithm is to find an innovation $\mathbf{s}^*$ of the type described, given an ordered basis $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ in which the first two vectors form a primitively (ρ, h)-minimal set. To do this, we find an element $\mathbf{q} \in \Lambda_2(\mathcal{B})$ such that $\neg L(\mathbf{w}, \mathbf{q};\ \mathbf{b}_2)$. If there is an element $\mathbf{v} \in \Lambda_3(\mathcal{B})$ such that $\neg L(\mathbf{q}, \mathbf{v};\ \mathbf{b}_2)$ then we find an element $\mathbf{r} \in \Lambda_3(\mathcal{B})$ such that $\neg L(\mathbf{w}, \mathbf{r};\ \mathbf{b}_2)$. Having found $\mathbf{q}$ and $\mathbf{r}$, we know that $\mathbf{s}^* = \mathbf{q}$ if $L(\mathbf{q}, \mathbf{r};\ \mathbf{b}_2)$, otherwise $\mathbf{s}^* = \mathbf{r}$.

We will see that we can do this in a fixed number of arithmetic operations for certain choices of the extended norms $\|\cdot\|^*$ and $\|\cdot\|^\dagger$. The advantage over the additive algorithm (Algorithm 4.1) is that, by so doing, we can "skip" a number of intermediate innovations and find best approximations more quickly.

Selecting $\mathbf{q}$ is simple. We test $\mathbf{b}_1 - \mathbf{b}_2$ and $\mathbf{b}_1 + \mathbf{b}_2$. If $L(\mathbf{b}_1 - \mathbf{b}_2, \mathbf{b}_1 + \mathbf{b}_2;\ \mathbf{b}_2)$ the $\mathbf{q} = \mathbf{b}_1 - \mathbf{b}_2$, otherwise $\mathbf{q} = \mathbf{b}_1 + \mathbf{b}_2$. Let us then turn our attention to a method for selecting $\mathbf{r}$. Since $\mathbf{r} \in \Lambda_3(\mathcal{B})$ we can write

$$\mathbf{r} = a_1^*\mathbf{b}_1 + a_2^*\mathbf{b}_2 + \mathbf{b}_3$$

where $a_1^*, a_2^* \in \mathbb{Z}$. Our method for selecting $\mathbf{r}$ is therefore a method for selecting the pair of integers a_1^* and a_2^*.

Suppose $\underline{\mathbf{b}}_1$ and $\underline{\mathbf{b}}_2$ are linearly independent. This can only be so if $n_1 = 2$, which is to say that the mapping denoted by the underline is onto $\mathbb{R}^2$. Consider the lattice Γ in $\mathbb{R}^2$ generated by the vectors $\underline{\mathbf{b}}_1$ and $\underline{\mathbf{b}}_2$. Suppose that the basis $\{\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2\}$ is Minkowski-reduced in the (generalised) sense that $\underline{\mathbf{b}}_1$ is the shortest vector in Γ with respect to the extended norm $\|\cdot\|^*$ and $\underline{\mathbf{b}}_2$ is the shortest vector in Γ which is linearly independent of $\underline{\mathbf{b}}_1$.

First of all, $\{\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2\}$ is a reduced basis of Γ if and only if $\|\underline{\mathbf{b}}_1 \pm \underline{\mathbf{b}}_2\|^* \leqslant \|\underline{\mathbf{b}}_2\|^*$. The necessity is obvious. The sufficiency can be proved in a similar manner to Lemma 3.5. Secondly, we have the following theorem of FURTWÄNGLER.

THEOREM 5.1. *Suppose $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a (generalised) Minkowski-reduced basis of a lattice Γ of rank 2 in $\mathbb{R}^2$ with respect to a strictly convex extended norm $\|\cdot\|$. Let $\mathbf{G}$ be the matrix with columns $\mathbf{v}_1$ and $\mathbf{v}_2$. If $\|\mathbf{y}\| \leqslant \|\mathbf{v}_2\|$ for $\mathbf{y} \in \mathbb{R}^2$ then $|x_2| < 2$ where $\mathbf{x} = \mathbf{G}^{-1}\mathbf{y}$.*

PROOF. With $\mathbf{y} = \mathbf{G}\mathbf{x}$, suppose $|x_2| \geqslant 2$ but $\|\mathbf{y}\| < \|\mathbf{v}_2\|$. If $|x_1| \leqslant 1$ then, using the simple identity

$$\mathbf{e}_2 = \frac{\mathbf{x} - x_1\mathbf{e}_1}{x_2},$$

where $\mathbf{e}_i$ is the i^{th} column of the identity matrix, we have, after premultiplication of both sides by $\mathbf{G}$,

$$\mathbf{v}_2 = \frac{\mathbf{y} - x_1\mathbf{v}_1}{x_2}$$

and so

$$\|\mathbf{v}_2\| < \frac{\|\mathbf{y}\| + |x_1|\,\|\mathbf{v}_1\|}{|x_2|} < \|\mathbf{v}_2\|\,.$$

If, on the other hand, $|x_1| > 1$ then we can use the identity

$$\operatorname{sgn}(x_1)\mathbf{e}_1 + \operatorname{sgn}(x_2)\mathbf{e}_2 = \frac{\operatorname{sgn}(x_1)(|x_2| - 1)\mathbf{e}_1 + \operatorname{sgn}(x_2)(|x_1| - 1)\mathbf{e}_2 + \mathbf{x}}{|x_1| + |x_2| - 1}$$

to show that

$$\|\operatorname{sgn}(x_1)\mathbf{v}_1 + \operatorname{sgn}(x_2)\mathbf{v}_2\| < \frac{(|x_2| - 1)\,\|\mathbf{v}_1\| + (|x_1| - 1)\,\|\mathbf{v}_2\| + \|\mathbf{y}\|}{|x_1| + |x_2| - 1} < \|\mathbf{v}_2\|\,,$$

contrary to the assumption that $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a reduced basis. □

To select $\mathbf{r}$ when the basis $\{\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2\}$ is reduced, we examine the value of f_2 where $\mathbf{f} = \mathbf{G}^{-1}\underline{\mathbf{b}}_3$ and $\mathbf{G} = (\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2)$. We require that $\|\underline{\mathbf{r}}\|^* < \|\underline{\mathbf{b}}_2\|^*$. Theorem 5.1 implies that $|x_2| < 2$ where $\mathbf{x} = \mathbf{G}^{-1}\underline{\mathbf{r}}$. Now $x_2 = f_2 + a_2^*$. There are clearly at most four values of a_2^* which can cause $|x_2| < 2$ to be satisfied and these are

$$c_{2k} = \lfloor k - f_2 \rfloor, \quad k = -2, -1, 0, 1. \tag{5.8}$$

For each value of c_{2k} it then remains to find a value for c_{1k} such that, for all $j \in \mathbb{Z}$,

$$\neg L(j\mathbf{b}_1 + c_{2k}\mathbf{b}_2 + \mathbf{b}_3, c_{1k}\mathbf{b}_1 + c_{2k}\mathbf{b}_2 + \mathbf{b}_3;\ \mathbf{b}_2). \tag{5.9}$$

From among the four candidate pairs of values of c_{1k} and c_{2k} we choose a_1^* and a_2^* so that

$$\neg L(c_{1k}\mathbf{b}_1 + c_{2k}\mathbf{b}_2 + \mathbf{b}_3, a_1^*\mathbf{b}_1 + a_2^*\mathbf{b}_2 + \mathbf{b}_3;\ \mathbf{b}_2).$$

A procedure for ensuring (5.9) involves operations from calculus which are often easily solved. It involves finding the points of intersection (if any) of the mapping $\underline{\mathbf{l}}(\lambda)$ of the line

$$\mathbf{l}(\lambda) = \lambda\mathbf{b}_1 + c_{2k}\mathbf{b}_2 + \mathbf{b}_3,$$

$\lambda \in \mathbb{R}$, with the set

$$\mathcal{S} = \{\mathbf{y} \in \mathbb{R}^{n_1} \mid \|\mathbf{y}\|^* \leqslant \|\underline{\mathbf{b}}_2\|^*\}. \tag{5.10}$$

If there is no intersection for a particular value of c_{2k} or the interval over λ contains no integer value then the choice for c_{1k} is that integer which minimises $\rho h(\mathbf{l}(\lambda))$. For those values of c_{2k} which yield intersections and for which the interval over λ of the intersection contains an integer, we minimise $h\rho(\mathbf{l}(\lambda))$. This procedure, for a given value of c_{2k}, we shall refer to as *minimiseL*. For many choices of $\|\cdot\|^*$ and $\|\cdot\|^\dagger$ (for

example, the Euclidean norm or the extended norm of Example 4.3), this involves only a constant number of arithmetic operations.

Now, on the other hand, suppose that either $\{\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2\}$ forms a basis of the lattice Γ, but not a reduced basis, or that $\underline{\mathbf{b}}_2$ is linearly dependent on $\underline{\mathbf{b}}_1$. Thus, either

$$\|\underline{\mathbf{b}}_1 + \underline{\mathbf{b}}_2\| < \|\underline{\mathbf{b}}_2\| \qquad \text{or} \qquad \|\underline{\mathbf{b}}_1 - \underline{\mathbf{b}}_2\| < \|\underline{\mathbf{b}}_2\|\,.$$

Consider again the set $\mathcal{S}$ defined in (5.10). Suppose $n_1 = 2$. For every point on the boundary of $\mathcal{S}$, there is a line which passes through the point such that $\mathcal{S}$ lies on one side of the line. This is an equivalent definition of convexity for sets in $\mathbb{R}^2$ (see HARDY & WRIGHT, 1979, pp. 31–32). Consider such (parallel) lines which pass through the points $\underline{\mathbf{b}}_2$ and $-\underline{\mathbf{b}}_2$. For every point $\mathbf{y} \in \mathbb{R}^2$ which lies outside the strip bounded by the two lines we have $\|\mathbf{y}\|^* > \|\underline{\mathbf{b}}_2\|^*$. We can express this by introducing a vector $\boldsymbol{\gamma} \in \mathbb{R}^2$ with $\boldsymbol{\gamma} \cdot \underline{\mathbf{b}}_2 \geqslant 0$ such that, for all $\mathbf{y} \in \mathbb{R}^{n_1}$,

$$|\boldsymbol{\gamma} \cdot \mathbf{y}| > \boldsymbol{\gamma} \cdot \underline{\mathbf{b}}_2 \;\Rightarrow\; \|\mathbf{y}\|^* > \|\underline{\mathbf{b}}_2\|^* \tag{5.11}$$

If $n_1 = 1$ then (5.11) is trivial when $\boldsymbol{\gamma}$ (which is a scalar) is set to $\underline{\mathbf{b}}_2$. The procedure for calculating $\boldsymbol{\gamma}$ we call *rhoTangent*. For some common choices of $\|\cdot\|$ (such as a p-norm) or when $n_1 = 1$, this procedure requires only a constant number of arithmetic operations.

Just as we found a vector $\boldsymbol{\gamma}$ so that (5.11) is true using the procedure *rhoTangent*, consider an analogous procedure *hTangent* which can find a vector $\boldsymbol{\delta}$ such that, for all $\mathbf{y} \in \mathbb{R}^{n_2}$,

$$|\boldsymbol{\delta} \cdot \mathbf{y}| > \boldsymbol{\delta} \cdot \left(\overline{\mathbf{b}}_1 + \overline{\mathbf{b}}_2\right) \;\Rightarrow\; \|\mathbf{y}\|^\dagger > \left\|\overline{\mathbf{b}}_1 + \overline{\mathbf{b}}_2\right\|^\dagger.$$

Let us suppose that $\boldsymbol{\gamma} \cdot \underline{\mathbf{b}}_1 \leqslant 0$. If this is not the case, replace $\mathbf{b}_1$ by $-\mathbf{b}_1$. Furthermore, suppose for a moment that $\boldsymbol{\gamma} \cdot \underline{\mathbf{b}}_1 < 0$. It can be easily verified using (5.11) that this implies that $\|\underline{\mathbf{b}}_1 + \underline{\mathbf{b}}_2\|^* < \|\underline{\mathbf{b}}_2\|^*$. Hence, $L(\mathbf{b}_1 + \mathbf{b}_2, \mathbf{b}_2;\ \mathbf{b}_2)$. Suppose that $L(\mathbf{r}, \mathbf{b}_1 + \mathbf{b}_2;\ \mathbf{b}_2)$. This implies

$$\rho h(\mathbf{r}) < \rho h(\mathbf{b}_2) \qquad \text{and} \qquad h\rho(\mathbf{r}) < h\rho(\mathbf{b}_1 + \mathbf{b}_2). \tag{5.12}$$

Let $\boldsymbol{\xi} = \boldsymbol{\gamma}^T \underline{\mathbf{B}}$ and $\boldsymbol{\eta} = \boldsymbol{\delta}^T \overline{\mathbf{B}}$ where $\mathbf{B}$ is the matrix of the basis vectors of $\mathcal{B}$ arranged as columns. To find $\mathbf{r} = a_1^* \mathbf{b}_1 + a_2^* \mathbf{b}_2 + \mathbf{b}_3$ which satisfies (5.12), we need

$$|a_1^* \xi_1 + a_2^* \xi_2 + \xi_3| \leqslant \xi_2$$

and

$$|a_1^* \eta_1 + a_2^* \eta_2 + \eta_3| \leqslant \eta_1 + \eta_2.$$

We know that $\xi_2 \geqslant 0$ and $\eta_2 \geqslant 0$ by definition of *rhoTangent* and *hTangent*. Also, $-\xi_2 \leqslant \xi_1 < 0$ since $0 < \|\underline{\mathbf{b}}_1\|^* \leqslant \|\underline{\mathbf{b}}_2\|^*$. Finally, $\eta_1 \geqslant 0$. If this were not so then we

would have

$$\rho(\mathbf{b}_1+\mathbf{b}_2)<\rho(\mathbf{b}_2) \qquad \text{and} \qquad h(\mathbf{b}_1+\mathbf{b}_2)<\max\{h(\mathbf{b}_1),h(\mathbf{b}_2)\}. \tag{5.13}$$

which would mean that $(\mathbf{b}_1,\mathbf{b}_2)$ was not primitively (ρ,h)-minimal. With these bounds on ξ_1, ξ_2, η_1 and η_2, we can determine that

$$\frac{\xi_1\eta_3-\xi_3\eta_1+\xi_1\eta_1}{\xi_2\eta_1-\xi_1\eta_2}-1\leqslant a_2^*\leqslant\frac{\xi_1\eta_3-\xi_3\eta_1-\xi_1\eta_1}{\xi_2\eta_1-\xi_1\eta_2}+1. \tag{5.14}$$

We also find that the difference between the bounds is

$$2-\frac{2\xi_1\eta_1}{\xi_2\eta_1-\xi_1\eta_2}\leqslant 2+\frac{2}{\eta_2/\eta_1-\xi_2/\xi_1}\leqslant 4.$$

Thus, as for the case where $\{\underline{\mathbf{b}}_1,\underline{\mathbf{b}}_2\}$ is a reduced basis, we would normally only have to check the four integer values, the c_{2k}, within the interval prescribed by (5.14), determining a value for c_{1k} in each case through the procedure *minimiseL*. From these integer pairs, the values for a_1^* and a_2^* are selected. However, the combination of several conditions simultaneously could conspire to force a fifth value of c_{2k} to be checked. This could only happen if $\xi_1=-\xi_2$, $\eta_2=0$ (that is, $h(\mathbf{b}_2)=0$) and the lower and upper bounds in (5.14) are integers.[2]

We assumed above that $\boldsymbol{\gamma}\cdot\underline{\mathbf{b}}_1<0$. If, instead, $\boldsymbol{\gamma}\cdot\underline{\mathbf{b}}_1=0$ then either of the procedures (that for the reduced case or that for the non-reduced case) can be used to select $\mathbf{r}$.

The method we have described has assumed that $(\mathbf{b}_1,\mathbf{b}_2)$ is primitively (ρ,h)-minimal. However, regardless of this assumption, we have set out a method for the selection of $\mathbf{q}\in\Lambda_2(\mathcal{B})$ such that $\neg L(\mathbf{w},\mathbf{q};\ \mathbf{b}_2)$ for all $\mathbf{w}\in\Lambda_2(\mathcal{B})$. The method also selects some $\mathbf{r}\in\Lambda_3(\mathcal{B})$ such that $\neg L(\mathbf{w},\mathbf{r};\ \mathbf{b}_2)$ for all $\mathbf{w}\in\Lambda_3(\mathcal{B})$, subject to the condition that there exists some $\mathbf{v}\in\Lambda_3(\mathcal{B})$ such that $\neg L(\mathbf{q},\mathbf{v};\ \mathbf{b}_2)$ and subject to the condition that

$$\rho h(\mathbf{q})\geqslant\rho h(\mathbf{b}_2) \qquad \text{and} \qquad h\rho(\mathbf{q})\geqslant\max\{h\rho(\mathbf{b}_1),h\rho(\mathbf{b}_2)\}.$$

This last condition does not arise when we assume $(\mathbf{b}_1,\mathbf{b}_2)$ is primitively (ρ,h)-minimal, but must now be considered because the arguments leading to (5.13) do not furnish a contradiction in this case.

5.3. The Accelerated Algorithm. We are now able to set out our accelerated algorithm for finding best approximations. The algorithm assumes that the simultaneous Diophantine approximation system consists of a lattice Ω of rank 3, for which we have an initial basis matrix $\mathbf{B}=(\mathbf{b}_1,\mathbf{b}_2,\mathbf{b}_3)$, and strictly convex, complementary radius and height functions ρ and h. We require the procedures *minimiseL*, *rhoTangent* and *hTangent*, which we have already introduced above. The algorithm also uses *orderBasis*, which orders the basis $\mathbf{B}$ so that $\rho h(\mathbf{b}_1)\leqslant\rho h(\mathbf{b}_2)\leqslant\rho h(\mathbf{b}_3)$.

[2]The author believes that, even in this case, it is probably of no consequence.

ALGORITHM 5.1.

```
1  begin
2    orderBasis(B);
3    while ρ(b_1) > ε do
4          γ := rhoTangent(b_2);
5          if γ · b_1 > 0 then b_1 := −b_1 fi;
6          if L(b_1 − b_2, b_1 + b_2; b_2) then q := b_1 − b_2
7          else q := b_1 + b_2 fi;
8          if ρ(b_1 + b_2) < ρ(b_2) then
9            δ := hTangent(b̄_1 + b̄_2);
10           ξ := γ^T B; η := δ^T B̄;
11           μ := (ξ_1η_3 − ξ_3η_1)/(ξ_2η_1 − ξ_1η_2);
12         else
13           G := (b_1, b_2);
14           f := G^{−1} b_3;
15           μ := −f_2;
16         fi;
17         r := b_3;
18         for k := −2 to 2 do
19             c_2k := ⌊μ + k⌋;
20             c_1k := minimiseL(c_2k, B);
21             if L(c_1k b_1 + c_2k b_2 + b_3, r; b_2) then
22               r := c_1k b_1 + c_2k b_2 + b_3 fi;
23         od;
24         if L(q, r; b_2) then b_2 := q; output(b_2)
25         else b_3 := r; output(b_3) fi;
26         orderBasis(B);
27   od;
28 end.
```

5.4. Analysis of the Accelerated Algorithm.

PROPOSITION 5.1. *Consider a simultaneous Diophantine approximation system consisting of a lattice Ω of rank 3 in $\mathbb{R}^m$ and strictly convex, complementary radius and height functions ρ and h. Suppose, at line 3 of Algorithm 5.1, $\mathbf{B}$ is a basis matrix of Ω and $(\mathbf{b}_1, \mathbf{b}_2)$ is primitively (ρ, h)-minimal. Then there exists a sequence of (ρ, h)-minimal sets, each an incremental successor of the previous one, such that when the primitively (ρ, h)-minimal sets of the first two elements of each (ρ, h)-minimal set are put in sequence and duplicates are removed, $(\mathbf{b}_1, \mathbf{b}_2)$ takes the value of each primitively (ρ, h)-minimal set in turn at line 3 on every subsequent iteration until the algorithm terminates.*

PROOF. The proof is by inspection of the algorithm and its reconciliation with the principles discussed in Section 5.2. □

It remains to show that the algorithm can produce a basis containing a primitively (ρ, h)-minimal set given an arbitrary basis.

PROPOSITION 5.2. *Consider a simultaneous Diophantine approximation system as in Proposition 5.1. If, at line 3 of Algorithm 5.1,* $\mathbf{B}$ *is a basis matrix of* Ω *for which* $(\mathbf{b}_1, \mathbf{b}_2)$ *is not primitively* (ρ, h)*-minimal then, on the subsequent iteration, the new values of* $\mathbf{b}_1$ *and* $\mathbf{b}_2$ *at line 3, denoted* $\mathbf{b}_1'$ *and* $\mathbf{b}_2'$*, satisfy*

$$\rho h(\mathbf{b}_2') \leqslant \rho h(\mathbf{b}_2)$$

and

$$\max\{h\rho(\mathbf{b}_1'), h\rho(\mathbf{b}_2')\} \leqslant \max\{h\rho(\mathbf{b}_1), h\rho(\mathbf{b}_2)\}$$

and one of these inequalities is satisfied strictly.

Moreover, after a finite number of iterations, Algorithm 5.1 either produces a basis at line 3 in which the first two elements constitute a primitively (ρ, h)*-minimal set or the algorithm terminates.*

PROOF. A straightforward consequence of Theorem 3.12 is that $(\mathbf{b}_1, \mathbf{b}_2)$ is primitively (ρ, h)-minimal if and only if

$$\rho h(\mathbf{w}) \geqslant \rho h(\mathbf{b}_2) \tag{5.15}$$

or

$$h\rho(\mathbf{w}) \geqslant \max\{h\rho(\mathbf{b}_1), h\rho(\mathbf{b}_2)\} \tag{5.16}$$

for all $\mathbf{w} \in \Lambda_2(\mathcal{B})$ and $\mathbf{w} \in \Lambda_3(\mathcal{B})$. If there is an element of $\Lambda_2(\mathcal{B})$ which fails to satisfy both (5.15) and (5.16) then $\mathbf{q}$ is assigned such an element at line 6 or 7. Lines 8 to 23 represent a formal description of the principles discussed in Section 5.2 for selecting $\mathbf{r} \in \Lambda_3(\mathcal{B})$. Therefore, if there exists an element of $\Lambda_3(\mathcal{B})$ which fails to satisfy both (5.15) and (5.16), where there is no such element in $\Lambda_2(\mathcal{B})$, then $\mathbf{r}$ is assigned such an element.

Because there are only finitely many lattice points which satisfy neither (5.15) nor (5.16), and this number strictly decreases from iteration to iteration, we conclude that the algorithm must produce an ordered basis in which $(\mathbf{b}_1, \mathbf{b}_2)$ is primitively (ρ, h)-minimal. □

Therefore, we can be assured that, for any best approximation $\mathbf{p}$ such that $\epsilon < \rho(\mathbf{p}) \leqslant \min\{\rho(\mathbf{b}_1), \rho(\mathbf{b}_2), \rho(\mathbf{b}_3)\}$ then, after a finite number of iterations, the algorithm will output $\mathbf{p}$ or an equivalent lattice point.

Algorithm 5.1 can be thought of as a generalisation of Furtwängler's algorithm. The notion of a primitively (ρ, h)-minimal set is a generalisation of FURTWÄNGLER's

approximation prism to a broader class of radius and height functions. As a result, it is not restricted to simultaneous Diophantine approximation using the sup-norm, nor is it restricted to simultaneous Diophantine approximation in the "traditional" sense, by which we mean approximation of a line by lattice points. It can also find best approximate integer relations and shortest integer relations, if they exist.

An unfortunate aspect of both Algorithm 4.1 and Algorithm 5.1 is that we cannot predict *a priori* how many intermediate calculations must be performed before a new best approximation is found. The best we can say is that, at any point in these algorithms, the number of iterations through the main loop that remain to be performed before a new best approximation is found can be bounded by a function of the radii of the basis elements, using the pigeonhole principle. In particular, we have the bound (3.14) in Theorem 3.4. For Algorithm 5.1, we can improve this bound by replacing $\mathbf{b}_3$ with $\mathbf{b}_2$.

Finally, we observe that the procedure *minimiseL* could be used (twice) in Algorithm 4.1 to replace the (two) loops on lines 14–17. This should result in some improvement in the speed of the algorithm.

5.5. Numerical Examples. We now present some numerical examples, drawn from FURTWÄNGLER (1927) and BRENTJES (1981), in order to demonstrate the correctness of Algorithm 5.1, and present some data which indicates that the time required by the algorithm to find best approximations of a given radius may be logarithmic in the inverse of the radius.

We begin with the two examples used by FURTWÄNGLER (1927). For clarity of exposition, the lattice used in each example is $\mathbb{Z}^3$. The examples are differentiated by the radius and height functions used. This is in contrast to the way the examples were originally presented. There, the examples used the same radius and height functions but different lattices.

Both of FURTWÄNGLER's examples use the sup-norm in the radius function (since his algorithm is formulated on that premise). However, the sup-norm is not strictly convex. Therefore, we replace the sup-norm with an extended norm which is extended from it. Consider the SORT NORM which we define for vectors $\mathbf{v} \in \mathbb{R}^n$ as

$$\|\mathbf{v}\|_s = \operatorname{sort}\{|v_1|, |v_2|, \ldots, |v_n|\}$$

where, as in Example 4.3, the function $\operatorname{sort}\{\cdot\}$ sorts its arguments in descending order. This extended norm is strictly convex. This difference aside, FURTWÄNGLER's algorithm and Algorithm 5.1 can be expected to produce identical outputs for the examples cited here because of the conjunction of the notions of a primitively (ρ, h)-minimal set and an approximation prism.

EXAMPLE 5.1. In this example, we execute Algorithm 5.1 with the identity matrix as its initial basis. The radius and height functions in this example are

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_s \qquad \text{and} \qquad h(\mathbf{x}) = |\boldsymbol{\alpha}\cdot\mathbf{x}|$$

where

$$\mathbf{P}^T = \begin{pmatrix} 1 & 0 & -\sqrt[3]{2} \\ 0 & 1 & -\sqrt[3]{4} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\alpha} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}.$$

TABLE 4. Outputs and important variables of Algorithm 5.1 in FURTWÄNGLER's first example.

It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{P}^T\mathbf{v}$
1	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$	
2	$(\mathbf{1},\mathbf{2},\mathbf{1})$	$(0,1,0)$	$(0,0,1)$	$(-0.260, 0.413)$
3	$(1,2,1)$	$(\mathbf{1},\mathbf{1},\mathbf{1})$	$(0,0,1)$	$(-0.260, -0.587)$
4	$(1,2,1)$	$(\mathbf{3},\mathbf{3},\mathbf{2})$	$(1,1,1)$	$(0.480, -0.175)$
5	$(\mathbf{4},\mathbf{5},\mathbf{3})$	$(1,2,1)$	$(1,1,1)$	$(0.220, 0.238)$
6	$(4,5,3)$	$(\mathbf{5},\mathbf{6},\mathbf{4})$	$(1,2,1)$	$(-0.0397, -0.350)$
7	$(4,5,3)$	$(\mathbf{6},\mathbf{8},\mathbf{5})$	$(5,6,4)$	$(-0.300, 0.0630)$
8	$(\mathbf{9},\mathbf{11},\mathbf{7})$	$(4,5,3)$	$(6,8,5)$	$(0.181, -0.112)$
9	$(\mathbf{15},\mathbf{19},\mathbf{12})$	$(9,11,7)$	$(4,5,3)$	$(-0.119, -0.0488)$
10	$(15,19,12)$	$(\mathbf{24},\mathbf{30},\mathbf{19})$	$(4,5,3)$	$(0.0615, -0.161)$
11	$(15,19,12)$	$(\mathbf{34},\mathbf{43},\mathbf{27})$	$(24,30,19)$	$(-0.177, 0.141)$
12	$(15,19,12)$	$(\mathbf{49},\mathbf{62},\mathbf{39})$	$(24,30,19)$	$(-0.137, 0.0914)$
13	$(\mathbf{58},\mathbf{73},\mathbf{46})$	$(15,19,12)$	$(49,62,39)$	$(0.0436, -0.204)$

Table 4 shows the output of Algorithm 5.1 for these inputs with $\epsilon = (0.1, 0)$. The format of Table 4 is similar to that which we used in Table 1 of Example 4.1. The table lists the state of $\mathcal{B}$ at line 3 at the beginning of each iteration of the algorithm for the first 13 iterations. The first column lists the iteration number. The next three columns list the values of the basis vectors. The vector in bold face is the innovation into the basis, which we denote $\mathbf{v}$. The rightmost column lists $\mathbf{P}^T\mathbf{v}$. The radius and height are not listed, but they are easily calculated. The height of $\mathbf{v}$ is simply v_3 and its radius is obtained by sorting the absolute values of $\mathbf{P}^T\mathbf{v}$.

Notice that, in contrast to Algorithm 4.1, the innovations into the basis always occur in the first or second position. This is because Algorithm 5.1 "skips" the intermediate innovations that would occur in the third position.

EXAMPLE 5.2. In this example, we again set

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_s \qquad \text{and} \qquad h(\mathbf{x}) = |\boldsymbol{\alpha}\cdot\mathbf{x}|$$

but with

$$\mathbf{P}^T = \begin{pmatrix} 1 & 0 & -\zeta \\ 0 & 1 & -\zeta^2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\alpha} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$

where $\zeta = 1.3248\ldots$ is the unique real solution of the equation $x^3 - x - 1 = 0$. The output of the algorithm for these inputs is shown in Table 5.

TABLE 5. Output and important variables of Algorithm 5.1 for FURTWÄNGLER's second example.

It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{P}^T\mathbf{v}$
1	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$	
2	$(\mathbf{1},\mathbf{2},\mathbf{1})$	$(0,1,0)$	$(0,0,1)$	$(-0.325, 0.245)$
3	$(1,2,1)$	$(\mathbf{2},\mathbf{2},\mathbf{1})$	$(0,1,0)$	$(0.675, 0.245)$
4	$(1,2,1)$	$(\mathbf{3},\mathbf{4},\mathbf{2})$	$(0,1,0)$	$(0.351, 0.490)$
5	$(\mathbf{4},\mathbf{5},\mathbf{3})$	$(1,2,1)$	$(3,4,2)$	$(0.0258, -0.265)$
6	$(4,5,3)$	$(\mathbf{5},\mathbf{7},\mathbf{4})$	$(3,4,2)$	$(-0.299, -0.0195)$
7	$(4,5,3)$	$(\mathbf{9},\mathbf{12},\mathbf{7})$	$(3,4,2)$	$(-0.273, -0.284)$
8	$(\mathbf{12},\mathbf{16},\mathbf{9})$	$(4,5,3)$	$(9,12,7)$	$(0.0775, 0.206)$
9	$(\mathbf{16},\mathbf{21},\mathbf{12})$	$(12,16,9)$	$(9,12,7)$	$(0.103, -0.0585)$
10	$(16,21,12)$	$(\mathbf{21},\mathbf{28},\mathbf{16})$	$(12,16,9)$	$(-0.195, -0.0780)$
11	$(16,21,12)$	$(\mathbf{28},\mathbf{37},\mathbf{21})$	$(21,28,16)$	$(0.181, 0.148)$
12	$(16,21,12)$	$(\mathbf{33},\mathbf{44},\mathbf{25})$	$(28,37,21)$	$(-0.118, 0.128)$
13	$(\mathbf{49},\mathbf{65},\mathbf{37})$	$(16,21,12)$	$(28,37,21)$	$(-0.0146, 0.0695)$

The outputs in Example 5.1 and Example 5.2 correspond almost exactly with those found by FURTWÄNGLER himself. However, there are a few minor differences. In the first example, FURTWÄNGLER erroneously lists $(2,2,1)$ as an output and fails to list $(6,8,5)$. In the second example, FURTWÄNGLER erroneously lists $(1,1,1)$. These errors are easily verified as such by calculation of their radii and heights.

To demonstrate the algorithm with another radius function, we return to the example of BRENTJES. We used this example to demonstrate Algorithm 4.1 in Example 4.1.

EXAMPLE 5.3. In this case, we have

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_2 \qquad \text{and} \qquad h(\mathbf{x}) = |\boldsymbol{\alpha}\cdot\mathbf{x}|$$

where $\mathbf{P}^T$ is given by (4.12) and $\boldsymbol{\alpha} = (1,0,0)$. The initial basis used is the identity.

The output of the algorithm for $\epsilon = 0.1$ is presented in Table 6.

Notice that, in this example, the final best approximation is found in seven fewer iterations than was required for Algorithm 4.1.

TABLE 6. Output and important variables of Algorithm 5.1 in BRENTJES' example.

It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{P}^T\mathbf{v}$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$			
2	$(\mathbf{1},\mathbf{2},\mathbf{3})$	$(0,1,0)$	$(0,0,1)$	$(0.290, 0.0760)$	0.300	1
3	$(1,2,3)$	$(\mathbf{1},\mathbf{1},\mathbf{3})$	$(0,0,1)$	$(-0.710, 0.0760)$	0.714	1
4	$(1,2,3)$	$(\mathbf{2},\mathbf{3},\mathbf{6})$	$(0,0,1)$	$(-0.420, 0.152)$	0.447	2
5	$(\mathbf{3},\mathbf{5},\mathbf{9})$	$(1,2,3)$	$(0,0,1)$	$(-0.130, 0.228)$	0.262	3
6	$(\mathbf{10},\mathbf{17},\mathbf{29})$	$(3,5,9)$	$(1,2,3)$	$(-0.100, -0.240)$	0.260	10
7	$(\mathbf{11},\mathbf{19},\mathbf{32})$	$(10,17,29)$	$(3,5,9)$	$(-0.190, 0.164)$	0.251	11
8	$(\mathbf{13},\mathbf{22},\mathbf{38})$	$(11,19,32)$	$(10,17,29)$	$(-0.230, -0.0122)$	0.230	13
9	$(\mathbf{14},\mathbf{24},\mathbf{41})$	$(13,22,38)$	$(11,19,32)$	$(0.0603, 0.0638)$	0.0878	14

EXAMPLE 5.4. To demonstrate the ability of the algorithm to generate a primitively (ρ, h)-minimal set from an arbitrary basis, we again use BRENTJES' example, but now we use a pseudo-randomly generated unimodular matrix instead of the identity matrix. We use the initial basis

$$\mathcal{B} = \begin{pmatrix} 13 & 16 & 2 \\ 24 & 29 & 4 \\ 54 & 65 & 9 \end{pmatrix}.$$

The outputs of the algorithm are presented in Table 7.

TABLE 7. Outputs and important variables of Algorithm 5.1 in BRENTJES' example with a pseudo-random initial basis.

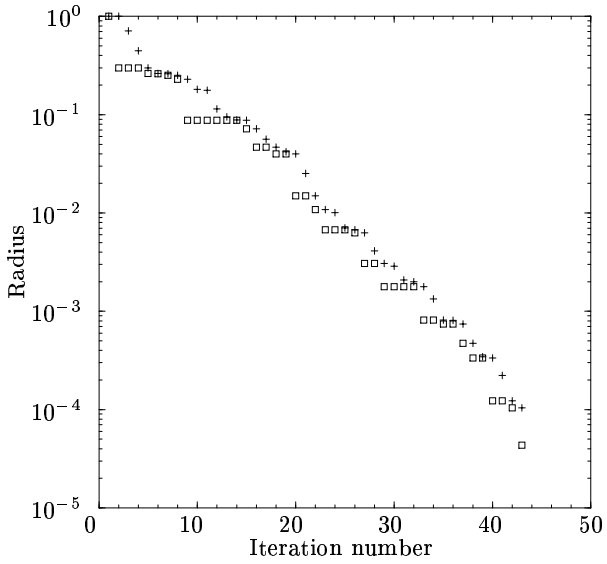
It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{P}^T\mathbf{v}$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(2,4,9)$	$(13,24,54)$	$(16,29,65)$			
2	$(\mathbf{0},\mathbf{1},\mathbf{2})$	$(2,4,9)$	$(13,24,54)$	$(1,2)$	2.24	0
3	$(\mathbf{1},\mathbf{1},\mathbf{3})$	$(0,1,2)$	$(2,4,9)$	$(-0.710, 0.0760)$	0.714	1
4	$(1,1,3)$	$(\mathbf{0},\mathbf{1},\mathbf{1})$	$(0,1,2)$	$(1,1)$	1.414	0
5	$(1,1,3)$	$(\mathbf{0},\mathbf{0},\mathbf{1})$	$(0,1,1)$	$(0,1)$	1	0
6	$(1,1,3)$	$(\mathbf{0},\mathbf{1},\mathbf{0})$	$(0,0,1)$	$(1,0)$	1	0
7	$(\mathbf{1},\mathbf{2},\mathbf{3})$	$(1,1,3)$	$(0,1,0)$	$(0.290, 0.0760)$	0.300	1
8	$(1,2,3)$	$(\mathbf{2},\mathbf{3},\mathbf{6})$	$(0,1,0)$	$(-0.420, 0.152)$	0.447	2

In this example, the algorithm has found the primitively (ρ, h)-minimal set

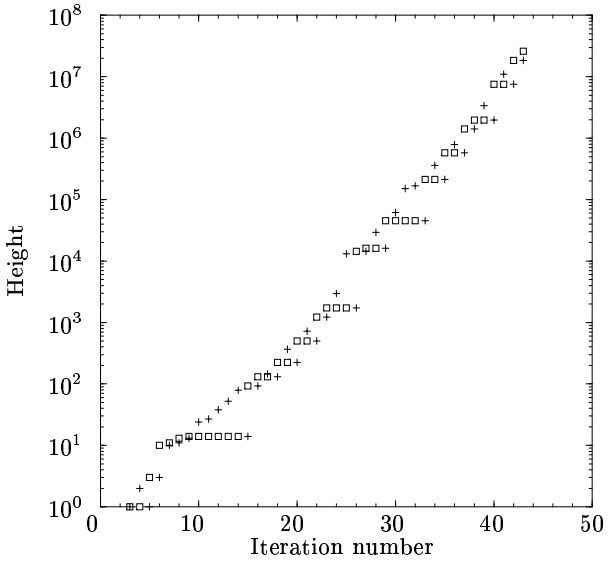
$$(\mathbf{b}_1, \mathbf{b}_2) = ((1,2,3),(1,1,3))$$

by the 7^{th} iteration. We know that this set is primitively (ρ, h)-minimal because the innovation $\mathbf{v} = (2,3,6)$ has $h\rho(\mathbf{v}) > \max\{h\rho(\mathbf{b}_1), h\rho(\mathbf{b}_2)\}$.

EXAMPLE 5.5. Finally, we present two plots in Figure 3 which show the speed with which the algorithm decreases the radii and increases the height of its approximations. The input to the algorithm used to generate the plots was again



(a) Radii of $\mathbf{b}_1$ and $\mathbf{b}_2$.



(b) Heights of $\mathbf{b}_1$ and $\mathbf{b}_2$.

FIGURE 3. Plots of decrease in radius and increase in height of the basis elements $\mathbf{b}_1$ ($\square$) and $\mathbf{b}_2$ ($+$) against the iteration number for BRENTJES' example.

BRENTJES' example with $\epsilon = 10^{-4}$. It appears that, for this input and for many others tested, the algorithm is able to find a best approximation with radius less

than ϵ in $O(\log \epsilon^{-1})$ iterations. However, its true complexity is unknown to the author at present.

6. Algorithms for Lattices of Higher Rank

6.1. Introductory Remarks. The algorithms we have developed in this chapter are limited to lattices of rank 2 and 3. It is not easy to see how these algorithms can be generalised to lattices of higher rank since the properties on which the algorithms rely do not carry over. In any case, it is almost certain that the computational complexity of finding best simultaneous Diophantine approximations will be very high. As we mentioned in the introduction to the chapter, LAGARIAS (1982) has shown that certain problems of this type are $\mathfrak{NP}$-hard.

Therefore, we seek algorithms which are able to give "good" simultaneous Diophantine approximations in a reasonable amount of time, say, in a time bounded by a polynomial of the input size. Before we describe in detail the recent discoveries which make this possible, it is instructive to review progress towards this goal.

The search for a higher-dimensional analogue of Euclid's algorithm has quite a long history. The algorithms which have been proposed usually go by the names "multi-dimensional Euclidean algorithm" or "multi-dimensional continued fraction algorithm." Historically, most algorithms have been developed solely for the approximation of a line by lattice points, rather than for the more general class of simultaneous Diophantine approximation systems we have proposed in Section 2.2. More recently, there has also been an emphasis on the problem of finding integer relations, which is a dual problem of the approximation of a line, as we shall see.

What constitutes a "multi-dimensional continued fraction algorithm" is not universally agreed. One possible, and rather loose, definition is that it is an algorithm which approximates a linear form (or simply line) by performing simple basis transformations, such as the replacement of one basis vector at a time with another. The aim of such algorithms has not solely been to produce best approximations. Authors of multi-dimensional continued fraction algorithms have frequently placed equal or greater weight on other characteristics of simple continued fractions which they hope to carry over into higher dimensions, such as the ability to uniquely describe the input according to a string of characters from a given alphabet (as the input of the s.c.f is uniquely described by the string of partial quotients) or the exhibition of periodicity for certain inputs (such as the s.c.f does for quadratic irrationalities; a property we have not examined in this thesis).

JACOBI (1868) was the first to propose a generalisation of Euclid's algorithm, although only to lattices of rank 3. However, PERRON (1907) proposed a further generalisation of this algorithm to lattices of arbitrary rank. BRUN (1919, 1920)[3] proposed another generalisation of Euclid's algorithm to arbitrary rank. We will

[3]BRUN's work is known to the author only through the description of BRENTJES (1981) and other secondary sources.

briefly describe his algorithm in the next subsection. A great many variations of these algorithms have since been proposed, but all have been of a rather *ad hoc* nature until quite recently, with the announcement by FERGUSON & FORCADE (1979) of their algorithm.

A problem with these algorithms has been to show that they are strongly convergent. We define a STRONGLY CONVERGENT algorithm as one that either finds a basis for the lattice in a finite number of iterations which contains a strict subset of vectors such that all points $\mathbf{x}$ in the lattice with $\rho(\mathbf{x}) = 0$ are linearly dependent on the vectors of the subset or

$$\lim_{j\to\infty} \max_{i=1,2,\ldots,n} \{\rho(\mathbf{b}_i^{(j)})\} = 0 \tag{6.1}$$

where $\mathbf{b}_i^{(j)}$ is the i^{th} basis vector of the basis produced by the algorithm on the j^{th} iteration and n is the rank of the lattice.

Let us consider this definition for a moment. Suppose, for some lattice Ω of rank n, there exists a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ such that if $\rho(\mathbf{x}) = \mathbf{0}$ and $\mathbf{x}$ lies in the real span of Ω then $\mathbf{x}$ is a linearly dependent on $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s\}$ where $s < n$. In this case, it can be shown that there exists come constant $\mathbf{c} > \mathbf{0}$ such that $\rho(\mathbf{v}) \geqslant \mathbf{c}$ whenever $\mathbf{v}$ is a lattice point which is independent of $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s$. Therefore, (6.1) could not possibly be achieved.

Consider the special case where $\Omega = \mathbb{Z}^n$ and the only points $\mathbf{x} \in \mathbb{R}^n$ such that $\rho(\mathbf{x}) = \mathbf{0}$ belong to the line $\mathbb{R}\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$. This is a problem of approximation of a line by lattice points: "traditional" simultaneous Diophantine approximation. Suppose also that we can find a basis of the type described above. That is, suppose there exists a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ such that $\boldsymbol{\alpha}$ is linearly dependent on $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s\}$, $s < n$. In this special case, the DUAL LATTICE of $\Omega = \mathbb{Z}^n$, the lattice generated by the rows of the inverse matrix of any basis matrix of Ω, is again Ω. Writing $\mathbf{C}^T = \mathbf{B}^{-1}$, where as usual $\mathbf{B}$ is the basis matrix corresponding to $\mathcal{B}$, then $\mathbf{C}$ is also a basis matrix of Ω. Moreover, the vectors $\mathbf{c}_{s+1}, \mathbf{c}_{s+2}, \ldots, \mathbf{c}_n$ are all orthogonal to the vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s$ and hence to $\boldsymbol{\alpha}$ also. Thus, the vectors $\mathbf{c}_{s+1}, \mathbf{c}_{s+2}, \ldots, \mathbf{c}_n$ are integer relations for $\boldsymbol{\alpha}$. The converse is also true: if there is an integer relation for $\boldsymbol{\alpha}$ then there exists a basis of the type described. Therefore, a strongly convergent algorithm applied to a simultaneous Diophantine system of this type will either find integer relations to $\boldsymbol{\alpha}$ or find increasingly good approximations to the line $\mathbb{R}\boldsymbol{\alpha}$. This explains the duality of the problem of approximation of a line by lattice points and the problem of finding approximate integer relations.

FERGUSON & FORCADE (1979, 1982) were the first to discover a strongly convergent algorithm. The algorithm was set up to find simultaneous Diophantine approximations of a line with respect to the sup-norm radius function. It is unknown whether the algorithm can find an approximation with radius less than some prescribed constant within a time bounded by a polynomial of the input size. BERGMAN (1980) proposed a variant of the algorithm of FERGUSON & FORCADE which found

approximations with respect to the Euclidean norm. This algorithm bears a striking resemblance to the later LLL algorithm of LENSTRA *et al.* (1982). The relationship between these two algorithms was clarified and the algorithms of BERGMAN and FERGUSON & FORCADE further developed by HASTAD *et al.* (1989). Their algorithms were aimed, in the first instance, at finding integer relations. They were able to show, using the ideas of LENSTRA *et al.*, that the algorithm could find short integer relations or prove that none exist in polynomial time. We will describe in detail their "Short Integer Relation Algorithm," which we call the HJLS algorithm, and the very similar PSLQ algorithm of FERGUSON & BAILEY (1991) and FERGUSON *et al.* (1996), in Section 6.3.

Finally in this section, we present some numerical examples comparing the outputs of these algorithms in lattices of rank 3 with the accelerated algorithm (Algorithm 5.1) we developed earlier in the chapter.

6.2. Brun's Algorithm. Brun's algorithm is a quite natural generalisation of Euclid's algorithm to lattices of higher rank. That it is a natural generalisation of Euclid's algorithm is attested by the fact that, as BRENTJES (1981) notes, the algorithm has been independently rediscovered by various authors many times since its original publication in 1919, and by the present author too! Brun's algorithm is specific to approximation of a line by lattice points. It doesn't appear that it was conceived with any particular radius function in mind, although BRUN proved certain (weak) convergence properties using the Euclidean norm.

We seek here to give Brun's algorithm a geometric interpretation. Consider a lattice Ω of rank n in $\mathbb{R}^n$ and a line $\mathbf{l}(\lambda) = \lambda\boldsymbol{\alpha}$, $\lambda \in \mathbb{R}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$, that is to be approximated by points of Ω. We are given a basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ of Ω. Now, $\boldsymbol{\alpha}$ is linearly dependent on the basis vectors so we can write

$$\boldsymbol{\alpha} = \mu_1\mathbf{b}_1 + \mu_2\mathbf{b}_2 + \cdots + \mu_n\mathbf{b}_n$$

where $\mu_1, \mu_2, \ldots, \mu_n \in \mathbb{R}$. Let us assume that each of the μ_i are non-negative. If this is not the case then, where $\mu_i < 0$, replace $\mathbf{b}_i$ with $-\mathbf{b}_i$. Furthermore, let us assume that at least two of the μ_i are non-zero. If all are zero then $\mathbf{l}(\lambda)$ does not represent a line. If only one of the μ_i is non-zero then one of the basis vectors lies on the line, so there can be no further best approximations.

Construct the parallelepiped with vertices at $\mathbf{0}$ and $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$. The line $\mathbf{l}(\lambda)$ intersects this parallelepiped at two points. The first point, as we increase λ, is the origin. Increasing λ further, the line passes through the parallelepiped and at last exits through one of its faces. The face through which the line exits is that face with vertices at $\mathbf{b}_s, \mathbf{b}_s + \mathbf{b}_1, \mathbf{b}_s + \mathbf{b}_2, \ldots, \mathbf{b}_s + \mathbf{b}_{s-1}, \mathbf{b}_s + \mathbf{b}_{s+1}, \ldots, \mathbf{b}_n$ where s is that index which maximises μ_i, $i = 1, 2, \ldots, n$. Now, consider placing another parallelepiped, displaced by $\mathbf{b}_s$ from the origin, so that it adjoins the original parallelepiped at the face through which the line exits. Consider the face through which the line exits in

the new parallelepiped. We can quickly confirm that it is the face opposite the face the line entered if $\mu_s > 2\mu_t$, where t is the index which maximises μ_i, $i = 1, 2, \ldots, n$, $i \neq s$. If this is the case, we adjoin other parallelepipeds successively in the same way until eventually $\mu_s \leqslant r\mu_t$ where r is the number of parallelepiped that have been so adjoined. The face through which the line now exits is that face with vertices at $r\mathbf{b}_s+\mathbf{b}_t, r\mathbf{b}_s+\mathbf{b}_t+\mathbf{b}_1, r\mathbf{b}_s+\mathbf{b}_t+\mathbf{b}_2, \ldots, r\mathbf{b}_s+\mathbf{b}_t+\mathbf{b}_{t-1}, r\mathbf{b}_s+\mathbf{b}_t+\mathbf{b}_{t+1}, \ldots, r\mathbf{b}_s+\mathbf{b}_t+\mathbf{b}_n$.

Therefore, if we replace $\mathbf{b}_t$ by $\mathbf{b}_t + r\mathbf{b}_s$ in the basis then, when we construct the parallelepiped as we have described from the new basis vectors, the line again intersects its body. Brun's algorithm is then to repeat this process until a sufficiently good approximation is found.

We illustrate this geometric interpretation in Figure 4. Here, we illustrate an

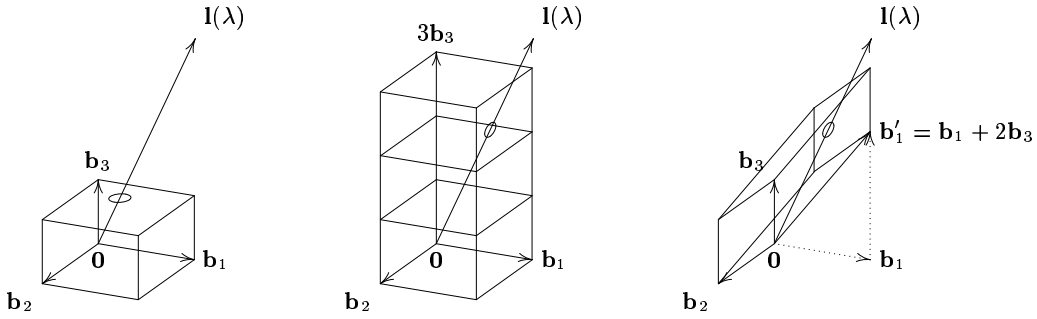


FIGURE 4. An iteration of Brun's algorithm on a lattice of rank 3.

iteration of Brun's algorithm on a lattice of rank 3. The line $\mathbf{l}(\lambda)$ to be approximated has the form $\mathbf{l}(\lambda) = \lambda\boldsymbol{\alpha}$ and

$$\boldsymbol{\alpha} = 0.4\mathbf{b}_1 + 0.3\mathbf{b}_2 + \mathbf{b}_3.$$

That is, $\mu_1 = 0.4$, $\mu_2 = 0.3$ and $\mu_3 = 1$. The line exits the parallelepiped $\mathbf{0}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ through the face $\mathbf{b}_3, \mathbf{b}_3 + \mathbf{b}_1, \mathbf{b}_3 + \mathbf{b}_2$. This is illustrated by the "hole" in the top face of the parallelepiped in the diagram at left in Figure 4. If we "stack" a similar parallelepiped on top of the first one then the line still exits through the top face. When we stack a third parallelepiped on to the first two then, as illustrated in the middle diagram, the line exits through the "side" face $2\mathbf{b}_3+\mathbf{b}_1, 2\mathbf{b}_3+\mathbf{b}_1+\mathbf{b}_2, 3\mathbf{b}_3+\mathbf{b}_1$. The diagram at right shows the new parallelepiped formed after $\mathbf{b}_1$ is replaced with $\mathbf{b}_1 + 2\mathbf{b}_3$ in the basis. Clearly, the same procedure can now be applied again to the new basis.

The following is a formal expression of the algorithm we have described.

ALGORITHM 6.1.

1 **begin**
2 $\boldsymbol{\mu} := \mathbf{B}^{-1}\boldsymbol{\alpha}$;
3 **for** $i := 1$ **to** n **do**
4 **if** $\mu_i < 0$ **then**

```
5        μ_i := −μ_i; b_i := −b_i fi;
6    od;
7    while min_{i=1,2,...,n} {ρ(b_i)} > ε do
8        s := arg max {μ_1, μ_2, ..., μ_n};
9        t := arg max {μ_1, μ_2, ..., μ_{s−1}, μ_{s+1}, ..., μ_n};
10       r := ⌊μ_s / μ_t⌋;
11       b_t := b_t + r b_s;
12       μ_s := μ_s − r μ_t;
13   od
14 end.
```

We conclude this subsection by summarising some results concerning the algorithm which were reported by Brentjes (1981). Brun showed that the algorithm is weakly convergent for lattices of rank 3 in that

$$\max\{\rho(\mathbf{b}_1'), \rho(\mathbf{b}_2'), \rho(\mathbf{b}_3')\} \leqslant \max\{\rho(\mathbf{b}_1), \rho(\mathbf{b}_2), \rho(\mathbf{b}_3)\}$$

on each iteration of the algorithm. Furthermore, he showed that the expansion enjoys a uniqueness property for lattices of rank 3, in that a string of characters from a certain alphabet can be used to represent the sequence of basis transformations performed by the algorithm and this string of characters is unique for any line (up to permutation of indices). Moreover, he showed that any string of characters from the alphabet uniquely describes a line with respect to the initial basis. Brentjes reports that Greiter (1977) has been able to extend these results (albeit with a slightly weaker notion of convergence) to lattices of any rank.

However, it is certainly not true that the algorithm is strongly convergent. Brentjes provides counterexamples to prove this. The first algorithm which was proved to be strongly convergent was discovered by Ferguson & Forcade (1979, 1982). We now discuss the closely-related HJLS algorithm.

6.3. The HJLS Algorithm and Its Variants. The paper of Hastad *et al.* (1989) contains a number of algorithms for finding integer relations and more general simultaneous Diophantine approximation problems. Algorithms are developed both for the arithmetic and bit complexity models. In this subsection, we will discuss one of these algorithms, which they call the "Short Integer Relation Algorithm." We refer to this algorithm as the HJLS algorithm. The HJLS algorithm is very similar to an algorithm proposed by Bergman (1980), which draws heavily on ideas used by Ferguson & Forcade (1979, 1982) for their algorithm. As such, it shares with them the property that it is strongly convergent. However, they extend the analysis, using ideas of Lenstra *et al.* (1982), to show that an integer relation for the input with length less than ϵ^{-1} can be proved not to exist, or a relation found with length

less that $2^{n/2-1}\epsilon^{-1}$, in an amount of time which is bounded by a polynomial in the rank of the lattice and $\log \epsilon^{-1}$.

Let us now formulate the algorithm for the simultaneous Diophantine approximation system consisting of the lattice $\mathbb{Z}^n$ in $\mathbb{R}^n$ and a radius function ρ and height function h which can be expressed

$$\rho(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_2 \qquad \text{and} \qquad h(\mathbf{x}) = |\boldsymbol{\alpha}\cdot\mathbf{x}|$$

where $\mathbf{P}^T$ is an $n-1\times n$ matrix and $\{\boldsymbol{\alpha}, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{n-1}\}$ is an orthonormal basis of $\mathbb{R}^n$. Therefore, a good approximation in this system is an integer vector which lies close to the line $\mathbb{R}\boldsymbol{\alpha}$ and an integer relation for $\boldsymbol{\alpha}$ exists if there is any non-zero lattice point $\mathbf{v}$ such that $h(\mathbf{v}) = 0$.

Let us denote by an underline a mapping of a vector by $\mathbf{P}^T$ and by an overline its mapping by $\boldsymbol{\alpha}^T$ so that $\underline{\mathbf{x}} = \mathbf{P}^T\mathbf{x}$ and $\overline{\mathbf{x}} = \boldsymbol{\alpha}\cdot\mathbf{x}$.

Consider the **QR** decomposition of $\underline{\mathbf{B}}$, where $\mathbf{B}$ is a basis matrix for $\mathbb{Z}^n$. Since $\underline{\mathbf{B}}$ does not have full column rank, consider the **QR** decomposition described for such matrices in Section 4. Recall that this means we can write $\underline{\mathbf{B}} = \mathbf{QR}$ where $\mathbf{Q}$ is an $(n-1)\times n$ column orthogonal matrix and $\mathbf{R}$ is an $n\times n$ upper triangular matrix with non-negative diagonal elements. One of the diagonal elements of $\mathbf{R}$ must be zero.

If $r_{j,j} = 0$ for some $1 \leqslant j < n$ then $\boldsymbol{\alpha}$ is a linear combination $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_j$. As we discussed in Section 6.1, this implies that $\mathbf{c}_{j+1}, \mathbf{c}_{j+2}, \ldots, \mathbf{c}_n$ are integer relations for $\boldsymbol{\alpha}$ where $\mathbf{C}^T = \mathbf{B}^{-1}$. Furthermore, there can only be one index j for which it is true that $r_{j,j} = 0$, for otherwise the basis vectors would not be linearly independent. Therefore, if $r_{n,n} \neq 0$ then $\mathbf{C}$ contains at least one integer relation.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have the identity

$$\mathbf{x}\cdot\mathbf{y} = \underline{\mathbf{x}}\cdot\underline{\mathbf{y}} + \overline{\mathbf{x}}\cdot\overline{\mathbf{y}}.$$

Now, for any integer relation, $\mathbf{v}$, for $\boldsymbol{\alpha}$ we have $\mathbf{v}\cdot\mathbf{x} = \underline{\mathbf{v}}\cdot\underline{\mathbf{x}}$. Furthermore, $\mathbf{v}\cdot\mathbf{x} \in \mathbb{Z}$ if $\mathbf{x} \in \mathbb{Z}^n$. For some index j, $\mathbf{v}\cdot\mathbf{b}_j \neq 0$. Let s be the smallest index for which this is true. This implies that $\underline{\mathbf{v}}\cdot\mathbf{q}_i = 0$ for all $1 \leqslant i < s$. We then have

$$\begin{aligned} 1 \leqslant |\mathbf{v}\cdot\mathbf{b}_s| &= |\underline{\mathbf{v}}\cdot\underline{\mathbf{b}}_s| \\ &= |r_{1,s}\underline{\mathbf{v}}\cdot\mathbf{q}_1 + r_{2,s}\underline{\mathbf{v}}\cdot\mathbf{q}_2 + \cdots + r_{s,s}\underline{\mathbf{v}}\cdot\mathbf{q}_s| \\ &= r_{s,s}|\underline{\mathbf{v}}\cdot\mathbf{q}_s| \leqslant r_{s,s}\left\|\underline{\mathbf{v}}\right\|_2 = r_{s,s}\left\|\mathbf{v}\right\|_2. \end{aligned}$$

Thus, we conclude that

$$\|\mathbf{v}\|_2 \geqslant \min\left\{r_{i,i}^{-1} \;\middle|\; r_{i,i} \neq 0;\ i = 1, 2, \ldots, n\right\}. \tag{6.2}$$

We have thus set out two important principles for finding integer relations. By examining the diagonal elements of $\mathbf{R}$, we can either discover an integer relation, if the last element is non-zero, or place a lower bound on the size of any such relation, as measured by the Euclidean norm, using (6.2).

Let us now set out a version of the HJLS algorithm.

ALGORITHM 6.2.

1 **begin**
2 $\quad \mathbf{B} := \mathbf{I}$;
3 $\quad$ *QRdecompose*$(\underline{\mathbf{B}}, \mathbf{Q}, \mathbf{R})$;
4 $\quad$ **while** $r_{n,n} = 0 \ \wedge \ \max\{r_{1,1}, r_{2,2}, \ldots, r_{n,n}\} > \epsilon$ **do**
5 $\quad\quad j := \arg\max\left\{2^1 r_{1,1}^2, 2^2 r_{2,2}^2, \ldots, 2^{n-1} r_{n-1,n-1}^2\right\}$;
6 $\quad\quad k := \left\lfloor \dfrac{r_{j,j+1}}{r_{j,j}} \right\rceil$;
7 $\quad\quad \mathbf{b}_{j+1} := \mathbf{b}_{j+1} - k\mathbf{b}_j$;
8 $\quad\quad$ *swap*$(\mathbf{b}_j, \mathbf{b}_{j+1})$;
9 $\quad\quad$ *QRdecompose*$(\underline{\mathbf{B}}, \mathbf{Q}, \mathbf{R})$;
10 $\quad$ **od**;
11 **end**.

The similarity of this algorithm to the version of the LLL algorithm we presented in Algorithm 7.1 of Chapter 3 is apparent at once. The clear identification of the similarities between these two algorithms was one of the stated aims of HASTAD *et al.*. Furthermore, it differs from the PSLQ algorithm and from Bergman's algorithm only in the details.

An important difference from the LLL algorithm is the choice of index at which to perform an exchange of basis vectors. The rule (on line 5) is due to BERGMAN (1980). We will refer to it as BERGMAN'S EXCHANGE RULE.

Just as for Algorithm 7.1 of Chapter 3, we observe that it is unnecessary and inefficient to perform a full **QR** decomposition at line 9. A Givens rotation can be substituted.

We now state two propositions concerning the properties of the algorithm which are adapted directly from the original paper. The first proposition concerns the properties which hold at termination.

PROPOSITION 6.1. *If Algorithm 6.2 terminates and* $0 < \epsilon < 1$ *then either* (6.2) *holds for any integer relation* $\mathbf{v}$ *for* $\boldsymbol{\alpha}$ *or* $\mathbf{c}_n$ *is an integer relation for* $\mathbf{v}$*, where* $\mathbf{C}^T = \mathbf{B}^{-1}$ *at termination. If the latter then*

$$\|\mathbf{c}_n\|_2^2 \leqslant 2^{n-2} \min\left\{\|\mathbf{v}\|_2^2, \epsilon^{-2}\right\}. \tag{6.3}$$

PROOF. The first part of the proposition has been established in the discussion prior to the statement of the algorithm. It remains to show (6.3). Consider the basis $\mathbf{B}$, and the associated matrices $\mathbf{Q}$ and $\mathbf{R}$, just prior to the final exchange step on line 8, if there was one, and let $\mathbf{B}'$, $\mathbf{Q}'$ and $\mathbf{R}'$ be their values after the exchange, that is, their terminal values. With this notation, let $\mathbf{C}^T = \mathbf{B}'^{-1}$ and $\mathbf{c}_n$ is an integer relation for $\boldsymbol{\alpha}$. If there was no final exchange step then the algorithm must have terminated before the first iteration which implies that $\mathbf{C} = \mathbf{I}$ and so (6.3)

holds trivially. So let us now assume there was a final exchange step. This must have involved an exchange of $\mathbf{b}_{n-1}$ with $\mathbf{b}_n$. That the index $n-1$ was chosen using Bergman's exchange rule implies that

$$r_{i,i}^2 \leqslant 2^{n-i-1} r_{n-1,n-1}^2$$

for all $1 \leqslant i \leqslant n-1$. Now, after the swap, $r'_{n-1,n-1} = 0$ and therefore $r'_{n,n} = r_{n-1,n-1}$. Hence,

$$r'_{n,n}{}^{-2} = r_{n-1,n-1}^{-2} \leqslant 2^{n-2} \min\left\{ r_{i,i}^{-2} \mid r_{i,i} \neq 0;\ i = 1, 2, \ldots, n \right\} \tag{6.4}$$

$$\leqslant 2^{n-2} \min\left\{ \|\mathbf{v}\|_2^2, \epsilon^{-2} \right\} \tag{6.5}$$

where $\mathbf{v}$ is an integer relation for $\boldsymbol{\alpha}$. Now, $\mathbf{c}_n$ is also an integer relation and $\mathbf{c}_n \cdot \mathbf{b}'_i = 0$ for $1 \leqslant i < n$ so $\underline{\mathbf{c}}_n \cdot \mathbf{q}'_i = 0$. However, $\mathbf{c}_n \cdot \mathbf{b}'_n = 1$ which implies that

$$1 = \mathbf{c}_n \cdot \mathbf{b}'_n = r'_{n,n} \underline{\mathbf{c}}_n \cdot \mathbf{q}'_n \leqslant r'_{n,n} \|\mathbf{c}_n\|_2$$

and so

$$\|\mathbf{c}_n\|_2 \leqslant r'_{n,n}{}^{-1} = r_{n-1,n-1}^{-1}. \tag{6.6}$$

Together, (6.5) and (6.6) imply (6.3). □

PROPOSITION 6.2. *If Algorithm 6.2 is executed with $\epsilon > 0$ then it terminates after $O(n^2(n + \log \epsilon^{-1}))$ iterations.*

PROOF. We prove this proposition in a similar fashion to the proof we used for the running time bounds for the LLL algorithm in Proposition 7.2 of Chapter 3. In that proof we showed that the square of the value of $D = d_1 d_2 \cdots d_n$ was diminished by at least a quarter after each exchange step, where $d_j = r_{1,1} r_{2,2} \cdots r_{j,j}$. We found that D was bounded above and below and so we were able to obtain the desired running time bound.

For this proof, consider again the value of $D = d_1 d_2 \cdots d_n$, but let us instead define the d_j as

$$d_j = m_1 m_2 \cdots m_j$$

where

$$m_i = \max\left\{ r_{i,i}, 2^{-n/2} \epsilon \right\}.$$

At each exchange step, the application of Bergman's exchange rule implies that the index j is selected so that

$$2^j r_{j,j}^2 \geqslant 2^i r_{i,i}^2$$

for $i = 1, 2, \ldots, n$. Since the algorithm did not terminate on a previous iteration, there exists some index, s, such that $r_{s,s} > \epsilon$. Hence, we find that

$$2^n r_{j,j}^2 \geqslant 2^j r_{j,j}^2 \geqslant 2^s r_{s,s}^2 > 2\epsilon^2$$

and so

$$r_{j,j}^2 > 2^{-n+1}\epsilon^2. \tag{6.7}$$

Furthermore, Bergman's exchange rule implies that $r_{j,j}^2 \geqslant 2r_{j+1,j+1}^2$ which means that we can use Proposition 7.1 of Chapter 3 to establish that

$$r'^2_{j,j} \leqslant \frac{3}{4}r_{j,j}^2, \qquad r'^2_{j+1,j+1} \leqslant r_{j,j}^2 \qquad \text{and} \qquad r'_{j,j}r'_{j+1,j+1} \leqslant r_{j,j}r_{j+1,j+1}. \tag{6.8}$$

We will show that the same inequalities are satisfied for m_j, m_{j+1}, m'_j and m'_{j+1}. Of course, $r'_{i,i} = r_{i,i}$ and hence $m'_i = m_i$ for $i = 1, 2, \ldots, n$ when $i \neq j$, $i \neq j+1$. As an immediate consequence,

$$d'_i = d_i \tag{6.9}$$

for all $i = 1, 2, \ldots, j-1$.

Now, (6.7) implies that $m_j^2 > 2^{-n+1}\epsilon^2$ and the leftmost inequality of (6.8) implies that $m'^2_j \leqslant \frac{3}{4}m_j^2$. Thus,

$$d'^2_j \leqslant \frac{3}{4}d_j^2. \tag{6.10}$$

The middle inequality of (6.8) clearly implies that ${m'_{j+1}}^2 \leqslant m_j^2$.

We will now show that $m'_j m'_{j+1} \leqslant m_j m_{j+1}$. If $r'_{j,j} \geqslant 2^{-n/2}\epsilon$ then

$$m'_j m'_{j+1} = r'_{j,j}m'_{j+1} \leqslant r_{j,j}m_{j+1} = m_j m_{j+1}.$$

Otherwise,

$$m'_j m'_{j+1} \leqslant m_{j+1}m'_{j+1} \leqslant m_{j+1}m_j.$$

It follows that

$$d'_i \leqslant d'_i \tag{6.11}$$

for all $i = j+1, j+2, \ldots, n$.

Together, (6.9), (6.10) and (6.11) imply that

$$D' \leqslant \frac{\sqrt{3}}{2}D,$$

just as we were able to show for the LLL algorithm. The algorithm begins with the initial basis matrix set to the identity. This implies that, initially, $r_{i,i} \leqslant 1$ for $i = 1, 2, \ldots, n$ and so $D \leqslant 1$. If, at any iteration, $m_i = 2^{-n/2}\epsilon$ for all $i = 1, 2, \ldots, n$ then the algorithm will terminate and so

$$D \geqslant \left(2^{-n/2}\epsilon\right)^{n(n+1)/2}.$$

Thus, the total number of iterations, E, is bounded by

$$E \leqslant n(n+1)\left[\tfrac{1}{2}n\log_{2/\sqrt{3}}2 + \log_{2/\sqrt{3}}\epsilon^{-1}\right]$$

and so the number of iterations performed by the algorithm before termination is $O(n^2(n + \log\epsilon^{-1}))$. □

Each iteration of the HJLS algorithm requires vector addition (subtraction) and a Givens rotation. Thus, $O(n)$ arithmetic operations are required in each iteration, so the total number of arithmetic operations required by the algorithm is $O(n^3(n + \log \epsilon^{-1}))$.

We now briefly describe two of the variants proposed by HASTAD *et al.*. They proposed variants for finding multiple linearly independent integer relations and for finding simultaneous integer relations. To find multiple, say k, linearly independent integer relations, it is necessary to continue the iterations until $r_{n-k,n-k} = 0$ or until it can be shown that there is no set of linearly independent integer relations all of which have Euclidean norm less than ϵ^{-1}. To do this, we check whether $\max\{r_{1,1}, r_{2,2}, \ldots, r_{n-k+1,n-k+1}\} > \epsilon$.

Consider the problem of finding simultaneous integer relations, by which we mean that the matrix $\mathbf{P}^T$ in the radius function has row rank $t < n - 1$ and $\boldsymbol{\alpha}^T$ in the height function is no longer a vector but an $(n - t) \times n$ matrix. A simultaneous integer relation $\mathbf{v}$ for $\boldsymbol{\alpha}$ is found when $h(\mathbf{v}) = 0$. There will always be $n - t$ zeros on the main diagonal of $\mathbf{R}$. To make the algorithm suitable for this new purpose, we continue the iterations until $r_{t,t} = 0$. If this occurs then $\mathbf{c}_n$ is a simultaneous integer relation for $\boldsymbol{\alpha}$. If instead $\max\{r_{1,1}, r_{2,2}, \ldots, r_{n,n}\} > \epsilon$ then we can conclude that there are no simultaneous integer relations with Euclidean norm less that ϵ^{-1}.

It is interesting to compare the HJLS algorithm with the similar PSLQ algorithm and Bergman's algorithm. The PSLQ algorithm differs from the HJLS algorithm in two key respects: the Bergman's exchange rule is parameterised so that the index j so that

$$j = \arg\max\left\{\gamma r_{1,1}, \gamma^2 r_{2,2}, \ldots, \gamma^n r_{n,n}\right\}$$

where $\gamma > 2/\sqrt{3}$ and full (but slightly modified) Hermite reduction is performed after each exchange step. The modification to Hermite reduction referred to is that the matrix $\mathbf{R}$ which dictates the operations on the basis is obtained in this case from the $\mathbf{QR}$ decomposition of $\underline{\mathbf{B}}$ rather than of $\mathbf{B}$. It is also worth mentioning that, in FERGUSON *et al.* (1996), the authors extended the PSLQ algorithm to approximation problems in complex and quaternion vector spaces. Bergman's algorithm differs from the HJLS algorithm in that it too performs modified Hermite reduction, and its termination criterion appears to be slightly different. In Bergman's algorithm, the emphasis is upon finding a basis consisting of sufficiently good approximations to the line $\mathbb{R}\boldsymbol{\alpha}$ or finding a single integer relation for $\boldsymbol{\alpha}$. Therefore, we could say that Bergman's algorithm terminates if $r_{n,n} > 0$ or $\max\{\rho(\mathbf{b}_1), \rho(\mathbf{b}_2), \ldots, \rho(\mathbf{b}_n)\} \leqslant \epsilon$.

As we remarked for the LLL algorithm, a numerically stable version of the HJLS algorithm requires full (modified) Hermite reduction at each exchange step in order to keep the size of the elements of $\mathbf{R}$ small. The penalty incurred is that the arithmetic complexity is further increased by a factor of n. This may explain the

superior numerical properties of the PSLQ algorithm which have been demonstrated by FERGUSON *et al.*.

Now, let us consider the properties of the HJLS algorithm with respect to approximation of the line $\mathbb{R}\boldsymbol{\alpha}$. If we allow the algorithm to run with some arbitrarily small ϵ then the algorithm either finds an integer relation or halts with $\max\{r_{1,1}, r_{2,2}, \ldots, r_{n,n}\} \leqslant \epsilon$. Suppose the latter. If we Hermite-reduce the terminal basis then, for each basis vector $\mathbf{b}_i$, $i = 1, 2, \ldots, n$, we have

$$\begin{aligned}\rho(\mathbf{b}_i) &= \|\underline{\mathbf{b}}_i\|_2 \\ &= \left(r_{1,j}^2 + r_{2,j}^2 + \cdots + r_{j,j}^2\right)^{1/2} \\ &\leqslant \left(\tfrac{1}{4}r_{1,1}^2 + \tfrac{1}{4}r_{2,2}^2 + \cdots + \tfrac{1}{4}r_{j-1,j-1}^2 + r_{j,j}^2\right)^{1/2} \\ &\leqslant \tfrac{1}{2}\epsilon\sqrt{n+3}.\end{aligned}$$

Therefore, if we Hermite-reduce the terminal basis of the HJLS algorithm then this new algorithm either detects an integer relation or produces a basis of arbitrarily good approximations of the line $\mathbb{R}\boldsymbol{\alpha}$. Therefore, the algorithm is strongly convergent.

How good are these approximations compared to the best approximations? The answer to this question is unknown but more recent algorithms of JUST (1992) and RÖSSNER & SCHNORR (1996) attempt to ensure that the approximations produced are not only small in radius but also in height. In order to show that the approximations are good with respect to this criterion, they appeal to the following theorem of DIRICHLET, which is the generalisation of Theorem 2.1 of Chapter 2 to arbitrary dimensions, and its consequences.

THEOREM 6.1. *Given any N real numbers $\alpha_1, \alpha_2, \ldots, \alpha_N$ and an integer $Q > 1$, there exist $N+1$ integers $p_1, p_2, \ldots, p_N, q$ such that*

$$0 < q < Q^N \qquad \textit{and} \qquad |q\alpha_i - p_i| \leqslant \frac{1}{Q} \tag{6.12}$$

for all $i = 1, 2, \ldots, N$.

PROOF. The proof makes use of the pigeon-hole principle and is a straightforward extension of the proof of Theorem 2.1 of Chapter 2. Consider the set of $Q^N + 1$ integer $(N+1)$-tuples which consists of $(0, 0, \ldots, 0)$, $(-1, 0, \ldots, 0)$ as well as $(\lfloor q\alpha_1 \rfloor, \lfloor q\alpha_2 \rfloor, \ldots, \lfloor q\alpha_N \rfloor, q)$ for $q = 1, 2, \ldots, Q^N - 1$. For any element of the set $(p_1, p_2, \ldots, p_N, q)$, it is clear that $(q\alpha_1 - p_1, q\alpha_2 - p_2, \ldots, q\alpha_N - p_N)$ lies within the unit hypercube $[0, 1]^N$. We then divide this hypercube into Q^N smaller hypercubes in the obvious way, the sides of each have length $1/Q$. Since we have $Q^N + 1$ elements in our set of integers $(N+1)$-tuples, there must be a pair of $(N+1)$-tuples $(p_1, p_2, \ldots, p_N, q)$ and $(p'_1, p'_2, \ldots, p'_N, q')$ such that

$$(q\alpha_1 - p_1, q\alpha_2 - p_2, \ldots, q\alpha_N - p_N) \quad \text{and} \quad (q'\alpha_1 - p'_1, q'\alpha_2 - p'_2, \ldots, q'\alpha_N - p'_N)$$

with q and q' not both zero that belong to the same small hypercube. Therefore,

$$|(q' - q)\alpha - (p'_i - p_i)| \leqslant \frac{1}{Q}$$

for all $i = 1, 2, \ldots, N$ and $0 < |q' - q| < Q^N$. □

COROLLARY 6.1. *Given N real numbers $\alpha_1, \alpha_2, \ldots, \alpha_N$ there is at least one solution in integers $p_1, p_2, \ldots, p_N, q$, $q \neq 0$, to*

$$|q\alpha_i - p_i| < q^{-1/N} \tag{6.13}$$

for all $i = 1, 2, \ldots, N$. If any of the α_i are irrational then there is an infinity of integer solutions.

PROOF. Theorem 6.1 implies that there must exist one solution, since for any $Q > 0$ we find a solution with

$$|q\alpha_i - p_i| \leqslant Q^{-1} < q^{-1/N}.$$

If, for some $1 \leqslant j \leqslant N$, $\alpha_j \notin \mathbb{Q}$ then for any solution to (6.13) we have $|q\alpha_j - p| = \epsilon > 0$ for all integers p, q, $q \neq 0$. By setting $Q > 1/\epsilon$ we can deduce the existence of a different solution. Thus, there must be an infinite sequence of solutions in this case. □

The following theorem is more complex to prove, so its proof is omitted (see CASSELS, 1957, Theorem III, p. 79).

THEOREM 6.2. *Corollary 6.1 is not true if the exponent $-1/N$ in* (6.13) *is replaced by any smaller constant.*

Finding solutions to (6.12) is a simultaneous Diophantine approximation problem of finding points in $\mathbb{Z}^{N+1}$ which lie close to the line $\mathbb{R}\boldsymbol{\alpha}$, as measured by the sup-norm, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N, 1)$. Corollary 6.1 implies that there exist approximations which lie sufficiently "close" to the line in the sense implied by (6.13). Theorem 6.2 implies that no better exponent can be substituted. For this reason, we refer to (6.13) as the DIRICHLET BOUND.

RÖSSNER & SCHNORR (1996), improving the analysis of JUST (1992), have announced that their algorithm produces simultaneous Diophantine approximations of this type which satisfy

$$|q\alpha_i - p_i| \leqslant \frac{2^{(N+3)/4}\sqrt{1 + \alpha_i^2}}{q^{1/N}}$$

for all $i = 1, 2, \ldots, N$. That is, they are, within a constant factor, as good as can be expected (although not necessarily best approximations). Their algorithm, which is essentially due to JUST, is again similar to the HJLS algorithm. The chief departures from the HJLS algorithm are the use of full Hermite reduction after each exchange step and the abandonment of Bergman's exchange rule. Instead, the exchange rule

of the LLL algorithm is applied with the restriction that exchanges of the last two vectors in the basis may only be performed when there is no other choice.

Finally in this subsection, we note that simultaneous Diophantine approximation and finding integer relations were two of the original applications of the LLL algorithm envisaged by Lenstra *et al.* (1982). They propose a method by which a lattice is constructed for a particular $\boldsymbol{\alpha}$ and ϵ and then Lovász-reduced. The simultaneous Diophantine approximation obtained in this way fulfills the Dirichlet bound, up to a constant factor. It is therefore unclear which method is to be preferred. However, the algorithms we have presented here appear to be a little more elegant in that, if a smaller value of ϵ is subsequently required, these algorithms need only undergo a few more iterations. The approach of Lenstra *et al.* requires a new lattice and hence a new **QR** decomposition. Furthermore, these algorithms (particularly the PSLQ algorithm) have, or can be easily modified to have, good numerical stability.

6.4. Numerical Examples. In this subsection, we revisit Example 4.2 to compare the performance of our accelerated algorithm (Algorithm 5.1) with Brun's algorithm (Algorithm 6.1) and the HJLS algorithm (Algorithm 6.2).

Example 6.1. We consider the problem of finding good or best approximate integer relations to $(e^2, e, 1)$. Here we compare the bases produced by our accelerated algorithm with those produced by Brun's algorithm and the HJLS algorithm. In Example 4.2, we used the additive algorithm to find all the best approximate integer relations to $(e^2, e, 1)$ through the simultaneous Diophantine approximation system consisting of the lattice $\mathbb{Z}^3$ and radius and height functions defined by

$$\rho(\mathbf{x}) = |\boldsymbol{\alpha} \cdot \mathbf{x}| \qquad \text{and} \qquad h(\mathbf{x}) = \left\|\mathbf{P}^T\mathbf{x}\right\|_2$$

where $\boldsymbol{\alpha} = (e^2, e, 1)$ and the columns $\mathbf{p}_1$ and $\mathbf{p}_2$ of $\mathbf{P}$ are an orthonormal basis of the orthogonal complement of $\boldsymbol{\alpha}$ in $\mathbb{R}^3$. Since the radius and height are strictly convex and complementary, we are assured that, for every best approximate integer relation, Algorithm 5.1 will find an equivalent lattice point.

Table 8 lists the state of the basis on each iteration in a format which is now familiar. As witnessed in the table, by the 14^{th} iteration, the accelerated algorithm has found the six best approximate integer relations for $\boldsymbol{\alpha}$ with radius greater than $\epsilon = 0.01$ as well as $(-6, 13, 9)$.

To compare Brun's algorithm, we have modified Algorithm 6.1 to output the updates to the inverse basis matrix, $\mathbf{C}^T = \mathbf{B}^{-1}$. We initially set $\mathbf{C} := \mathbf{I}$ and we augment the line $\mathbf{b}_t := \mathbf{b}_t + r\mathbf{b}_s$ (line 11) with $\mathbf{c}_s := \mathbf{c}_s - r\mathbf{c}_t$. It can be checked that the condition $\mathbf{C}^T\mathbf{B} = \mathbf{I}$ is thus always maintained. Table 9 lists the state of the basis at the beginning of each iteration of this modification to Brun's algorithm. We have chosen to terminate the algorithm at the point at which the best approximate integer relation $(-6, 13, 9)$ is discovered by the algorithm. To this point, the algorithm has

TABLE 8. Outputs and important variables of Algorithm 5.1 applied to $\boldsymbol{\alpha} = (e^2, e, 1)$.

It.	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(0,0,1)$	$(0,1,0)$	$(1,0,0)$		
2	$(0,0,1)$	$(\mathbf{0},\mathbf{1},\mathbf{-1})$	$(1,0,0)$	1.72	1.40
3	$(\mathbf{0},\mathbf{1},\mathbf{-2})$	$(0,0,1)$	$(1,0,0)$	0.718	2.23
4	$(0,1,-2)$	$(\mathbf{1},\mathbf{-2},\mathbf{-1})$	$(0,0,1)$	0.952	2.45
5	$(\mathbf{-1},\mathbf{2},\mathbf{2})$	$(0,1,-2)$	$(1,-2,-1)$	4.72×10^{-2}	3.00
6	$(-1,2,2)$	$(\mathbf{0},\mathbf{-1},\mathbf{3})$	$(0,1,-2)$	0.282	3.16
7	$(-1,2,2)$	$(\mathbf{1},\mathbf{-3},\mathbf{1})$	$(0,1,-2)$	0.234	3.32
8	$(-1,2,2)$	$(\mathbf{2},\mathbf{-5},\mathbf{-1})$	$(0,1,-2)$	0.187	5.48
9	$(-1,2,2)$	$(\mathbf{3},\mathbf{-7},\mathbf{-3})$	$(0,1,-2)$	0.139	8.19
10	$(-1,2,2)$	$(\mathbf{1},\mathbf{1},\mathbf{-10})$	$(3,-7,-3)$	0.107	10.1
11	$(\mathbf{2},\mathbf{-8},\mathbf{7})$	$(-1,2,2)$	$(1,1,-10)$	3.19×10^{-2}	10.8
12	$(\mathbf{-3},\mathbf{10},\mathbf{-5})$	$(2,-8,7)$	$(1,1,-10)$	1.56×10^{-2}	11.6
13	$(\mathbf{3},\mathbf{-3},\mathbf{-14})$	$(-3,10,-5)$	$(2,-8,7)$	1.23×10^{-2}	14.6
14	$(\mathbf{-6},\mathbf{13},\mathbf{9})$	$(3,-3,14)$	$(2,-8,7)$	3.33×10^{-3}	16.9

TABLE 9. Outputs and important variables of Brun's algorithm applied to $\boldsymbol{\alpha} = (e^2, e, 1)$.

It.	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$		
2	$(\mathbf{1},\mathbf{-2},\mathbf{0})$	$(0,1,0)$	$(0,0,1)$	1.95	2.22
3	$(1,-2,0)$	$(\mathbf{-1},\mathbf{3},\mathbf{0})$	$(0,0,1)$	0.766	3.16
4	$(\mathbf{1},\mathbf{-2},\mathbf{-1})$	$(-1,3,0)$	$(0,0,1)$	0.952	2.45
5	$(1,-2,-1)$	$(-1,3,0)$	$(\mathbf{-1},\mathbf{2},\mathbf{2})$	4.75×10^{-2}	3.00
6	$(\mathbf{2},\mathbf{-5},\mathbf{-1})$	$(-1,3,0)$	$(-1,2,2)$	0.187	5.48
7	$(2,-5,1)$	$(\mathbf{-9},\mathbf{23},\mathbf{4})$	$(-1,2,2)$	1.90×10^{-2}	25.0
8	$(\mathbf{5},\mathbf{-11},\mathbf{-7})$	$(-9,23,4)$	$(-1,2,2)$	4.42×10^{-2}	14.0
9	$(5,-11,7)$	$(-9,23,4)$	$(\mathbf{-6},\mathbf{13},\mathbf{9})$	3.33×10^{-3}	16.9

discovered only three of the seven best approximations with radius greater than or equal to that of $(-6, 13, 9)$.

Finally, we demonstrate the operation of the HJLS algorithm. Table 10 lists the state of the (inverse) basis $\mathbf{C}$ at the beginning of each iteration. The algorithm was executed with $\epsilon = 0.01$ and terminates after the 11^{th} iteration, concluding that there is no integer relation $\mathbf{x}$ for $\boldsymbol{\alpha}$ with $h(\mathbf{x}) \leqslant 100$. To this point, it has discovered four of the seven best approximations with radius greater than or equal to that of $(-6, 13, 9)$ as well as one, $(8, -28, 17)$, with smaller radius. Notice that two of

TABLE 10. Outputs and important variables of the HJLS algorithm applied to $\boldsymbol{\alpha} = (e^2, e, 1)$.

It.	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$	$\rho(\mathbf{v})$	$h(\mathbf{v})$
1	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$		
2	$(1,0,0)$	$(0,0,1)$	$(\mathbf{0},\mathbf{-1},\mathbf{3})$	0.282	3.16
3	$(0,0,1)$	$(\mathbf{-1},\mathbf{0},\mathbf{8})$	$(0,-1,3)$	0.611	8.06
4	$(0,0,1)$	$(0,-1,3)$	$(\mathbf{-1},\mathbf{2},\mathbf{2})$	4.75×10^{-2}	3.00
5	$(0,-1,3)$	$(\mathbf{0},\mathbf{-4},\mathbf{11})$	$(-1,2,2)$	0.127	11.7
6	$(0,-1,3)$	$(-1,2,2)$	$(\mathbf{-3},\mathbf{10},\mathbf{-5})$	1.56×10^{-2}	11.6
7	$(-1,2,2)$	$(\mathbf{-6},\mathbf{13},\mathbf{9})$	$(-3,10,-5)$	3.33×10^{-3}	16.9
8	$(-1,2,2)$	$(-3,10,-5)$	$(-6,13,9)$		
9	$(-3,10,-5)$	$(\mathbf{8},\mathbf{-28},\mathbf{17})$	$(-6,13,9)$	5.58×10^{-4}	33.7
10	$(-3,10,-5)$	$(-6,13,9)$	$(8,-28,17)$		
11	$(-6,13,9)$	$(\mathbf{-27},\mathbf{55},\mathbf{50})$	$(8,-28,17)$	9.86×10^{-4}	79.1

the iterations only involve an exchange: the partial Hermite reduction involves no operations on the basis (that is, $k := 0$ at line 6).

CHAPTER 5

PROBABILITY OF INTERCEPT

1. Introduction

Intercept time problems are those in which one wishes to obtain information about the simultaneous coincidence of two or more periodic events. They are interesting mathematical problems and common to many physical systems, but they are particularly relevant to the design of equipment for electronic support measures (ESM), such as radar warning receivers. In designing a radar warning receiver, it often happens that we can only observe a given part of the environment periodically for a short time. For example, this will be the case if we use a rotating, directional antenna or we use a swept-frequency superheterodyne receiver. In addition, the radar we wish to observe might only be transmitting periodically for a short time. A good radar warning receiver should observe a radar very soon after it first begins transmitting, so in designing our radar warning receiver we would like to ensure that the intercept time is low or the probability of intercept after a specified time is high.

We can formulate these problems as problems of determining the time at which several periodic pulse trains coincide. For instance, in the case of rotating, directional antennas, we can associate a function to each antenna that is equal to 1 or **true** whenever that antenna is pointing at the other antenna (to within some tolerance as specified, perhaps, by the main beam width) and 0 or **false** at other times. Both functions are PERIODIC PULSE TRAINS. Both have a fixed period, which we call the PULSE REPETITION INTERVAL or PRI corresponding to the time required for one revolution of the antenna and a pulse width corresponding to the tolerance in angle. The antennas are "looking at each other" only when both functions are 1 (or **true**) simultaneously.

In certain situations, we may have to consider more than two periodic processes of this type. For example, the transmitting antenna may be emitting a periodic train of radar pulses at a particular carrier frequency as it rotates and the receiving antenna may be searching (scanning) periodically through a range of carrier frequencies. Thus, the receiver will receive energy from the transmitter only when the four pulse trains associated with the problem coincide.

The analysis of interception of pulse trains has been investigated sporadically over the last fifty years. RICHARDS (1948) was the first to publish a detailed analysis of the probability of intercept of two strictly periodic pulse trains. His construction of the problem is essentially similar to ours. He discovered a good approximation

for the probability and attempted to account for possible uncertainties in the parameters of the pulse trains. He demonstrated the relationship between the ratio of PRIs and the Farey series, as we also will. Miller & Schwarz (1953) and subsequently Friedman (1954) and Hawkes (1983) showed how intercept time could be predicted for rational PRI ratios using linear congruence. Using a statistical description of pulse trains due to Stein & Johansen (1958), Self & Smith (1985) derived an expression for the probability of intercept when the pulse widths and time differences between pulses are random variables. They claimed that the expression can be used as an approximation in the case where these parameters are fixed and known. Their results seem to have gained acceptance amongst practitioners in the ESM community because of their simplicity, their applicability to cases involving more than two pulse trains, their accuracy in some situations and because of the orientation of their paper towards ESM problems. However, their assumption that the probabilities of intercept in small, disjoint intervals are independent is invalid in the cases considered in this chapter. Most recently, Kelly *et al.* (1996) derived an exact expression for the probability of intercept where one phase is known. One of the objectives of this chapter is to present their results in the language of Diophantine approximation.

Suppose we have n pulse trains. Throughout this chapter, we assume that the time-of-arrival (TOA) of the i^{th} pulse from the k^{th} pulse train occurs at the time $iT_k + \phi_k$ where T_k is the PRI and ϕ_k is the phase. We will sometimes refer to the integer i as the pulse index. The pulses from each pulse train have associated with them a pulse width τ_k. We define the i^{th} pulse from the k^{th} pulse train to be "on" at time t when

$$iT_k + \phi_k - \tfrac{1}{2}\tau_k \leqslant t \leqslant iT_k + \phi_k + \tfrac{1}{2}\tau_k.$$

A coincidence or intercept occurs when all n pulse trains are simultaneously on. This is illustrated for the case of three pulse trains in Figure 1.

Let us now briefly summarise the contents of this chapter. We will consider a number of intercept time problems. Firstly, we will do this for two pulse trains only. We will consider the following variations: where the phases are known and equal, where they are known and unequal and where one or both are random variables. We will show that the problem can be simply stated and solved using the theory and algorithms for Diophantine approximation that we developed in Chapter 2. We will then consider intercept time problems involving more than two pulse trains. We will see that this is a simultaneous Diophantine approximation problem. Therefore, we can apply the theory and algorithms of Chapter 4. We find that many of the calculations which can be done easily for two pulse trains are difficult problems for arbitrary numbers of pulse trains. To conclude, we will contrast our approach to others which have appeared in the literature.

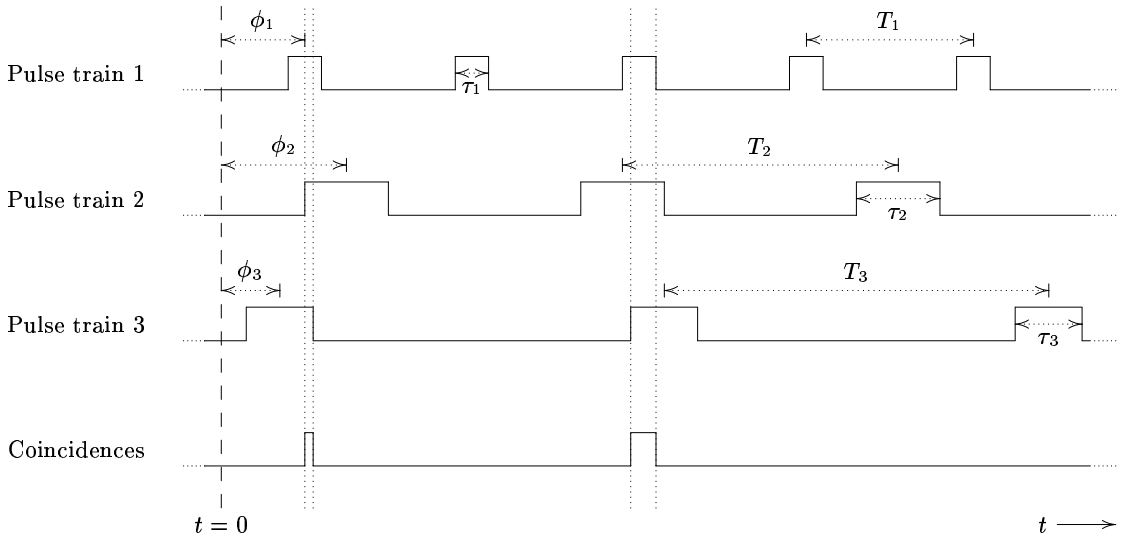


FIGURE 1. Coincidence (intercept) of three pulse trains.

Let us now consider these topics in a little more detail. In Sections 2–5, we consider problems involving two pulse trains only. For the problem of calculating intercept time of two pulse trains, we assume that the phases of the pulse trains are known *a priori*. We then want to find an algorithm for computing when the first intercept will occur and when subsequent intercepts will occur. For the probability of intercept, we assume that one or both phases are random and we want to find the probability that at least one intercept has occurred after a certain number of pulses or after a certain time.

We will firstly revisit the intercept time problem in Section 2. Unlike MILLER & SCHWARZ (1953) and similar work which exploits the properties of linear congruence, we will not restrict the ratio of PRIs to being rational numbers. We formulate the problem as a Diophantine approximation problem. We find that, by considering the simple continued fraction expansion of the PRI ratio and examining the convergents of that expansion, we can compute the intercept time. We present a means for finding the times of further intercepts with a recurrence equation. We believe that these techniques offer insights into the problem which have not previously come to light and they provide efficient methods for computation.

In Section 3, we will examine the probability of intercept between two periodic pulse trains. We show how the intercept probability expression of KELLY *et al.* where one phase is known can be reinterpreted and simplified by considering the number theoretic results obtained for the intercept time. We then consider the problem of RICHARDS, where neither phase is known and derive an exact expression for the probability of intercept in this case. As the exact expression is rather complex, we show that the expression for the earlier case, where one phase is known, can be adapted and used as a good approximation. In Section 4, we derive expressions for the mean time to intercept.

In Section 5, we examine the dependence of the probability of intercept of two pulse trains on the PRI parameters. We explore the relationship between the probability of intercept and the Farey series and outline how a recursive algorithm can be constructed to exactly calculate average probabilities of intercept.

We will then discuss the problem of interception of three or more pulse trains in Section 6. Although we can obtain satisfactory answers for some problems involving three pulse trains (using the theory and algorithms developed in Chapter 4), we will see that the properties which we relied upon for cases involving two pulse trains quickly evaporate as we increase the number of pulse trains.

Finally, in Section 7, we present a short critique of the approaches arising from linear congruence and the from the stochastic representation of pulse trains.

2. Intercept Time of Two Pulse Trains

In this section, we will discuss the problem of the intercept time of two pulse trains. Initially, we will consider the pulse trains divorced from their pulse widths, as if they were a sequence of points or impulses. We will solve the equivalent problem of APPROXIMATE COINCIDENCE. As we discussed earlier, the TOAs of the pulse trains are defined as $iT_1 + \phi_1$ for the first pulse train and $jT_2 + \phi_2$ for the second, where i and j are integers. Approximate coincidence occurs to within a tolerance δ when

$$|iT_1 + \phi_1 - jT_2 - \phi_2| \leqslant \delta. \tag{2.1}$$

Hence, we have formulated the problem as a problem in Diophantine approximation.

In our original problem, that of finding the first intercept time, the pulse trains have pulse widths, τ_1 and τ_2, associated with them. We assume here that the TOA of a pulse is defined as occurring in the middle of the pulse. The problem of finding intercepts reduces to the problem of approximate coincidence stated above where $\delta = \frac{1}{2}(\tau_1 + \tau_2)$. A closely related approximate coincidence problem is that where, in order to register an intercept, the intercept must last at least a length of time d. If an intercept occurs between the two pulse trains for pulse indices i and j then the length of the intercept is $\min\left\{\frac{1}{2}(\tau_1 + \tau_2) - |\zeta|, \tau_1, \tau_2\right\}$ where $\zeta = iT_1 + \phi_1 - jT_2 - \phi_2$. Therefore an intercept of length d or greater occurs if and only there is an approximate coincidence of the pulse trains within a tolerance $\delta = \frac{1}{2}(\tau_1 + \tau_2) - d$.

We will begin by discussing the case where $\phi_1 = \phi_2$. The inequality of (2.1) reduces to

$$|iT_1 - jT_2| \leqslant \epsilon. \tag{2.2}$$

We will refer to this condition as the IN PHASE problem. This is a homogeneous Diophantine approximation problem. We have already studied the theory of Diophantine approximation in Chapter 2. We will show how the simple continued

fraction expansion of the ratio of the PRIs and Euclid's algorithm can be directly used to find the first approximate coincidence.

We will then consider the case where $\phi_1 - \phi_2 \neq 0$. We will refer to this as the ARBITRARY PHASE problem. This is an inhomogeneous Diophantine approximation problem. We will show that the first approximate coincidence can be found using Cassels' algorithm.

To conclude this section, we will consider the problem of finding further approximate coincidences, having found one. We will show that a recurrence equation can be used to find all further coincidences.

2.1. In Phase Initial Conditions. If we write $\alpha = T_2/T_1$ then a solution to (2.2) is equivalent to finding a solution in integers p, q to

$$|q\alpha - p| \leqslant \epsilon$$

where $\epsilon = \delta/T_1$. Now, $(p, q) = (0, 0)$ is a trivial solution. If

$$\epsilon \geqslant \min\{1, \alpha\} \tag{2.3}$$

then either $(p, q) = (1, 0)$ or $(p, q) = (0, 1)$ is a solution. Suppose that (2.3) does not hold. We seek the *first* approximate coincidence, by which we mean the non-trivial solution in positive integers to (2.2) with least pulse index with respect to one of the pulse trains. It does not matter to which pulse train the criterion of least pulse index is applied: the first approximate coincidence will be the same, as we shall see. However, for the moment, suppose we seek the approximate coincidence with least positive pulse index for the second pulse train. Then we seek a best homogeneous Diophantine approximation of α in the absolute sense. From Theorem 3.6 of Chapter 2 we know that the correspondence of best approximations in this sense and the convergents of the simple continued fraction expansion of α is nearly one-to-one. We can state the following theorem.

THEOREM 2.1. *The minimum $i > 0$ or $j > 0$ such that $|iT_1 - jT_2| \leqslant \delta$, $T_1, T_2 > 0$ when $0 < \delta < \min\{T_1, T_2\}$ is given by $i = p_{n(\epsilon)}, j = q_{n(\epsilon)}$ where*

$$n(\epsilon) = \min_{n \geqslant 0} \{n \mid |\eta_n| \leqslant \epsilon\} \tag{2.4}$$

and (p_n, q_n) is the n^{th} convergent of the s.c.f. expansion of $\alpha = T_2/T_1$, η_n is the corresponding approximation error and $\epsilon = \delta/T_1$.

PROOF. The proof is a consequence of Theorem 3.6 and subsequent Remarks. So long as α is not a half-integer, its s.c.f. expansion contains all the best approximations of α in the absolute sense. If α is a half-integer then the best approximation $(\lfloor\alpha\rfloor, 1)$ will appear as a convergent but $(\lceil\alpha\rceil, 1)$ will be missed. This is of no consequence since the two approximations have identical absolute approximation error and the former has the lesser pulse index for the first pulse train.

The approximation errors of the convergents are a strictly decreasing sequence, so it is appropriate to choose the minimum n such that $|\eta_n| \leqslant \epsilon$. In this way, we find the best approximation we require and we are assured that the pulse index for the second pulse train is minimal. We will now see that it is minimal in the first pulse index also. Suppose there exists some $0 < i < p_{n(\epsilon)}$ for which there can be found some $j > 0$ to produce an approximate coincidence. Then $j > q_{n(\epsilon)}$ because $\left(p_{n(\epsilon)}, q_{n(\epsilon)}\right)$ is the best approximation with smallest denominator with an absolute approximation error less than or equal to ϵ. But then $j\alpha - i \geqslant \eta_n + \alpha + 1 \geqslant \min\{\alpha, 1\} > \epsilon$ and this is a contradiction. □

We recall that formulation of the problem as a homogeneous Diophantine approximation problem allows the use of Euclid's algorithm to find the first approximate coincidence. Hence, we can find first approximate coincidences in $O(\log \epsilon^{-1})$ arithmetic operations.

2.2. Arbitrary Phase Initial Conditions. The problem can be generalised to include the situation where the difference in phases or RELATIVE PHASE $\phi_1 - \phi_2 = \phi$ is arbitrary. Again, we wish to find the first time of approximate coincidence. Hence, we want to find the first $i, j \geqslant 0$ such that $|iT_1 - jT_2 + \phi| \leqslant \delta$. Note that it is now possible, but no longer necessary, that the pulse trains could approximately coincide at $i = j = 0$. If there is an approximate coincidence with $i = 0$ or $j = 0$ then it can be easily detected by determining if

$$\left| -\left\lfloor \frac{\phi}{T_2} \right\rceil T_2 + \phi \right| \leqslant \delta \qquad \text{or} \qquad \left| \left\lfloor \frac{-\phi}{T_1} \right\rceil T_1 + \phi \right| \leqslant \delta,$$

respectively. In what follows, we assume there are no approximate coincidences when $i = 0$ or $j = 0$.

If, as with the in phase initial conditions, we set $\alpha = T_2/T_1$ and $\epsilon = \delta/T_1$ and now also set $\beta = \phi/T_1$ then finding an approximate coincidence involves finding two positive integers p and q which satisfy

$$|q\alpha - p - \beta| \leqslant \epsilon.$$

The problem is thus one of inhomogeneous Diophantine approximations. The first approximate coincidence is thus a best inhomogeneous Diophantine approximation.

From Theorem 5.1 of Chapter 2, we know that all the best approximations can be obtained from the outputs of Cassels' algorithm. We state the following theorem.

THEOREM 2.2. *Consider solutions to*

$$|iT_1 - jT_2 + \phi| \leqslant \delta$$

with $T_1, T_2, \delta > 0$ in integer i, j with $j > 0$. If a solution exists then the solution with the minimum positive value for j and associated value for i are given by

$$(i, j) = \left(P_{m(\epsilon)} - kp_{m(\epsilon)-1}, Q_{m(\epsilon)} - kq_{m(\epsilon)-1}\right) \tag{2.5}$$

where

$$m(\epsilon) = \min_{m \geqslant 0} \{m \mid |\zeta_m| \leqslant \epsilon\},$$

$$k = \min \left\{ b_{m(\epsilon)}, \left\lfloor \frac{\epsilon - |\zeta_{m(\epsilon)}|}{|\eta_{m(\epsilon)-1}|} \right\rfloor \right\}, \tag{2.6}$$

$\epsilon = \delta/T_1$ and the p_n, q_n are the convergents of the s.c.f. expansion of $\alpha = T_2/T_1$, the η_n are the associated homogeneous approximation errors and the P_n and Q_n are the auxiliary convergents with respect to $\beta = \phi/T_1$, the ζ_n are their inhomogeneous approximation errors and the b_n are the auxiliary partial quotients as output by Cassels' algorithm (Algorithm 5.1 of Chapter 2).

Before proving this theorem, we recall, from Corollary 5.1 of Chapter 2, the behaviour of Cassels' algorithm for inputs α and β. It either terminates with $\zeta_n = 0$ or $\eta_n = 0$ or it produces a non-terminating sequence of outputs with $\lim_{n\to\infty} \zeta_n = 0$. If it does not terminate or it terminates with $|\zeta_n| \leqslant \epsilon$ then $m(\epsilon)$ is defined and an approximate coincidence occurs. If it terminates with $|\zeta_n| > \epsilon$ then no approximate coincidence occurs. If the algorithm terminates in this condition then $\eta_n = 0$ which means that the PRIs satisfy some rational relation. Hence, we interpret this condition as that in which the pulse trains are SYNCHRONISED, but "out of step" with one another.

PROOF. The proof follows from Theorem 5.1 of Chapter 2. First of all, the appearance of $b_{m(\epsilon)}$ in (2.6) is allowable because we can be sure that the algorithm must be in state $(\mathcal{A}_n)$ on the $(m(\epsilon)-1)^{\text{th}}$ iteration. Otherwise, from statement (vi) of Remark 5.1 of Chapter 2, $\zeta_{m(\epsilon)} = \zeta_{m(\epsilon)-2}$. Now, since $0 \leqslant k \leqslant b_{m(\epsilon)}$, the expression for (i, j) in (2.5) is an intermediate auxiliary convergent of α with respect to β (or an auxiliary convergent if $k = b_{m(\epsilon)}$). It is easily checked that the absolute inhomogeneous approximation error of the integer pair so determined is less than or equal to ϵ. Consider the arrangement of auxiliary convergents and intermediate auxiliary convergents defined in the proof of Theorem 5.1 of Chapter 2. According to this arrangement, the intermediate auxiliary convergent (or auxiliary convergent) selected by (2.5) is the first in the sequence that has an absolute inhomogeneous approximation error less than or equal to ϵ. Therefore, we conclude that it is the inhomogeneous best approximation of α with respect to β with least "denominator" such that the absolute inhomogeneous approximation error is less than or equal to ϵ. □

We recall that Cassels' algorithm (a modification of Euclid's algorithm) allows us to find the first approximate coincidence in the sense implied by Theorem 2.2, or to disprove its existence, in $O(\log \epsilon^{-1})$ arithmetic operations.

2.3. Finding Further Intercepts. It may be of interest not only to find the first approximate coincidence, but to find subsequent ones also. We will present

a recurrence equation for finding all approximate coincidences after the first. The recurrence equation is valid regardless of whether the in phase or arbitrary phase initial conditions hold.

Henceforth, we will abuse notation slightly and use the shorthand $p(\epsilon)$, $q(\epsilon)$ and $\eta(\epsilon)$ respectively to denote the numerator, denominator and homogeneous approximation error of the convergent $\big(p_{n(\epsilon)}, q_{n(\epsilon)}\big)$ where $n(\epsilon)$ is defined as in (2.4) of Theorem 2.1.

Suppose $\eta(\epsilon) \neq 0$. We define an intermediate fraction $(p'(\epsilon), q'(\epsilon))$ where

$$\begin{aligned}(p'(\epsilon), q'(\epsilon)) &= \big(p_{n(\epsilon)+1} - kp_{n(\epsilon)}, q_{n(\epsilon)+1} - kq_{n(\epsilon)}\big) \\ &= \big(p_{n(\epsilon)-1} + \big(a_{n(\epsilon)+1} - k\big)p_{n(\epsilon)}, q_{n(\epsilon)-1} + \big(a_{n(\epsilon)+1} - k\big)q_{n(\epsilon)}\big)\end{aligned} \tag{2.7}$$

and

$$k = \left\lfloor \frac{\epsilon - \big|\eta_{n(\epsilon)+1}\big|}{\big|\eta_{n(\epsilon)}\big|} \right\rfloor. \tag{2.8}$$

Let $\eta'(\epsilon)$ denote the approximation error $|q'(\epsilon)\alpha - p'(\epsilon)|$. We can quickly confirm that $|\eta'(\epsilon)| \leqslant \epsilon$ and $|\eta'(\epsilon) - \eta(\epsilon)| > \epsilon$. Now, $k \geqslant 0$ since $\big|\eta_{n(\epsilon)+1}\big| < \big|\eta_{n(\epsilon)}\big| \leqslant \epsilon$ and $k < a_{n(\epsilon)+1}$ since if $k \geqslant a_{n(\epsilon)+1}$ then $|\eta'(\epsilon)| \geqslant \big|\eta_{n(\epsilon)-1}\big| > \epsilon$. Thus, $(p'(\epsilon), q'(\epsilon))$ is an intermediate fraction between the $n(\epsilon)^{\text{th}}$ and $(n(\epsilon)+1)^{\text{th}}$ convergents unless $k = 0$, in which case it is the $(n(\epsilon)+1)^{\text{th}}$ convergent. Furthermore, we see that $\eta'(\epsilon) \neq 0$ since this would imply that $\eta_{n(\epsilon)+1} = 0$ and $k = 0$, which is impossible since $\big|\eta_{n(\epsilon)}\big| \leqslant \epsilon$.

We can now state the following theorem regarding further intercepts.

THEOREM 2.3. *Suppose there exists a pair of pulse indices (P, Q) which define an approximate coincidence $|PT_1 - QT_2 + \phi| \leqslant \delta$ for some $T_1, T_2, \delta > 0$ with $\delta < \min\{T_1, T_2\}$. Let $\alpha = T_2/T_1$, $\beta = \phi/T_1$, $\epsilon = \delta/T_1$, $\zeta = Q\alpha - P - \beta$ and*

$$(R, S) = \begin{cases} (p(2\epsilon), q(2\epsilon)) & \text{if } |\zeta + \eta(2\epsilon)| \leqslant \epsilon, \\ (p'(2\epsilon), q'(2\epsilon)) & \text{if } |\zeta + \eta'(2\epsilon)| \leqslant \epsilon, \\ (p(2\epsilon) + p'(2\epsilon), q(2\epsilon) + q'(2\epsilon)) & \text{otherwise.} \end{cases} \tag{2.9}$$

Then $(P + R, Q + S)$ defines the pulse indices of the next approximate coincidence, by which it is meant that there are no approximate coincidences for which the second pulse index is greater than Q but less than $Q + S$.

We remark that if $\eta(2\epsilon) = 0$ then $p'(2\epsilon)$ and $q'(2\epsilon)$ are not defined. In this case, (2.9) reduces to $(R, S) = (p(2\epsilon), q(2\epsilon))$.

PROOF. Consider the pair of pulse indices $(P + p, Q + q)$ where p and q are non-negative integers. If the pair define an approximate coincidence then $q > 0$ since $\delta < \min\{T_1, T_2\}$. In order for them to define an approximate coincidence we must

have

$$|(Q+q)\alpha - (P+p) - \beta| \leqslant \epsilon$$

which implies that, with $\eta = q\alpha - p$,

$$|\eta| \leqslant |\zeta| + \epsilon \leqslant 2\epsilon.$$

Hence, if $|\zeta + \eta(2\epsilon)| \leqslant \epsilon$ then $(R, S) = (p(2\epsilon), q(2\epsilon))$ since this is the best approximation with least denominator such that its absolute approximation error is less than or equal to 2ϵ. Suppose this is not the case.

We know, from the definition of $p'(\epsilon)$, $q'(\epsilon)$ and $\eta'(\epsilon)$ at the beginning of this subsection, that $|\eta(2\epsilon) - \eta'(2\epsilon)| > 2\epsilon$. Furthermore, from statement (ii) of Proposition 3.1 of Chapter 2, we deduce that $\eta(2\epsilon)$ and $\eta'(2\epsilon)$ have opposite sign. Thus, in order for the pair $(P+p, Q+q)$ to form an approximate coincidence, we must have

$$(\eta - \eta(2\epsilon))(\eta - \eta'(2\epsilon) + \eta(2\epsilon)) < 0.$$

We can then apply Proposition 3.3 and Proposition 3.4 of Chapter 2 to show that the smallest positive value of q for which this is satisfied, and the corresponding value of p, is given by $(p, q) = (p'(2\epsilon), q'(2\epsilon))$. Hence, if $|\zeta + \eta'(2\epsilon)| \leqslant \epsilon$ then this is the integer pair which should be chosen for (R, S).

Suppose that both $|\zeta + \eta(2\epsilon)| > \epsilon$ and $|\zeta + \eta'(2\epsilon)| > \epsilon$. We know that $\eta(2\epsilon)$ and $\eta'(2\epsilon)$ have opposite sign so we now require that

$$(\eta - \eta(2\epsilon))(\eta - \eta'(2\epsilon)) < 0.$$

Again applying Proposition 3.3 and Proposition 3.4 from Chapter 2, we find that $(p, q) = (p(2\epsilon) + p'(2\epsilon), q(2\epsilon) + q'(2\epsilon))$ is the integer pair with least denominator which satisfies this inequality. Without loss of generality, suppose $\eta(2\epsilon) > 0$ and $\eta'(2\epsilon) < 0$. Then $\zeta + \eta(2\epsilon) > \epsilon$ which implies that $\zeta + \eta(2\epsilon) + \eta'(2\epsilon) > -\epsilon$ since $\eta'(2\epsilon) \geqslant -2\epsilon$. Similarly, we have $\zeta + \eta'(2\epsilon) < -\epsilon$ which implies that $\zeta + \eta'(2\epsilon) + \eta(2\epsilon) < \epsilon$. Thus, $|\zeta + \eta(2\epsilon) + \eta'(2\epsilon)| < \epsilon$. We conclude that if $(R, S) \neq (p(2\epsilon), q(2\epsilon))$ and $(R, S) \neq (p'(2\epsilon), q'(2\epsilon))$ then $(R, S) = (p(2\epsilon) + p'(2\epsilon), q(2\epsilon) + q'(2\epsilon))$. □

The expression (2.9) of Theorem 2.3 defines a recurrence equation by which all further approximate coincidences after the first can be found. For the case of in phase initial conditions, we may set $P = Q = 0$.

3. Probability of Intercept of Two Pulse Trains

We now discuss how the number theoretic solution used in the previous section can be applied to the solution of probability of intercept problems involving two pulse trains. In the probability of intercept problem, one or both of the phases are assumed to be uniform random variables with ranges equal to their respective PRIs.

Two subproblems are now analysed. The first subproblem is the case in which only the phase of the second pulse train, ϕ, is a random variable, and we want to

know the probability of intercept after N pulses from the first pulse train. That is, we want to know the probability of at least one coincidence occurring with one of the first N pulses from second pulse train. We shall sometimes refer to this as the DISCRETE TIME problem. The second subproblem is the case in which both phases are random, and we want to know the probability of intercept over the time interval $[0, t]$. We shall sometimes refer to this as the CONTINUOUS TIME problem and to the interval $[0, t]$ as the OBSERVATION INTERVAL. The method of solution of the first subproblem leads to the solution of the second.

3.1. Uniformly Random Phase for One Pulse Train. We wish to find the probability that at least one approximate coincidence to within δ has occurred with pulse train 1 after N pulses from pulse train 2. We assume that we know the phase of pulse train 2. We can set the time origin so that pulses from pulse train 2 occur at the times jT_2, where j is a non-negative integer, without loss of generality. The relative phase, Φ, is unknown and is assumed to be a random variable, uniformly distributed over the interval $(-T_1, 0]$. At this point, we should justify this assumption. In an ESM scenario involving a simple transmitter and receiver, what we want to calculate is some measure of confidence of intercepting a pulse train within a certain number of "pulses" or "looks" from our receiver, and these looks constitute the second pulse train. The time at which our receiving equipment is turned on (the first look; pulse index $j = 0$) is known to us and is not random. We define the point $t = 0$ to be at the centre of this first look, in accordance with our usual construction. We assume that the pulse train from the transmitter which we wish to intercept — the first pulse train — is present at this time, which is to say that at least one pulse from the second pulse train occurred at some time $t \leqslant 0$. We assume that we have no control over when the transmitter begins operating. If the distribution of the "turn-on" time for the transmitter relative to that of the receiver exists and is sufficiently smooth and broad then the distribution of the time-of-arrival of the pulse from the transmitter immediately preceding the first from the receiver will be approximately uniform. Thus, we assign the pulse index $i = 0$ to this pulse and assume that the relative phase is uniform, in order to arrive at an indicative probability of intercept. Otherwise, we can view the results we will describe not as a probability in the strict sense but simply as a proportion of relative phases in $(-T_1, 0]$ that would have led to an intercept after the prescribed number of pulses from the second pulse train.

Let us now consider how to determine the probability of intercept. We again normalise all the parameters with respect to T_1 and define $\alpha = T_2/T_1$ and $\epsilon = \delta/T_1$. Let β be an instance of the random variable B, where $B = \Phi/T_2$. Therefore, $B \sim U(-1, 0)$. An approximate coincidence with tolerance ϵ occurs with one of the first N pulses from pulse train 2 if there exists some $0 \leqslant q < N$ such that

$$|q\alpha - p - \beta| \leqslant \epsilon$$

for some integer p. Let $\mathcal{I}_{p,q}$ be the interval on $\mathbb{R}$ of length 2ϵ defined by

$$\mathcal{I}_{p,q} = \{x \in \mathbb{R} \mid |q\alpha - p - x| \leqslant \epsilon\}.$$

Thus, an intercept occurs if

$$\beta \in \bigcup_{\substack{p,q\in\mathbb{Z};\\ 0\leqslant q<N}} \mathcal{I}_{p,q}.$$

Let $\mathcal{C}_N(\beta)$ be the characteristic function of this union. That is, $\mathcal{C}_N(\beta) = 1$ if there exists some $p, q \in \mathbb{Z}$, $0 \leqslant q < N$ such that $\beta \in \mathcal{I}_{p,q}$ and $\mathcal{C}_N(\beta) = 0$ otherwise. Let $\mathcal{P}_N$ be the probability of intercept after N pulses from the second pulse train. Then

$$\mathcal{P}_N = \int_{-1}^{0} \mathcal{C}_N(\beta)\, d\beta. \tag{3.1}$$

That is, the probability of intercept is that proportion of the range of possible relative phases (from -1 to 0) which is covered by the intervals $\mathcal{I}_{p,q}$ with $0 \leqslant q < N$. Now, $\mathcal{C}_N(\beta)$ is periodic with period 1 so we could replace the interval of integration in (3.1) with any interval of length 1. Notice that $\mathcal{P}_N = 1$ for all $N > 0$ if $\epsilon \geqslant \frac{1}{2}$. Therefore, we will assume that $\epsilon < \frac{1}{2}$.

Let us now consider the increase in the probability of intercept as we increment N. That is, we consider $\mathcal{P}_{N+1} - \mathcal{P}_N$ for $N \geqslant 0$. We define $\mathcal{P}_0 = 0$. We write

$$\mathcal{P}_{N+1} - \mathcal{P}_N = \int_{N\alpha-\frac{1}{2}}^{N\alpha+\frac{1}{2}} \mathcal{C}_{N+1}(\beta) - \mathcal{C}_N(\beta)\, d\beta. \tag{3.2}$$

Now, $\mathcal{C}_{N+1}(\beta) - \mathcal{C}_N(\beta) = 1$ if and only if there exists an integer p such that $x \in \mathcal{I}_{p,N}$ but $x \notin \mathcal{I}_{p,q}$ for any other choice of p or $0 \leqslant q < N$. Otherwise, $\mathcal{C}_{N+1}(\beta) - \mathcal{C}_N(\beta) = 0$. Furthermore, the interval $\mathcal{I}_{0,N}$ is the only interval of the form $\mathcal{I}_{p,N}$, $p \in \mathbb{Z}$, which is contained within the limits of integration in (3.2). All others lie completely outside. Therefore, the value of $\mathcal{P}_{N+1} - \mathcal{P}_N$ is the length of that portion of the interval $\mathcal{I}_{0,N}$ which does not overlap other intervals of the form $\mathcal{I}_{p,q}$ with $p, q \in \mathbb{Z}$, $0 \leqslant q < N$.

An overlap of $\mathcal{I}_{0,N}$ with $\mathcal{I}_{i,j}$, $i \in \mathbb{Z}$, $0 \leqslant j < N$ occurs if $|(N-j)\alpha - i| \leqslant 2\epsilon$. Therefore, we know that no such overlap can occur if $0 \leqslant N < q(2\epsilon)$. In this case,

$$\mathcal{P}_{N+1} - \mathcal{P}_N = 2\epsilon. \tag{3.3}$$

If $\eta(2\epsilon) = 0$ then

$$\mathcal{I}_{0,N} = \mathcal{I}_{-p(2\epsilon),N-q(2\epsilon)}.$$

Thus, if $\eta(2\epsilon) = 0$ and $N \geqslant q(2\epsilon)$ then

$$\mathcal{P}_{N+1} - \mathcal{P}_N = 0. \tag{3.4}$$

Suppose $\eta(2\epsilon) \neq 0$ and $N \geqslant q(2\epsilon)$. In this case, we know that $\mathcal{I}_{0,N}$ overlaps with $\mathcal{I}_{-p(2\epsilon),N-q(2\epsilon)}$ but the overlap is not complete. If it overlaps with this interval only then the length of the subinterval of $\mathcal{I}_{0,N}$ that is not overlapped — the "exposed" subinterval — is $|\eta(2\epsilon)|$. Suppose $\mathcal{I}_{0,N}$ overlaps other intervals as well. Suppose it

overlaps $\mathcal{I}_{i,j}$. Let $(p,q) = (-i, N-j)$ and let $\eta = q\alpha - p$. If η has the same sign as $\eta(2\epsilon)$ and greater absolute value then there is no effect on the length of the exposed subinterval. Recalling that $\eta'(2\epsilon) - \eta(2\epsilon)$ has opposite sign to $\eta(2\epsilon)$ and its absolute value is greater than 2ϵ, we deduce that an overlap with $\mathcal{I}_{i,j}$ will only affect the length of the exposed subinterval if

$$(\eta - \eta(2\epsilon))(\eta - \eta'(2\epsilon) + \eta(2\epsilon)) < 0.$$

From Proposition 3.3 of Chapter 2, we know that this can only be satisfied if $q \leqslant 0$ or $q \geqslant q'(2\epsilon)$. Therefore, we conclude that if $\eta(2\epsilon) \neq 0$ and $q(2\epsilon) \leqslant N < q'(2\epsilon)$ then no other overlaps occur which affect the length of the exposed subinterval and therefore

$$\mathcal{P}_{N+1} - \mathcal{P}_N = |\eta(2\epsilon)|. \tag{3.5}$$

Suppose $\eta(2\epsilon) \neq 0$ and $N \geqslant q'(2\epsilon)$. We know that $\mathcal{I}_{0,N}$ overlaps on one side with $\mathcal{I}_{-p(2\epsilon),N-q(2\epsilon)}$ and on the other side with $\mathcal{I}_{-p'(2\epsilon),N-q(2\epsilon)}$. These are the closest overlaps on either side so long as

$$(\eta - \eta(2\epsilon))(\eta - \eta'(2\epsilon)) \geqslant 0$$

for all admissible intervals $\mathcal{I}_{i,j}$ where, as before, $(p,q) = (-i, N-j)$ and $\eta = q\alpha - p$. Applying Proposition 3.3 of Chapter 2 once more, we conclude that this is the case when $q'(2\epsilon) \leqslant N < q(2\epsilon) + q'(2\epsilon)$. Consider the length of the exposed subinterval of $\mathcal{I}_{0,N}$ in this case. Suppose that $\eta(2\epsilon) > 0$ and $\eta'(2\epsilon) < 0$. The subinterval of $\mathcal{I}_{0,N}$ which is not already covered by the other intervals is then

$$(N\alpha + \eta'(2\epsilon) + \epsilon, N\alpha + \eta'(2\epsilon) - \epsilon).$$

If the signs are reversed then the interval is

$$(N\alpha + \eta(2\epsilon) + \epsilon, N\alpha + \eta(2\epsilon) - \epsilon).$$

In either case, the length is $|\eta(2\epsilon) - \eta'(2\epsilon)| - 2\epsilon$. We know that $\mathcal{I}_{0,N}$ is not completely covered since $|\eta(2\epsilon) - \eta'(2\epsilon)| > 2\epsilon$. Therefore, if $\eta(2\epsilon) \neq 0$ and $q'(2\epsilon) \leqslant N < q(2\epsilon) + q'(2\epsilon)$ then

$$\mathcal{P}_{N+1} - \mathcal{P}_N = |\eta(2\epsilon)| + |\eta'(2\epsilon)| - 2\epsilon. \tag{3.6}$$

Finally, suppose $\eta(2\epsilon) \neq 0$ and $N \geqslant q(2\epsilon) + q'(2\epsilon)$. In this case $\mathcal{I}_{0,N}$ is overlapped by

$$\mathcal{I}_{-p(2\epsilon),N-q(2\epsilon)}, \quad \mathcal{I}_{-p(2\epsilon)-p'(2\epsilon),N-q(2\epsilon)-q'(2\epsilon)} \quad \text{and} \quad \mathcal{I}_{-p'(2\epsilon),N-q'(2\epsilon)}.$$

The distances between the centres of these intervals are $|\eta'(2\epsilon)|$ for the former two and $|\eta(2\epsilon)|$ for the latter two. Since both of these distances are less than or equal to 2ϵ, we conclude that $\mathcal{I}_{0,N}$ is completely covered by the other intervals and so

$$\mathcal{P}_{N+1} - \mathcal{P}_N = 0 \tag{3.7}$$

in this case.

We can now express the probability of intercept for $N > 0$ as

(3.8)
$$\mathcal{P}_N = \begin{cases} 2\epsilon N & \text{if } 0 \leqslant N \leqslant q(2\epsilon), \\ 2\epsilon q(2\epsilon) & \text{if } N > q(2\epsilon) \text{ and } \eta(2\epsilon) = 0, \\ |\eta(2\epsilon)|N & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]q(2\epsilon) & \text{if } q(2\epsilon) < N \leqslant q'(2\epsilon) \text{ and } \eta(2\epsilon) \neq 0, \\ [|\eta(2\epsilon)| + |\eta'(2\epsilon)| - 2\epsilon]N & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]q(2\epsilon) & \\ \quad +[2\epsilon - |\eta'(2\epsilon)|]q'(2\epsilon) & \text{if } q'(2\epsilon) < N \leqslant q(2\epsilon) + q'(2\epsilon) \text{ and } \eta(2\epsilon) \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

The expression (3.8) arises from the summation of the terms (3.3), (3.4), (3.5), (3.6) and (3.7). We need only show that the last subexpression for $\mathcal{P}_N$ is correct, namely that $\mathcal{P}_N = 1$ when $\eta(2\epsilon) \neq 0$ and $N > q(2\epsilon) + q'(2\epsilon)$. Consider the value of $\mathcal{P}_N$ when $N = q(2\epsilon) + q'(2\epsilon)$. We have

$$\begin{aligned} \mathcal{P}_{q(2\epsilon)+q'(2\epsilon)} &= |\eta'(2\epsilon)|q(2\epsilon) + |\eta(2\epsilon)|q'(2\epsilon) \\ &= |\eta'(2\epsilon)q(2\epsilon) - \eta(2\epsilon)q'(2\epsilon)| \\ &= \left|\left(\eta_{n(2\epsilon)+1} - k\eta_{n(2\epsilon)}\right)q_{n(2\epsilon)} - \eta_{n(2\epsilon)}\left(q_{n(2\epsilon)+1} - kq_{n(2\epsilon)}\right)\right| \\ &= \left|\eta_{n(2\epsilon)+1}q_{n(2\epsilon)} - \eta_{n(2\epsilon)}q_{n(2\epsilon)+1}\right| \end{aligned}$$

where k is defined in (2.8). We then use statement (*viii*) of Proposition 3.1 of Chapter 2 to show that this expression must be equal to 1. Equation (3.7) then implies that $\mathcal{P}_N = 1$ for all $N > q(2\epsilon) + q'(2\epsilon)$ when $\eta(2\epsilon) \neq 0$.

Figure 2 illustrates the value of the characteristic function $\mathcal{C}_N(\beta)$ over the interval $[-1, 0]$ for $N = 5$, $N = 9$ and $N = 14$ where $\alpha = 0.217$ and $\epsilon = 0.05$. In this illustration, the interval over β is fixed whereas in the preceding discussion we allowed it to move in order to center $\mathcal{I}_{0,N}$. Nevertheless, from the topmost illustration, we can see that, for $N \leqslant 5$, the intervals $\mathcal{I}_{p,N}$ are separate and do not overlap. Therefore, the rate of growth of the probability of intercept is at its greatest. We call this the stage of NO OVERLAP. In the middle illustration, we see that, for $5 < N \leqslant 9$, the intervals $\mathcal{I}_{o,N}$ overlap on one side only. We call this the SINGLE OVERLAP stage. Finally, in the illustration at bottom, we see that, for $9 < N \leqslant 14$, the intervals $\mathcal{I}_{p,N}$ overlap previous intervals on both sides, filling in the last of the "gaps" in the integration interval. We call this the DOUBLE OVERLAP stage. For $N > 14$, the value of $\mathcal{C}_N(\beta) = 1$ everywhere and all new intervals $\mathcal{I}_{p,N}$ are completely overlapped by previous intervals. We call this the COMPLETE OVERLAP stage.

In Figure 3(a), we present a graph of the probability of intercept, $\mathcal{P}_N$, as a function of N using the same parameters that were used in Figure 2, *i.e.*, $\alpha = 0.217$

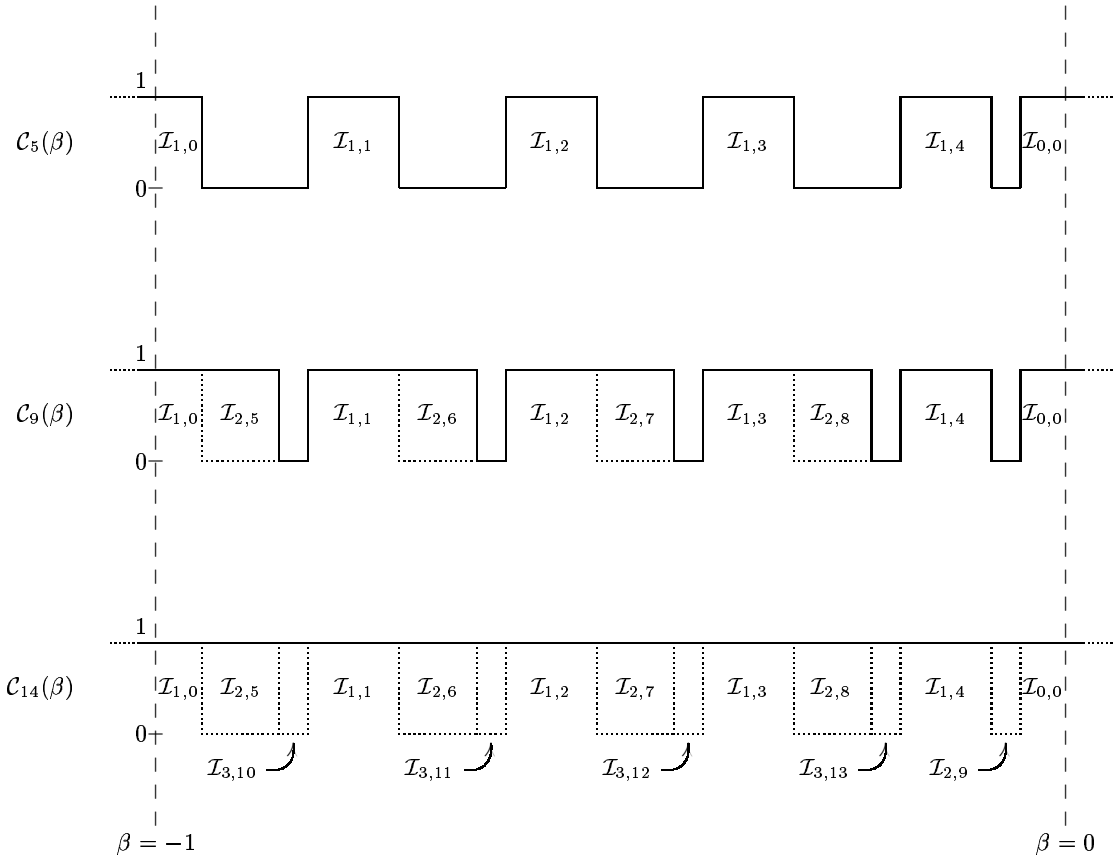


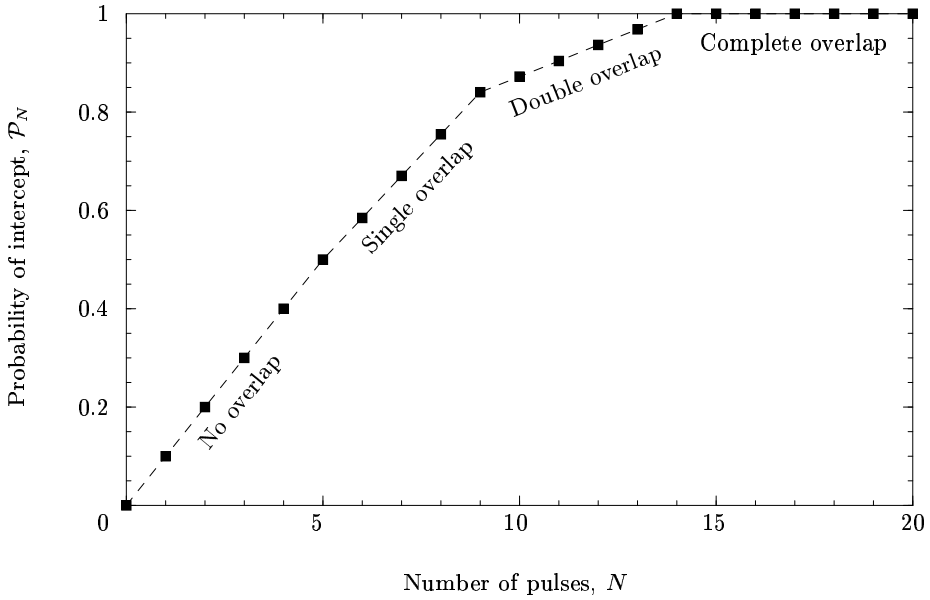
FIGURE 2. The value of the characteristic function $\mathcal{C}_N(\beta)$ for $N = 5$, $N = 9$ and $N = 14$.

and $\epsilon = 0.05$. The four linear segments are clearly visible in the graph and each has been labelled according to the corresponding stage.

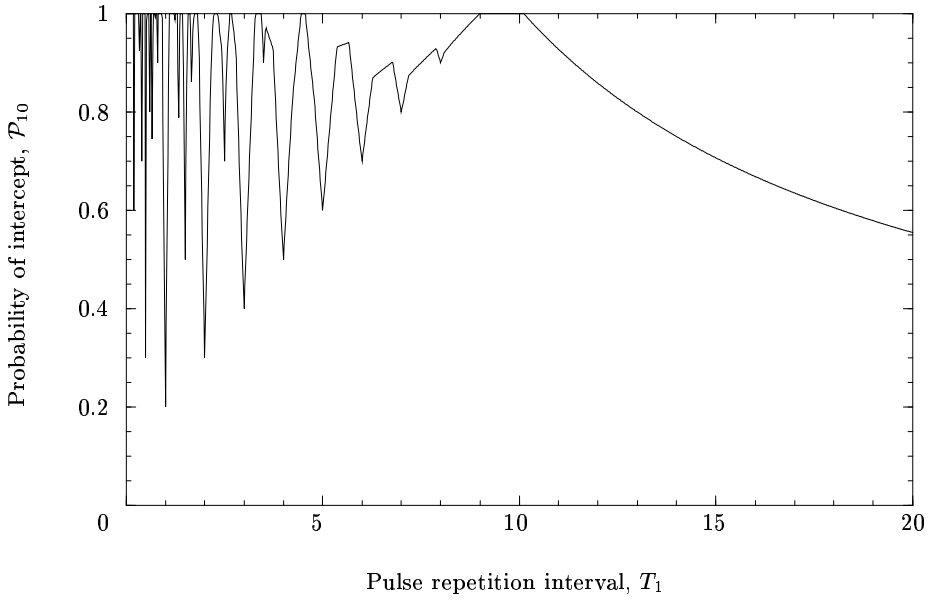
Note that the form of (3.8) enables efficient computation of plots of probability of intercept for a range of PRI ratios. For example, consider two pulse trains for which the PRI, T_2, and pulse width, τ_2, of the second pulse train is known, but only the DUTY CYCLE $\lambda = \tau_1/T_1$ of the first pulse train is known. We wish to find the probability of intercept after a prescribed number of pulses, N, from the second pulse train. By writing $\epsilon = (\lambda T_1 + \tau_2)/(2T_1)$, we can apply (3.8) and plot the probability as a function of T_1. A plot illustrating this appears as Figure 3(b) with both duty cycles set at 10% (*i.e.*, $\epsilon = 0.05 + 0.05\alpha^{-1}$) and $T_2 = 1$. From the plot, it is clear that the probability is highly erratic when $T_1 < 10$ before following a smooth decay for $T_1 > 10$.

Even faster methods for drawing this kind of graph, and computing averages from it, are described in Section 5.

3.2. Uniformly Random Phase for Both Pulse Trains. Another problem of interest is the case where the total observation time is known, but the phases of the pulse trains are not. We can interpret this problem as one where no control is exercised over the phases of either pulse train. RICHARDS (1948) discusses this



(a) Probability of intercept as a function of number of pulses from pulse train 2 with $\alpha = 0.217$ and $\epsilon = 0.05$.



(b) Probability of intercept as a function of the PRI of pulse train 2 with $N = 10$, $T_2 = 1$ and both duty cycles at 10%.

FIGURE 3. Plots of probability of intercept with a random time offset for one pulse train

problem and finds an approximate expression for the form of the probability of intercept.

RICHARDS shows that the solution to the problem can be found by considering the "phase space" of the pulse trains. Let us now consider this approach. Suppose

both pulse trains have a negative phase, which is to say that at least one pulse from each pulse train occurred prior to time $t = 0$. At any given time $t \geqslant 0$, both pulse trains are on (an intercept occurs) if there exists some $i, j \in \mathbb{Z}$ such that

$$|iT_1 + \phi_1 - t| \leqslant \tfrac{1}{2}\tau_1 \qquad \text{and} \qquad |jT_2 + \phi_2 - t| \leqslant \tfrac{1}{2}\tau_2. \tag{3.9}$$

Let us assume that the pulse which arrived just prior to $t = 0$ is labelled the 0^{th} for both pulse trains and that this labelling causes the distribution of the phases to be uniform. That is, if we let Φ_k, $k = 1, 2$, be the random variable representing the phase of the k^{th} pulse train then $\Phi_k \sim U(-T_k, 0)$. Let $\mathcal{I}_{i,j}(t)$ be the interval in $\mathbb{R}^2$ defined by

$$\mathcal{I}_{i,j}(t) = \big\{(x, y) \in \mathbb{R}^2 \;\mid\; |iT_1 + x - t| \leqslant \tfrac{1}{2}\tau_1;\; |jT_2 + y - t| \leqslant \tfrac{1}{2}\tau_2\big\}.$$

Clearly, $\mathcal{I}_{i,j}(t)$ has area $\tau_1\tau_2$. An intercept occurs at the time instant t if $(\phi_1, \phi_2) \in \mathcal{I}_{i,j}(t)$ for any $i, j \in \mathbb{Z}$. Therefore, the probability of intercept at that instant is $(\tau_1\tau_2)/(T_1T_2)$. Consider the probability of at least one intercept occurring over the time interval $[0, t]$. Let $\mathcal{C}(x, y;\, t)$ be the characteristic function of the set

$$\bigcup_{\substack{i,j\in\mathbb{Z};\\ 0\leqslant u\leqslant t}} \mathcal{I}_{i,j}(u).$$

An intercept occurs during the interval $[0, t]$ for a particular pair of phase ϕ_1 and ϕ_2 if and only if $\mathcal{C}(\phi_1, \phi_2;\, t) = 1$. Thus, the probability of intercept $\mathcal{P}(t)$ over the time interval $[0, t]$ is

$$\mathcal{P}(t) = \frac{1}{T_1T_2} \int_{-T_2}^{0} \int_{-T_1}^{0} \mathcal{C}(\phi_1, \phi_2;\, t)\, d\phi_1\, d\phi_2. \tag{3.10}$$

We observe that, since the characteristic function is periodic in its first two arguments with periods T_1 and T_2, respectively, the intervals of integration in (3.10) could be replaced by any intervals of lengths T_1 and T_2 for the appropriate integrands. We call the Cartesian product of any such choice of integration intervals the PHASE SPACE or PHASE PLANE, since it contains all possible choices of phase, up to periodicities.

We now discuss a geometric interpretation of (3.10). Consider again the sets $\mathcal{I}_{i,j}(t)$. Let i and j be fixed as we vary t. The rectangle described by $\mathcal{I}_{i,j}(t)$ moves an equal distance up and across as we increase t. Therefore, for fixed i and j, the union

$$\bigcup_{0\leqslant u\leqslant t} \mathcal{I}_{i,j}(t)$$

is a hexagonal region formed by "dragging" the rectangle $\mathcal{I}_{i,j}(0)$ diagonally up and across by t. Now, since the characteristic function we are integrating in (3.10) is periodic in its first two arguments and the length of the integration interval is the period for each integrand, we can think of the characteristic function being integrated over a torus, rather than a rectangle. The support of the characteristic function is

the hexagonal region wrapped around the torus. We illustrate this in Figure 4. It

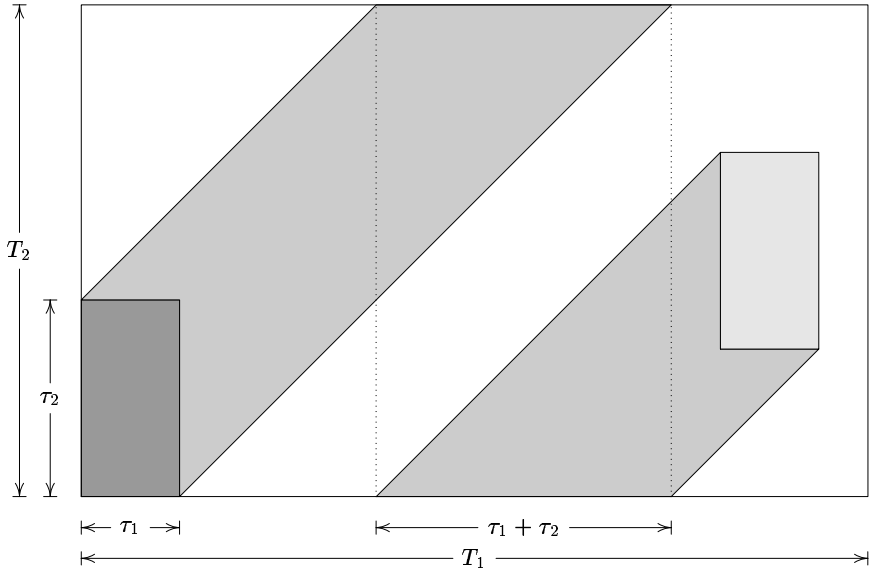


FIGURE 4. Region of the phase space within which pulse coincidence can occur

shows an example of how the rectangle $\mathcal{I}_{i,j}(0)$ is dragged over the torus. The most darkly shaded area shows where the rectangle started at time 0, which we call the INITIAL rectangle, the medium shading is the area covered, ending at the lightly shaded rectangle at time t, which we call the LEADING rectangle. The probability of intercept $\mathcal{P}(t)$ is the ratio of the shaded areas to the area of the larger rectangle (torus).

In contrast to the discrete time problem, we note that $\mathcal{P}(0) = (\tau_1\tau_2)/(T_1T_2)$ and is in general non-zero. On the other hand, we will now see that otherwise the problem bears a close resemblance to the earlier problem. Observe how the "swathe" cut by the leading rectangle crosses the boundary of the integration rectangle as t is increased. Assume for definiteness that the boundary under consideration is a boundary along which ϕ_2 is constant. In Figure 4, this is an edge of the integration rectangle with length T_1. The width of the swathe across the boundary is $\tau_1 + \tau_2$. Each time the rectangle returns to the boundary, it will be a distance T_2 further along. This is very similar to the construction we employed in solving the discrete time probability of intercept problem. That is, if we consider only the boundary crossings then they behave like accumulation of the intervals $\mathcal{I}_{p,q}$ on $\mathbb{R}$ in the discrete time problem. Therefore, we should be able to use the discrete time solution to determine the number of crossings before single overlaps occur, the number before double overlaps occur and the number before the entire length of the boundary has been covered. We know that the time between crossings is T_2.

We therefore expect the probability of intercept over time to consist of four linear segments corresponding to the four segments in the expression for $\mathcal{P}_N$ in (3.8). Instead of the integer index N, the probability of intercept is now a function

of the continuous time argument, t, and we expect the transitions between segments to occur at integer multiples of T_2. However, there is a slight advance or delay according to when the leading rectangle overlaps with the initial rectangle. The exact amount can be worked out by simple geometrical considerations. In addition to the slight advance or delay, there will also be a short period of time when the rectangles are "meshing" during which the rate of growth of the probability is non-linear (in fact it is quadratic).

Let us again normalise all the parameters with respect to T_1. Thus, we have $\alpha = T_2/T_1$, $\epsilon_1 = \tau_1/(2T_1)$ and $\epsilon_2 = \tau_2/(2T_1)$. Let $\epsilon = \epsilon_1 + \epsilon_2$. The symbols $p(2\epsilon)$, $q(2\epsilon)$, $\eta(2\epsilon)$, and so on, have their usual meanings with regard to the s.c.f. expansion of α. We can now write an expression for the probability of intercept thus:

(3.11)

$$\mathcal{P}(t) = \frac{h(t)}{T_1 T_2} + \frac{1}{T_2} \begin{cases} 4T_1\epsilon_1\epsilon_2 + 2\epsilon t & \text{if } 0 \leqslant t \leqslant t_1, \\ 4T_1\epsilon_1\epsilon_2 + 2\epsilon t_1 & \text{if } \eta(2\epsilon) = 0 \text{ and } t > t_1, \\ 4T_1\epsilon_1\epsilon_2 + |\eta(2\epsilon)|t & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]t_1 & \text{if } \eta(2\epsilon) \neq 0 \text{ and } t_1 < t \leqslant t_2, \\ 4T_1\epsilon_1\epsilon_2 & \\ \quad +[|\eta(2\epsilon)| + |\eta'(2\epsilon)| - 2\epsilon]t & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]t_1 & \\ \quad +[2\epsilon - |\eta'(2\epsilon)|]t_2 & \text{if } \eta(2\epsilon) \neq 0 \text{ and } t_2 < t \leqslant t_3, \\ T_2 - \nu^2/T_1 & \text{if } \eta(2\epsilon) \neq 0 \text{ and } t > t_3, \end{cases}$$

where

$$\begin{aligned} t_1 &= T_2 q(2\epsilon) - T_1 \min\{2\epsilon_1 + \eta(2\epsilon), 2\epsilon_2\}, \\ t_2 &= T_2 q'(2\epsilon) - T_1 \min\{2\epsilon_1 + \eta'(2\epsilon), 2\epsilon_2\}, \\ t_3 &= T_2[q(2\epsilon) + q'(2\epsilon)] - T_1 \min\{2\epsilon_1 + \eta(2\epsilon) + \eta'(2\epsilon), 2\epsilon_2\} \end{aligned}$$

and

$$h(t) = \begin{cases} 0 & \text{if } t \leqslant t^\star, \\ (t - t^\star)(2\nu - t + t^*) & \text{if } t^\star < t \leqslant t^* + \nu, \\ \nu^2 & \text{if } t > t^\star + \nu, \end{cases}$$

where

$$\nu = T_1 \min\{|2\epsilon_1 + \omega_1|, |2\epsilon_2 - \omega_1|, |2\epsilon_1 + \omega_2|, |2\epsilon_2 - \omega_2|\}$$

and

$$(\omega_1, \omega_2, t^*) = \begin{cases} (\eta(2\epsilon), 0, t_1) & \text{if } -2\epsilon_1 \leqslant \eta(2\epsilon) \leqslant 2\epsilon_2, \\ (0, \eta'(2\epsilon), t_2) & \text{if } -2\epsilon_1 \leqslant \eta'(2\epsilon) \leqslant 2\epsilon_2, \\ (\eta(2\epsilon), \eta'(2\epsilon), t_3) & \text{otherwise.} \end{cases}$$

The function $h(t)$ represents the quadratic segment which occurs when the leading rectangle meshes with the initial rectangle. Notice that if $\tau_1\tau_2 \ll T_1T_2$ then $h(t)$ will be negligible. Furthermore, t_1, t_2 and t_3 will approximate $q(2\epsilon)T_2$, $q'(2\epsilon)T_2$ and $[q(2\epsilon) + q'(2\epsilon)]T_2$, respectively.

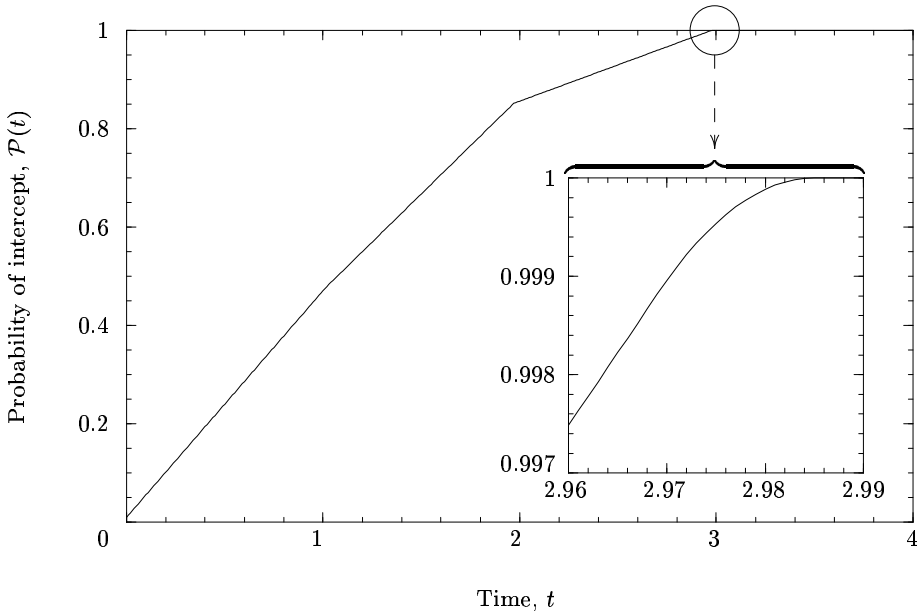
Figure 5(a) is a plot of a similar style to Figure 3(a), using similar parameters. The PRIs and pulse widths have been selected to ensure that $\alpha = 0.217$ and $\epsilon = 0.05$. It shows that the forms of the probabilities are very similar, with the exception that we are now dealing with a continuous quantity (time) as our independent axis rather than a discrete quantity (number of pulses). It is quite difficult to discern the extra, quadratic segment in the plot because it is very small in this case. For this reason, the quadratic segment is show in the inset.

Figure 5(b) is nearly identical to Figure 3(b). Indeed, inspection of the values reveals that the difference between the functions is less than 0.009 at any point. Hence, it would appear that, with some small modifications, the expression for the probability of intercept which was derived for the discrete case could be used to approximate the probability of intercept in the continuous case to a high degree of accuracy. This is especially true when the pulse widths are small compared to the PRIs. We now discuss the construction of such an approximation.

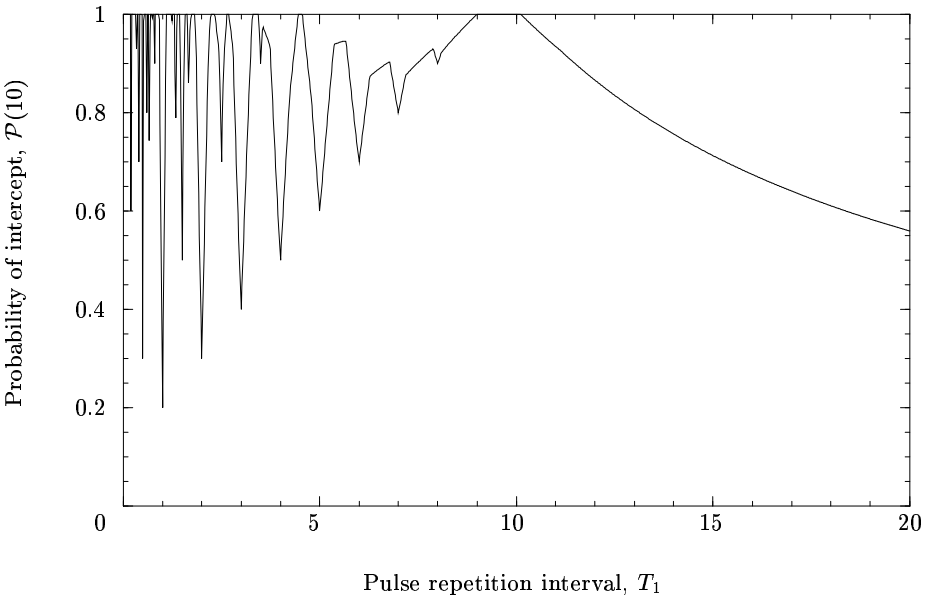
3.3. Approximation for the Probability of Intercept. We have seen from Figure 5(b) that it should be possible to approximate the expression of the probability of intercept when both phases are random in (3.11) with a simpler expression resembling that of the probability of intercept when only one phase is random. Such an expression is now given:

(3.12)

$$\hat{\mathcal{P}}(t) = \frac{1}{T_2} \begin{cases} 2\epsilon t & \text{if } 0 \leqslant t \leqslant q(2\epsilon)T_2, \\ 2\epsilon q(2\epsilon) & \text{if } \eta(2\epsilon) = 0 \text{ and } t > q(2\epsilon)T_2, \\ |\eta(2\epsilon)|t & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]q(2\epsilon)T_2 & \text{if } \eta(2\epsilon) \neq 0 \text{ and } q(2\epsilon)T_2 < t \leqslant q'(2\epsilon)T_2, \\ [|\eta(2\epsilon)| + |\eta'(2\epsilon)| - 2\epsilon]t & \\ \quad +[2\epsilon - |\eta(2\epsilon)|]q(2\epsilon)T_2 & \\ \quad +[2\epsilon - |\eta'(2\epsilon)|]q'(2\epsilon)T_2 & \text{if } \eta(2\epsilon) \neq 0 \\ & \text{and } q'(2\epsilon)T_2 < t \leqslant [q(2\epsilon) + q'(2\epsilon)]T_2, \\ T_2 & \text{if } \eta(2\epsilon) \neq 0 \text{ and } t > [q(2\epsilon) + q'(2\epsilon)]T_2. \end{cases}$$



(a) Probability of intercept as a function of time with $T_1 = 1$, $T_2 = 0.217$, $\tau_1 = 0.07$ and $\tau_2 = 0.03$.



(b) Probability of intercept as a function of the PRI of pulse train 2 with $t = 10$, $T_2 = 1$ and both duty cycles at 10%.

FIGURE 5. Plots of probability of intercept with random time offsets for both pulse trains

Note the close similarity between the expression for $\hat{\mathcal{P}}(t)$ in (3.12) and that for $\mathcal{P}_N$ in (3.8). The expression for the former is now simply a linear interpolation between the discrete points of the latter, and with a scaling in the time axis by T_2.

Consider the error of our approximate expression for the continuous time probability of intercept. There are five sources of difference between the approximate expression (3.12) and the true expression (3.11): the absence of the quadratic segment $h(t)$, the absence of the initial probability $\mathcal{P}(0) = (\tau_1\tau_2)/(T_1T_2)$ and the differences in the (three) time boundaries between the linear segments. In each case it can be shown that the total contribution to the error is less that $(\tau_1 + \tau_2)^2/(T_1T_2)$ and so we conclude that

$$\left|\hat{\mathcal{P}}(t) - \mathcal{P}(t)\right| \leqslant 5\frac{(\tau_1 + \tau_2)^2}{T_1T_2}.$$

Hence, so long as $\tau_1 \ll T_1$ and $\tau_2 \ll T_2$, *i.e.* both duty cycles are small, then we can use the approximation with only a very small error.

4. Mean Time to Intercept of Two Pulse Trains

Calculating the mean time[1] to intercept is a straightforward extension of the solution to the probability of intercept problem discussed previously. From the expression for the probability of intercept in the discrete time case of (3.8), we can express the mean time to intercept as

(4.1)
$$\mathrm{E}[N] = \tfrac{1}{2} + \tfrac{1}{2}|\eta'(2\epsilon)|q(2\epsilon)^2 + \tfrac{1}{2}|\eta(2\epsilon)|q'(2\epsilon)^2 + [|\eta(2\epsilon)| + |\eta'(2\epsilon)| - 2\epsilon]q(2\epsilon)q'(2\epsilon).$$

A similar expression can be derived for the continuous time case, although the resulting expression is not as neat. From consideration of (3.11), we get

$$\mathrm{E}[t] = [T_1\epsilon\big(t_1^2 + t_2^2 - t_3^2\big) + T_1|\eta(2\epsilon)|\big(t_3^2 - t_1^2\big) + T_1|\eta'(2\epsilon)|\big(t_3^2 - t_2^2\big) + 2t^*\nu^2 + \tfrac{2}{3}\nu^3]/(2T_1T_2).$$

Figure 6 shows the mean time to intercept plotted using (4.1) for a range of PRI ratios and $\epsilon = 0.05$ in the discrete time case. The plot indicates that the minimum mean time is 5.5. This occurs because the rate of growth of the probability of intercept cannot exceed 2ϵ. That is, $\mathcal{P}_N \leqslant 2\epsilon N$ and so we can deduce that the mean time to intercept satisfies

(4.2)
$$\mathrm{E}[N] \geqslant \sum_{n=1}^{\left\lfloor \frac{1}{2}\epsilon^{-1}\right\rfloor} 2\epsilon n = \epsilon\left\lfloor \tfrac{1}{2}\epsilon^{-1}\right\rfloor\left(\left\lfloor \tfrac{1}{2}\epsilon^{-1}\right\rfloor + 1\right).$$

Substituting for ϵ, we find that (4.2) is in agreement with the observed minimum in Figure 6. Notice the symmetry of the plot about $\alpha = 0.5$. We can also see that the mean time becomes very large at several points on the graph. In fact, it approaches infinity. The points at which this occurs are from a Farey series, and their relevance is now discussed in greater detail.

[1]To be precise, we should write "mean number of pulses from pulse train 2 to intercept" in the discrete time case, but we write "mean time" for convenience.

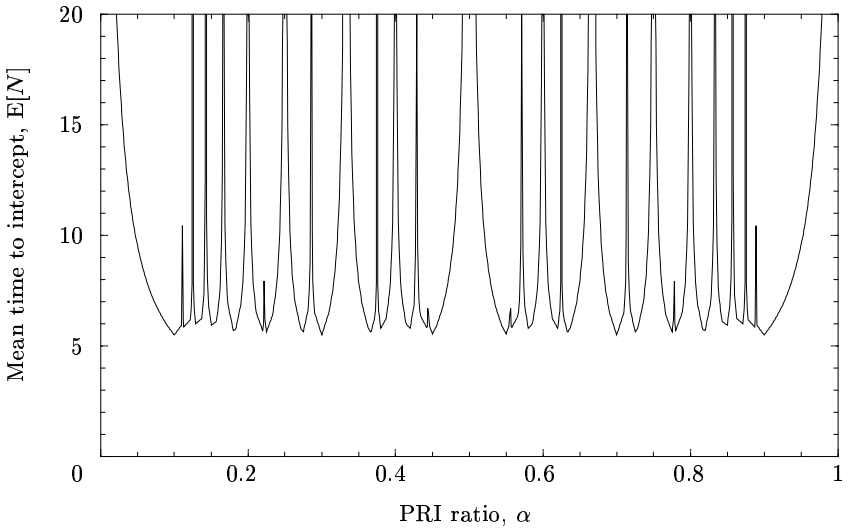


FIGURE 6. Mean time to intercept, E[N], plotted against the PRI ratio α with $\epsilon = 0.05$.

5. Relationship with Farey Series

We now discuss the situation in which the PRIs of the pulse trains are not known exactly *a priori*, but are known to lie within some range. We assume that, although a PRI may be unknown, it is a constant. We firstly discuss how the probability of intercept changes as we vary a PRI, holding all other parameters constant.

Consider how $q(2\epsilon)$ changes as we vary the PRI ratio α. If we write $q(2\epsilon)$ as a function of α also, *i.e.* as $q(2\epsilon, \alpha)$ then we can show that $q(2\epsilon, \alpha)$ is piecewise constant. We will show that the intervals on which the function is constant surround points in a Farey series of the appropriate order. To see this, recall Theorem 7.2 from Chapter 2. Let $\mathfrak{F}(\epsilon)$ denote the Farey series of order $\lceil \epsilon^{-1} \rceil - 1$. From Theorem 7.2, we see that if $h/k < h'/k'$ are adjacent elements in the Farey series $\mathfrak{F}(2\epsilon)$ and $h/k \leqslant \alpha \leqslant h'/k'$ then

$$\frac{p(2\epsilon)}{q(2\epsilon)} \in \left\{ \frac{h}{k}, \frac{h'}{h'} \right\}.$$

Deciding between the two elements is simple and follows directly from the definition of a best approximation. Suppose $k < k'$. If $|k\alpha - h| \leqslant 2\epsilon$ then $p(2\epsilon)/q(2\epsilon) = h/k$, otherwise h'/k'. On the other hand, suppose $k' < k$. If $|k'\alpha - h'| \leqslant 2\epsilon$ then $p(2\epsilon)/q(2\epsilon) = h'/k'$, otherwise h/k. Note that $k \neq k'$ since the order of the Farey series in question is greater than 1 because $\epsilon < \frac{1}{2}$.

Therefore, we see that $p(2\epsilon, \alpha)$ and $q(2\epsilon, \alpha)$ is piecewise constant over α. The points of transition between constant values are dictated by the elements of $\mathfrak{F}(2\epsilon)$. It is clear that $\eta(2\epsilon)$ will therefore be piecewise linear.

In order to calculate the probability of intercept over a range of α, it is also necessary to find $q'(2\epsilon, \alpha)$ and $\eta'(2\epsilon, \alpha)$ when $\eta(2\epsilon) \neq 0$. We make the following definition.

DEFINITION 5.1. The left and right PARENT of the Farey point h/k, $k > 1$, are those two Farey points which are adjacent in a lower order such that h/k is their mediant. The left parent is the lesser of the two parents.

To find $q'(2\epsilon)$ and $\eta'(2\epsilon)$ for a given α from information in the Farey series, it is sufficient to find $p_{n(2\epsilon)-1}/q_{n(2\epsilon)-1}$. From Theorem 7.4 of Chapter 2, we know that $p_{n(2\epsilon)-1}/q_{n(2\epsilon)-1}$ is one of the parents of $p(2\epsilon)/q(2\epsilon)$. Specifically, it is the left parent if $\alpha < p(2\epsilon)/q(2\epsilon)$ or the right parent otherwise. Given these two convergents, we can calculate $p'(2\epsilon)$, $q'(2\epsilon)$ and $\eta'(2\epsilon)$ directly from (2.7).

Given adjacent elements $h/k < h'/k'$ in $\mathfrak{F}(2\epsilon)$ and the right parent of h/k and the left parent of h'/k', we have enough information to directly calculate the probability of intercept for α over the entire interval $[h/k, h'/k']$, with all other parameters being held constant. Let H/K be the right parent of h/k and H'/K' be the left parent of h'/k'. An expression for the probability of intercept, $\mathcal{P}_N(\alpha)$, over the interval is

$$\mathcal{P}_N(\alpha) = \begin{cases} 2\epsilon N & \text{if } N \leqslant k \text{ and } x_0 \leqslant \alpha \leqslant x_1, \\ |\eta(2\epsilon,\alpha)|N & \\ \quad +[2\epsilon - |\eta(2\epsilon,\alpha)|]q(2\epsilon,\alpha) & \text{if } N > k \text{ and } x_0 \leqslant \alpha \leqslant d_1, \\ [|\eta(2\epsilon,\alpha)| + |\eta'(2\epsilon,\alpha)| - 2\epsilon] & \\ \quad +[2\epsilon - |\eta(2\epsilon,\alpha)|]q(2\epsilon,\alpha) & \\ \quad +[2\epsilon - |\eta'(2\epsilon,\alpha)|]q'(2\epsilon,\alpha) & \text{if } N > k \text{ and } d_1 < \alpha \leqslant f_1, \\ 1 & \text{if } N > k \text{ and } f_1 < \alpha \leqslant x_1, \\ 1 & \text{if } N > k' \text{ and } x_1 \leqslant \alpha < f_2, \\ [|\eta(2\epsilon,\alpha)| + |\eta'(2\epsilon,\alpha)| - 2\epsilon] & \\ \quad +[2\epsilon - |\eta(2\epsilon,\alpha)|]q(2\epsilon,\alpha) & \\ \quad +[2\epsilon - |\eta'(2\epsilon,\alpha)|]q'(2\epsilon,\alpha) & \text{if } N > k' \text{ and } f_2 \leqslant \alpha < d_2, \\ |\eta(2\epsilon,\alpha)|N & \\ \quad +[2\epsilon - |\eta(2\epsilon,\alpha)|]q(2\epsilon,\alpha) & \text{if } N > k' \text{ and } d_2 \leqslant \alpha \leqslant x_2, \\ 2\epsilon N & \text{if } N \leqslant k' \text{ and } x_1 \leqslant \alpha \leqslant x_2. \end{cases} \tag{5.1}$$

where

$$x_0 = \frac{h}{k},$$

$$x_1 = \begin{cases} \dfrac{h+2\epsilon}{k} & \text{if } k < k', \\ \dfrac{h'-2\epsilon}{k'} & \text{otherwise,} \end{cases}$$

$$x_2 = \frac{h'}{k'}$$

and

$$(p(2\epsilon,\alpha),q(2\epsilon,\alpha)) = \begin{cases} (h,k) & \text{if } x_0 \leqslant \alpha < x_1, \\ (h',k') & \text{if } x_1 < \alpha \leqslant x_2, \end{cases}$$

$$(p'(2\epsilon,\alpha),q'(2\epsilon,\alpha)) = \begin{cases} (H+\kappa(2\epsilon,\alpha)h, K+\kappa(2\epsilon,\alpha)k) & \text{if } x_0 \leqslant \alpha < x_1, \\ (H'+\kappa(2\epsilon,\alpha)h', K'+\kappa(2\epsilon,\alpha)k') & \text{if } x_1 < \alpha \leqslant x_2. \end{cases}$$

The values for the approximation errors $\eta(2\epsilon,\alpha)$ and $\eta'(2\epsilon,\alpha)$ are calculated in the usual way. The intermediate points d_1, f_1, f_2 and d_2 define the boundaries between single overlap, double overlap, complete overlap, double overlap and single overlap, in ascending order of α. Expressions for these points are as follows:

$$d_1 = \min\left\{x_0 + \frac{1-2\epsilon k}{kq'(2\epsilon,\alpha)}, x_1\right\}, \tag{5.2}$$

$$f_1 = \min\left\{x_0 + \frac{1-2\epsilon k}{kq'(2\epsilon,\alpha)-k^2}, x_1\right\},$$

$$f_2 = \max\left\{x_2 - \frac{1-2\epsilon k'}{k'q'(2\epsilon,\alpha)-k'^2}, x_1\right\}$$

and

$$d_2 = \max\left\{x_2 - \frac{1-2\epsilon k'}{k'q'(2\epsilon,\alpha)}, x_1\right\}. \tag{5.3}$$

The value of $\kappa(2\epsilon,\alpha)$ is constant on the intervals $(d_1,f_1]$ and $[f_2,d_2)$ and on these intervals can be expressed as

$$\kappa(2\epsilon,\alpha) = \begin{cases} \left\lceil \dfrac{N-K}{k} \right\rceil & \text{if } \alpha \in (d_1,f_1], \\ \left\lceil \dfrac{N-K'}{k'} \right\rceil & \text{if } \alpha \in [f_2,d_2). \end{cases}$$

It should be noted that the choice of strict or weak inequalities is rather arbitrary in the above expressions. This is because the probability of intercept is a continuous function of α and is insensitive to which case is used on the boundary points. However, we have used strict inequalities in (5.1) to prevent the expression of $p_N(\alpha)$ from becoming any more awkward than it is already. We have used the strict inequalities in (5.1) in conjunction with the $\min\{\cdot\}$ and $\max\{\cdot\}$ notation of (5.2)–(5.3) as a shorthand way of testing whether the boundaries between regions occur on the "correct" side of x_1, and thereby to determine if these regions exist at all.

Also observe that we can adapt the probability of intercept expression of (5.1) to serve as an approximation to the probability of intercept in the continuous time case by everywhere replacing occurrences of N with t/T_2.

Figure 7 plots the probability of intercept as a function of the number of pulses from the first pulse train, N, and the PRI ratio α with $\epsilon = 0.05$. The plot consists

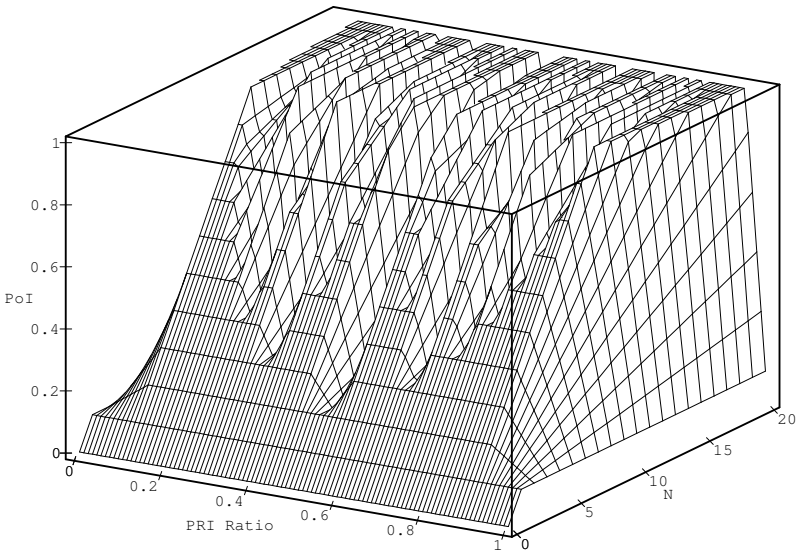


FIGURE 7. Plot of the discrete time probability of intercept as a function of the PRI ratio, α, and number of pulses, N, with $\epsilon = 0.05$.

of a sloping face for small N, levelling out when the probability reaches unity. The face has several valleys gouged out around certain PRIs. These valleys are centered about the Farey points, as we discussed above.

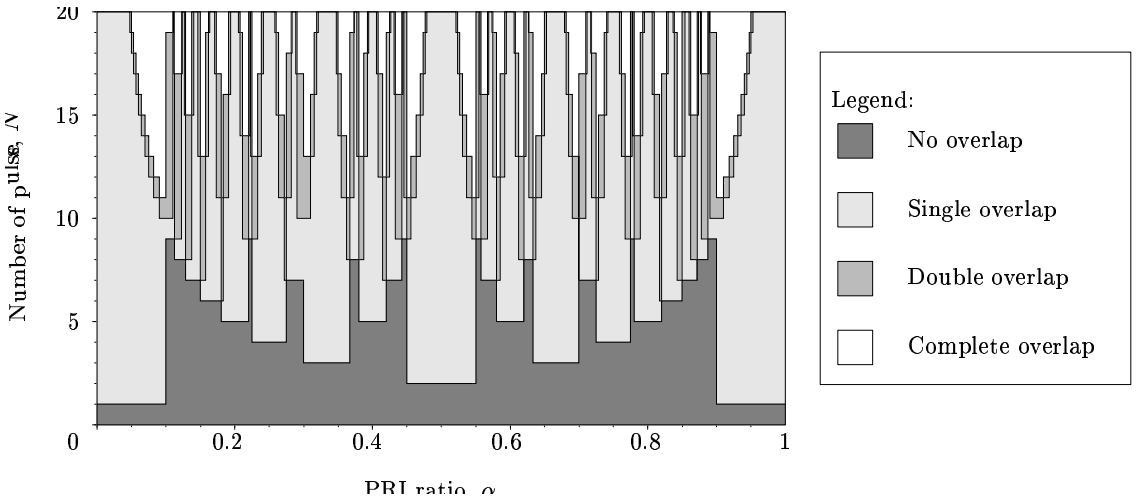


FIGURE 8. Probability of intercept shown by regions

Figure 8 shows how the discrete time probability of intercept as plotted in Figure 7 can be interpreted in terms of "regions." For any given N and α, it shows whether the probability of intercept lies in the region of no overlap, single or double overlap or complete overlap. The boundaries of these regions were computed using the expression (5.1). The probabilities are linear within these regions, so integration or averaging becomes a simple task once the boundaries are known. The recursive Algorithm 7.1 from Chapter 2 can be modified for this purpose.

6. Simultaneous Coincidence of More Than Two Pulse Trains

The computational problem of the approximate coincidence of many pulse trains is made difficult by the computational complexity of simultaneous Diophantine approximation, to which each of these problems can be reduced. In this section, we will examine the intercept time problem and the probability of intercept problem. For arbitrary numbers of pulse trains, we cannot say very much about the form of the solutions. For three pulse trains, we can use the algorithms of Chapter 4 to obtain solutions in some instances.

6.1. Intercept Time. Simultaneous approximate coincidence of n pulse trains occurs when there is a group of pulses, one from each pulse train, that have TOAs which are sufficiently close to one another. That is, simultaneous approximate coincidence occurs when

$$k_1T_1 + \phi_1 \approx k_2T_2 + \phi_2 \approx \cdots \approx k_nT_n + \phi_n$$

where, as usual, T_j and ϕ_j are the PRI and phase of the j^{th} pulse train and k_j is the (integer) pulse index. More particularly, an approximate coincidence occurs if

$$k_jT_j + \phi_j - t \approx 0$$

at some time t. To pose this problem a little more precisely, we could construct an inhomogeneous lattice Ω in $\mathbb{R}^n$ defined so that

$$\Lambda = \{(k_1T_1 + \phi_1, k_2T_2 + \phi_2, \ldots, k_nT_n + \phi_n) \mid k_1, k_2, \ldots, k_n \in \mathbb{Z}\}.$$

Approximate coincidence occurs if there exists some $\mathbf{v} \in \Omega$ such that $\mathbf{v} - t\mathbf{1} \approx \mathbf{0}$. Thus, according to this formulation, the problem is one of finding points of Λ which lie sufficiently close to the line $\mathbb{R}\mathbf{1}$. We can reformulate this in terms of a homogeneous lattice

$$\Omega = \{(k_1T_1, k_2T_2, \ldots, k_nT_n) \mid k_1, k_2, \ldots, k_n \in \mathbb{Z}\}$$

and the inhomogeneous line $\mathbb{R}\mathbf{1} - (\phi_1, \phi_2, \ldots, \phi_n)$. With the choice of an appropriate "distance function" from the lattice points to the line, the problem becomes well-posed.

When we associate the pulse width τ_j to each of the pulses of the j^{th} pulse train, a simultaneous coincidence occurs at time t when

$$\begin{aligned} |k_1T_1 + \phi_1 - t| &\leqslant \tfrac{1}{2}\tau_1, \\ |k_2T_2 + \phi_2 - t| &\leqslant \tfrac{1}{2}\tau_2, \\ &\vdots \\ |k_nT_n + \phi_n - t| &\leqslant \tfrac{1}{2}\tau_n. \end{aligned}$$

In the lattice formulation, the choice of distance function is a scaling of the sup-norm defined so that

$$\|\mathbf{x}\| = \max\left\{\left|\frac{x_1}{\tau_1}\right|, \left|\frac{x_2}{\tau_2}\right|, \ldots, \left|\frac{x_n}{\tau_n}\right|\right\}.$$

A simultaneous coincidence occurs if $\|\mathbf{v} - t\mathbf{1}\| \leqslant 2$ for some lattice point $\mathbf{v} \in \Omega$ and some $t \in \mathbb{R}$. Geometrically, we can picture a rectangular prism centred about each of lattice points, with sides of length $\tau_1, \tau_2, \ldots, \tau_n$. The problem is to find the intersection of these prisms with the line $\mathbb{R}\mathbf{1}$.

By taking the projection along $\mathbb{R}\mathbf{1}$ onto an orthogonal hyperplane, we discover that a simultaneous coincidence occurs if and only if

$$|k_i T_i + \phi_i - k_j T_j - \phi_j| \leqslant \tfrac{1}{2}(\tau_i + \tau_j)$$

for all $1 \leqslant i, j \leqslant n$. Of course, we can deduce this directly by noticing that a simultaneous coincidence occurs if and only if there is a coincidence between every pair of pulse trains i and j.

For two pulse trains, we were able to calculate intercept time for both in phase and arbitrary phase initial conditions: in the former case using Euclid's algorithm and in the latter case using Cassels' algorithm. We were also able to find further intercepts from a given one using a recurrence relation. The author knows of no simple analogue of these results for multiple pulse trains, save only for the case of in phase initial conditions with three pulse trains.

For the case of three pulse trains where $\phi_1 = \phi_2 = \phi_3$ — the in phase initial conditions — we are able to construct a simultaneous Diophantine approximation system, as defined in Definition 2.6 of Chapter 4, and apply either Algorithm 4.1 or Algorithm 5.1 of the same chapter to find the first intercept with respect to one of the pulse indices. The details of this construction were given in Example 4.3 of Chapter 4.

6.2. Probability of Intercept. For the problem involving two pulse trains, we distinguished two subproblems of probability of intercept: the discrete time problem where the phase of one of the pulse trains was known and the other uniformly random over the range of its PRI, and the continuous time problem where both phases were uniformly random. For more than two pulse trains, we have a greater choice. We can specify that m out of n phases are known and the rest are uniformly random. For the problem of two pulse trains, we were able to determine that the probability of intercept as a function of pulse index or time consisted of at most four linear segments (plus a fifth quadratic segment in the continuous time subproblem). We were also able to identify the relationship with the Farey series, allowing us to examine the behaviour of the probability of intercept as the ratio of PRIs was varied. Can any of these properties be carried over into problems involving more than two pulse trains? Again, the author knows of no simple analogue to these results. However, we are

able to prove a negative result. We will see that the number of linear segments in the expression for the probability of intercept is unbounded.

We consider two problems which we will call the discrete time and continuous time problems for the probability of intercept of more than two pulse trains. In the discrete time problem, one phase is known and the rest are uniformly random over the ranges of their respective PRIs. In the continuous time problem, all are assumed to be uniformly random.

Consider the discrete time problem. Let ϕ_n be the phase that is known and suppose it is 0. Let the other phases be instances of the uniform random variable $\Phi_j \sim U(-T_j, 0)$. Consider the sets

$$\begin{aligned}\mathcal{I}_{k_1,k_2,\ldots,k_n} = \{\mathbf{x} \in \mathbb{R}^{n-1} \mid\ & |k_i T_i + x_i - k_j T_j - x_j| \leqslant \tfrac{1}{2}(\tau_i + \tau_j); \\ & |k_i T_i + x_i - k_n T_n| \leqslant \tfrac{1}{2}(\tau_i + \tau_n);\ 1 \leqslant i, j < n\}\end{aligned}$$

which are solid polyhedra in $\mathbb{R}^{n-1}$. For $n = 3$, they are hexagons. If, for a given instance of the random variables, $(\phi_1, \phi_2, \ldots, \phi_{n-1}) \in \mathcal{I}_{k_1,k_2,\ldots,k_n}$ for some $k_1, k_2, \ldots, k_n \in \mathbb{Z}$ and $0 \leqslant k_n < N$ then an intercept occurs within the first N pulses from pulse train n. If we set $\mathcal{C}_N(\mathbf{x})$ to be the characteristic function of the union

$$\bigcup_{\substack{k_1,k_2,\ldots,k_n \in \mathbb{Z}; \\ 0 \leqslant k_n < N}} \mathcal{I}_{k_1,k_2,\ldots,k_n}$$

then the probability of intercept after N pulses from pulse train n is

$$\mathcal{P}_N = \frac{1}{T_1 T_2 \cdots T_{n-1}} \int_{-T_{n-1}}^{0} \cdots \int_{-T_2}^{0} \int_{-T_1}^{0} \mathcal{C}_N(\phi_1, \phi_2, \ldots, \phi_{n-1})\, d\phi_1\, d\phi_2\, \cdots\, d\phi_{n-1}.$$

Because $\mathcal{P}_N(\mathbf{x})$ is periodic in each of its arguments with periods $T_1, T_2, \ldots, T_{n-1}$, respectively, and these are also the lengths of the integration intervals for each of the integrands, we can replace these intervals with any intervals of the appropriate length. That is, we can regard the Cartesian product of these intervals as an $(n-1)$-dimensional hypertorus.

Figure 9 illustrates the situation for a problem involving three pulse trains after $N = 30$ pulses from pulse train 3. The parameters used were $T_1 = 1$, $T_2 = 5^{-1/3}$, $T_3 = 5^{-2/3}$ and $\tau_1 = \tau_2 = \tau_3 = 0.05$. The shaded hexagons are the sets $\mathcal{I}_{k_1,k_2,k_3}$ for $0 \leqslant k_3 < 30$. The more darkly shaded hexagon has index $k_3 = 29$ and overlaps the hexagon with index $k_3 = 0$. That is, for $N < 30$, the probability of intercept has been in the no overlap stage but is about to enter the single overlap stage.

However, it is certain that the probability of intercept does not enjoy the property of a fixed number of linear segments. To see that this is not so, consider the case for three pulse trains where $T_1 = T_2$ but the ratio T_3/T_1 is irrational. Moreover,

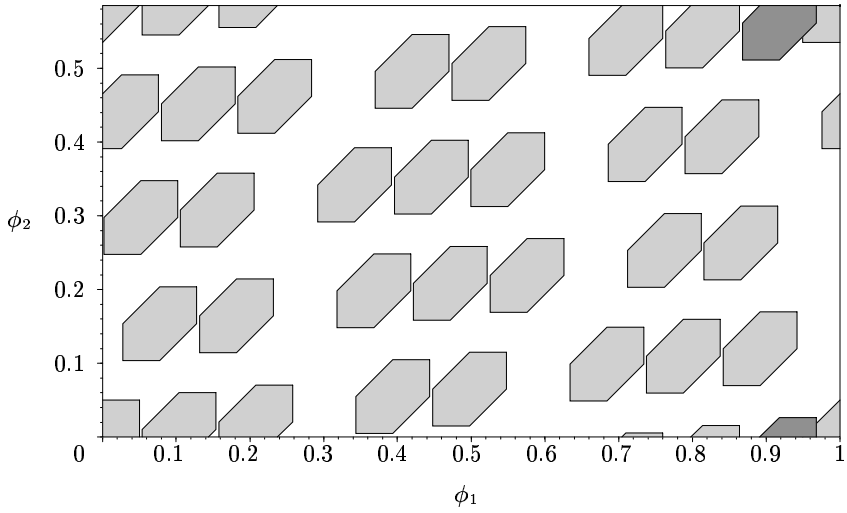


FIGURE 9. Phase space for calculation of probability of intercept for three pulse trains with $T_1 = 1$, $T_2 = 5^{-1/3}$, $T_3 = 5^{-2/3}$ and $\tau_1 = \tau_2 = \tau_3 = 0.05$.

suppose $\tau_3 = 0$ but $0 < \tau_1, \tau_2 < \frac{1}{2}T_1$. The sets $\mathcal{I}_{k_1,k_2,k_3}$ are now rectangles:

$$\mathcal{I}_{k_1,k_2,k_3} = \big\{\mathbf{x} \in \mathbb{R}^2 \;\big|\; |k_1T_1 + x_1 - k_3T_3| \leqslant \tau_1;\; |k_2T_1 + x_2 - k_3T_3| \leqslant \tau_2\big\}.$$

Now consider

$$\mathcal{P}_{N+1} - \mathcal{P}_N = \frac{1}{T_1T_2}\int_{NT_3-\frac{1}{2}T_1}^{NT_3+\frac{1}{2}T_1}\int_{NT_3-\frac{1}{2}T_1}^{NT_3+\frac{1}{2}T_1} \mathcal{C}_{N+1}(\boldsymbol{\phi}) - \mathcal{C}_N(\boldsymbol{\phi})\; d\phi_1\; d\phi_2.$$

The set $\mathcal{I}_{0,0,N}$ is the only set with index $k_3 = N$ which lies either wholly or partly within the limits of integration. The integral of the difference in the characteristic function measures that part of the set $\mathcal{I}_{0,0,N}$ which is not a part of any of the sets $\mathcal{I}_{k_1,k_2,k_3}$ with $0 \leqslant k_3 < N$. Suppose an overlap occurs with $\mathcal{I}_{k_1,k_2,k_3}$ for some k_1, k_2, k_3. Then

$$|k_1T_1 + (N - k_3)T_3| \leqslant 2\tau_1 \qquad \text{and} \qquad |k_2T_1 + (N - k_3)T_3| \leqslant 2\tau_2.$$

This implies that $k_1 = k_2$. Therefore, an overlap occurs whenever $|q\alpha - p| \leqslant \epsilon$, where $q = N - k_3$, $p = -k_1 = -k_2$, $\alpha = T_3/T_1$ and $\epsilon = \min\{\tau_1, \tau_2\}/T_1$. We know that α is irrational and $0 < q \leqslant N$. Therefore, it is impossible for $\mathcal{I}_{0,0,N}$ to overlap completely with an earlier set. Even if it overlaps partially with earlier sets, the overlaps cannot completely cover $\mathcal{I}_{0,0,N}$. This is because all overlapping sets must be offset diagonally from it by some non-zero amount which means that two (opposite) corners of $\mathcal{I}_{0,0,N}$ will remain exposed, regardless of N. Therefore, $\mathcal{P}_{N+1} - \mathcal{P}_N > 0$ for all $N \geqslant 0$. Since $\mathcal{P}_N \leqslant 1$, we conclude that there cannot be a finite number of linear segments in the expression for $\mathcal{P}_N$ in this case.

Now consider the continuous time problem. At any time instant t, a simultaneous coincidence occurs if the instances $\phi_1, \phi_2, \ldots, \phi_n$ of the random variables representing

the phases happen to satisfy the inequalities

$$|k_1T_1 + \phi_1 - t| \leqslant \tfrac{1}{2}\tau_1,\ |k_2T_2 + \phi_2 - t| \leqslant \tfrac{1}{2}\tau_2,\ \ldots,\ |k_nT_n + \phi_n - t| \leqslant \tfrac{1}{2}\tau_n.$$

Since the Φ_j are uniformly distributed over the range of their PRIs, the probability of intercept at any time instant is $(\tau_1\tau_2\cdots\tau_n)/(T_1T_2\cdots T_n)$. If we now define the set

$$\mathcal{I}_{k_1,k_2,\ldots,k_n}(t) = \left\{\mathbf{x} \in \mathbb{R}^n \ \middle|\ |k_iT_i + x_i - t| \leqslant \tfrac{1}{2}\tau_i;\ 1 \leqslant i \leqslant n\right\}$$

then an intercept occurs in the interval $[0, t]$ if $\boldsymbol{\phi} \in \mathcal{I}_{k_1,k_2,\ldots,k_n}(u)$ for any $u \in [0, t]$ or $k_1, k_2, \ldots, k_n \in \mathbb{Z}$. With $\mathcal{C}(\boldsymbol{\phi};\ t)$ defined to be the characteristic function of the union

$$\bigcup_{\substack{k_1,k_2,\ldots,k_n\in\mathbb{Z};\\ 0\leqslant u\leqslant t}} \mathcal{I}_{k_1,k_2,\ldots,k_n}(u),$$

the probability of intercept can be found from

$$\mathcal{P}(t) = \frac{1}{T_1T_2\cdots T_n}\int_{-T_n}^{0}\cdots\int_{-T_2}^{0}\int_{-T_1}^{0}\mathcal{C}(\phi_1, \phi_2, \ldots, \phi_n;\ t)\ d\phi_1\ d\phi_2\ \cdots\ d\phi_n. \tag{6.1}$$

Again, the Cartesian product of the integration intervals can be regarded as a hypertorus because of the periodicity of the characteristic function $\mathcal{C}(\boldsymbol{\phi};\ t)$ in its phase arguments and because the lengths of the intervals are equal to the periods. Geometrically, the probability is the proportion of the rectangular prism (the phase space) with sides of lengths $T_1, T_2, \ldots, T_n$ (with opposite faces identified) which is traced out by the movement of a rectangular prism with sides of length $\tau_1, \tau_2, \ldots, \tau_n$. The leading prism moves at a rate which is equal and positive along all axes. We can picture it as the obvious generalisation of Figure 4 to n dimensions. Eventually, the leading prism will overlap with the initial prism but, until this occurs, the rate of growth of the probability of intercept is equal to half the surface area of the leading prism. This is because, for any two opposite sides of the prism, only one is "exposed." Therefore, until overlap occurs, the probability of intercept is the sum of the volume of the prism with sides of length $\tau_1, \tau_2, \ldots, \tau_n$ plus the product of half its surface area with the observation time, normalised by the volume of the phase space. That is,

$$\mathcal{P}(t) = \frac{\tau_1\tau_2\cdots\tau_n(1 + t/\tau_1 + t/\tau_2 + \cdots + t/\tau_n)}{T_1T_2\cdots T_n} \tag{6.2}$$

for values of t which are sufficiently close to 0.

It is difficult to say more than this from a theoretical point of view. From a computational point of view, it seems unlikely that we could derive an algorithm which can efficiently calculate the probability of intercept for any given length of observation time or any other related statistics because of the apparent computational difficulties with best simultaneous Diophantine approximation. On the other hand, we can conceive of an algorithm that computes (6.1) by "monitoring" the progress

of the leading prism as it traverses phase space. The time required to compute the integral is then proportional to the length of the observation interval, which is unfortunate but perhaps unavoidable.

7. Other Approaches

In this chapter, we have developed a theory for predicting intercept times and calculating the probability of intercept for periodic pulse trains which is founded upon the theory of Diophantine approximation. For two pulse trains, this led to expressions for intercept time and probability of intercept which made use of the convergents of the simple continued fraction expansion of the PRI ratio, or upon the position of points in a Farey series of appropriate order. We found that the problem is much more difficult to treat both theoretically and computationally for more than two pulse trains.

It is therefore worthwhile to examine some other approaches to intercept time problems which have appeared in the literature. We divide these approaches into two categories: those which exploit linear congruence and those which replace the assumption of periodicity with stochastic behaviour. We will examine the approaches based on linear congruence first.

7.1. Exploitation of Linear Congruence. The properties of linear congruence were first exploited for the analysis of intercept time of periodic pulse trains by Miller & Schwarz (1953) and has been subsequently developed by Friedman (1954), Hawkes (1983) and Slocumb (1993).

For this approach, we assume the existence of some fundamental unit of time Δ, of which all the parameters of the problem are integer multiples. That is, we write

$$T_j = m_j\Delta, \qquad \tau_j = u_j\Delta \quad \text{and} \quad \phi_j = v_j\Delta$$

where $m_j, u_j, v_j \in \mathbb{Z}$ for $j = 1, 2, \ldots, n$ and n is the number of pulse trains. Assume for simplicity that all the u_j are odd and that the k^{th} pulse from the j^{th} pulse train is on whenever $kT_j + \phi_j - \frac{1}{2}(\tau_j - 1) \leqslant t = d\Delta \leqslant kT_j + \phi_j + \frac{1}{2}(\tau_j - 1)$. Simultaneous coincidences can then be found by solving the simultaneous linear congruences

$$k_i m_i \equiv v_i - v_j + w_i - w_j \pmod{m_j}$$

for $1 \leqslant i < j \leqslant n$ where w_i is allowed to be any integer in the range

$$-\tfrac{1}{2}(u_i - 1), \ldots, \tfrac{1}{2}(u_i - 1).$$

Some solutions can then be expressed in terms of the greatest common divisor or least common multiple of the periods. However, this approach is essentially a special case of the approach we have used throughout the chapter. The greatest common divisors and least common multiples are calculated using Euclid's algorithm and therefore has direct parallels with the process of obtaining the simple continued fraction expansion of real numbers.

7.2. Statistical Description of the Pulse Train. Since we quickly reach a point where we can say very little theoretically about the intercept times or the probability of intercept of multiple pulse trains as we increase the number of pulse trains involved and the exhaustive computation required to obtain accurate results may be unacceptable, some authors have sought to replace the assumption of strict periodicity of the pulse trains with a stochastic behaviour in the hope of obtaining a good approximate solution. Of course, in many physical systems it is perfectly valid to assume such behaviour in preference to periodicity.

The approach adopted by Stein & Johansen (1958) is to describe the frequency of pulses from a pulse train with pulse widths not exceeding a given value as a function of that value. That is, we associate a function $Q_j(x)$ with the j^{th} pulse train which represents the expected number of pulses per unit time from this pulse train with pulse widths $\tau \leqslant x$. The expected number of pulses per unit time, regardless of pulse width, is $\lim_{x\to\infty} Q_j(x)$. Obviously, $Q_j(x)$ is a non-decreasing function of x, defined for $x \geqslant 0$. From this, we obtain the FREQUENCY DENSITY FUNCTION

$$q_j(x) = \frac{d}{dx}\, Q_j(x).$$

Notice that this description of the pulse trains is not complete. No information is given about the interarrival times of pulses. Notwithstanding, Stein & Johansen are able to derive the frequency distribution function of the pulse train of coincidences, given only the frequency distributions of each pulse train, provided we make the important assumption that the occurrence of pulses any given pulse train is statistically independent of the occurrence of pulses from any other pulse train. The frequency distribution function of the coincidences can then be used to obtain the expected number of coincidences per unit time and the average duration of coincidences when they occur. Unfortunately, the assumption of statistical independence between the occurrence of pulses from different pulse trains means that the analysis is invalid for periodic pulse trains, for if the pulse trains are synchronised then the occurrence of a pulse from one pulse train determines to some extent the probability of the occurrence of pulses from other pulse trains, depending upon the phase relationships.

Nevertheless, Self & Smith (1985) apply the approach of Stein & Johansen to the prediction of intercept time in electronic warfare scenarios in which the pulse trains are often periodic. They derive an expression for the probability of intercept of a number of pulse trains over a prescribed observation interval. To do this, they again make use of the assumption of statistical independence of the occurrence of pulses between pulse trains at any time instant but extend it further by also assuming that occurrences in adjacent time intervals are also independent. That is, by breaking up the observation interval into a number of adjacent subintervals between each of which the probability of intercept is assumed to be independent,

the coincidence process becomes a Bernoulli process and the probability of intercept over the whole observation interval can be accumulated in the usual way. Taking the limit as the size of the component subintervals tends to zero, SELF & SMITH obtain the approximate expression

$$\mathcal{P}(t) = 1 - K\exp(-\lambda t) \tag{7.1}$$

for the probability of intercept, where

$$K = 1 - \mathcal{P}(0) = 1 - \frac{\tau_1\tau_2\cdots\tau_n}{T_1T_2\cdots T_n}$$

and

$$\lambda = \frac{\tau_1\tau_2\cdots\tau_n(1/\tau_1 + 1/\tau_2 + \cdots + 1/\tau_n)}{T_1T_2\cdots T_n}.$$

From our discussion in the previous section, we know that the probability of intercept of multiple periodic pulse trains is very difficult to determine. However, for short observation intervals, the probability of intercept is a linear function of the observation length, as described by (6.2). We observe that the expression (7.1) yields the same value for $\mathcal{P}(0)$ and, for small pulse widths, close agreement for $d\mathcal{P}(t)/dt$ at $t = 0$. Therefore, they are very similar for short observation intervals. This is demonstrated by SELF & SMITH in numerical simulations. Indeed, their simulations show the expression (6.2) to be superior, as we would expect. However, for long observation intervals both expressions for the probability of intercept are misleading and will give incorrect values for periodic pulse trains. Similarly, statistics derived from these expressions such as expected time to intercept may be very inaccurate. On the other hand, if the pulse trains are not strictly periodic but exhibit CUMULATIVE JITTER, which is to say the sequence of TOAs from the j^{th} pulse train is a random walk with a mean step length of T_j, then as the amount of jitter is increased, the expression (7.1) appears to become increasingly valid. KELLY *et al.* (1996) presents some numerical simulations in support of this.

We conclude by mentioning the similar work of DZIECH (1993). In his book, he develops an expression for the probability of intercept for stochastic pulse trains. He also requires a degree of independence between the occurrence of pulses between different pulse trains. The expression he derives requires that the statistics of the interarrival times of pulses from the pulse train of coincidences is known. In general, it would appear that this is a statistic that is rather difficult to obtain.

CHAPTER 6

PARAMETER ESTIMATION OF A PERIODIC PULSE TRAIN

1. Introduction

Periodic pulse trains are a common feature of many physical systems. In this chapter, we consider a situation in which a single periodic pulse train is observed and the times-of-arrival (TOAs) of pulses are measured. However, the pulses which are observed are not consecutive. It is assumed that some (and perhaps many) of the pulses were not observed. Additionally, we assume that TOAs are not measured perfectly, but are subject to random errors. The problem is to estimate the period of the pulse train from the data recorded and associate each of the measured TOAs with a pulse number or index, relative to the first observed, which takes account of the intervening missing pulses.

The motivating problem in this instance is passive radar surveillance. Typically, a radar emits a train of pulses in a periodic sequence. In the simplest and most frequently encountered case, the pulse train is purely periodic (the sequence has a length of one). Receivers for passive radar surveillance can often make use of the period or pulse repetition interval (PRI) to identify the emitter and its mode of operation. However, for many conceivable reasons, it may not be possible or desirable to measure the TOA of each consecutive pulse. It may not be possible because the signal strength of received pulses varies or because pulses from different sources overlap, and it may not be desirable because of the need to maintain surveillance over a range of parameters, such as angle-of-arrival or carrier frequency, which is wider than the receiver is capable of at any one time. Therefore, pulses may be missing from the record. For certain receiver types, for example a scanning superheterodyne receiver, the record of pulses from a given pulse train may be extremely sparse. Additionally, measurement of the TOA will be subject to a variety of errors, such as thresholding effects caused by thermal noise or variability in received power or simply poor time resolution.

Very little has been published regarding the problem of estimating the period of a pulse train from sparse, noisy measurements. Indeed, the papers of Casey & Sadler (1995, 1996) are the only works in the open literature of which the author is aware. In those papers, a number of generalised Euclidean algorithms are proposed to recover the period. They demonstrate that, even for a very sparse record in which 99% of pulses are missing, the period can be reliably estimated. We have

identified a number of areas in which their results can be improved. Firstly, we formulate statistical models for the measurement process. We make use of a fairly standard multi-dimensional Euclidean algorithm, derived from the LLL algorithm of Lenstra *et al.* (1982) for lattice reduction (see Chapter 3), and explain its relationship to a method of maximum likelihood for estimation and association. Furthermore, the algorithm we propose is capable of reliably estimating the PRI and associating pulse indices for extremely sparse, noisy and short records.

In Section 2, we introduce two statistical models for the measurement process: a simple model and an extended model. The simple model assumes very little prior information about the way in which pulses are missing from the record, while the extended model assumes that the indices of the observed pulses are random variables with a known distribution.

We discuss the method of maximum likelihood estimation of the PRI and phase (time offset of the first observed pulse from an arbitrary but fixed time origin) for the simple model in Section 3. For the extended model, we also introduce the problem of *joint maximum likelihood estimation and association* (JMLEA) of the PRI, phase and pulse indices.

In Section 4, we show how the estimation and association problems can be formulated as a problem of simultaneous Diophantine approximation (see Chapter 4). We use the theory of simultaneous Diophantine approximation to conclude that, in general, no maximum likelihood estimates of PRI or phase exist for the simple model. We also prove the existence of a JMLEA for the extended model, and discuss conditions under which the JMLEA corresponds to a best approximation in the relative sense in this formulation. The algorithm we propose for obtaining a JMLEA, an adaptation LLL algorithm for simultaneous Diophantine approximation, is discussed in detail in Section 5.

A relationship with the maximisation of a certain trigonometric sum is elucidated in Section 6. The trigonometric sum we examine can be regarded as a periodogram of impulses at the observation times. This allows us to interpret the behaviour of our algorithm in the frequency domain. We discover that the algorithm offers an efficient means of finding peaks in the periodogram.

Finally, we present some numerical results in Section 7. We find that our algorithm is able to make correct associations (and hence obtain statistically efficient estimates) with experimental frequencies in excess of 99%, even when the number of expected number of missing pulses is 99.9%, the noise variance on the TOA measurements, as a proportion of the PRI, is as high as 0.01 and as few as 9 pulses are observed. We also present some numerical evidence which suggests that the algorithm is robust to uncertainties in the known parameters. Finally, we illustrate the behaviour of the algorithm in the frequency domain.

2. Signal Model

We will consider two signal models: a simple model and an extended model.

Consider a purely periodic pulse train with pulse repetition interval T and phase θ. Under this model, pulses are emitted at the times $iT + \theta$, $i \in \mathbb{Z}$. We assume that our record consists of n observations of pulses which are corrupted by noise and which are not necessarily consecutive. Thus, our observations are of random variables, which we shall denote $Z_1, Z_2, \ldots, Z_n$, such that

$$Z_i = s_i T + \theta + X_i$$

where the $s_i \in \mathbb{Z}$ are the indices of the observed pulses and the X_i are independent, identically distributed (i.i.d.) normal random variables representing the observation errors (noise) with zero mean and variance σ^2. We assume that the only ADMISSIBLE pulse indices are those which satisfy

$$0 = s_1 < s_2 < \cdots < s_n. \tag{2.1}$$

We will call this model the SIMPLE model. An example of a received pulse train generated by this model is illustrated in Figure 1. The "true" pulse train is depicted

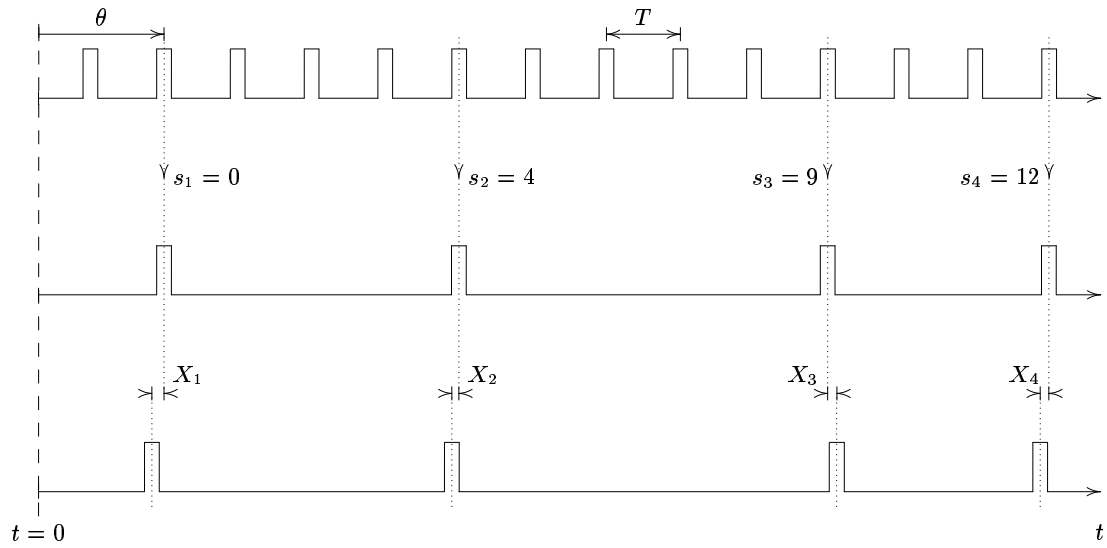


FIGURE 1. A record of pulses generated by the simple model.

at top. Of the pulses in the true pulse train, only four are recorded, with pulse indices $s_1 = 0$, $s_2 = 4$, $s_3 = 9$ and $s_4 = 12$. The pulse train which would have been recorded in the absence of measurement errors is depicted at centre. A representation of the pulses with noisy TOAs are shown at bottom.

We will find that, in general, no maximum likelihood estimate of the parameters exists in the simple model. We require more prior information to obtain such an estimate. For this reason, we propose an EXTENDED model, in which we will also assume that the indices of the observed pulses are random variables. We assume

that our TOA observations $Z_1, Z_2, \ldots, Z_n$ now have the form

$$Z_i = S_i T + \theta + X_i$$

where $S_1 = 0$ (a degenerate random variable),

$$S_{i+1} - S_i = Y_i + 1$$

for $i > 1$ and the Y_i are i.i.d. random variables from a geometric distribution with parameter λ. Furthermore, the X_i and the Y_i are assumed to be mutually independent. The parameter λ can be interpreted as the probability of a given pulse being observed.

The assignment of a distribution to the observed indices can be justified on physical grounds for certain receiver types. Consider a sensitive receiver which must search for signals in some parameter space (for example a rotating, receiver searching in angle and carrier frequency) and suppose its search strategy is fixed. It is reasonable to suggest that we might analyse the search strategy to discover relative frequencies with which pulses are observed or missed *a priori*. If, in addition, the strategy has a random or pseudo-random component, then it makes sense to consider the observed pulse indices as random variables. Of course, it is unlikely that this distribution will be completely independent of the PRI or that the differences in observed indices arise from a geometric distribution. Nevertheless, we believe that our extended model, while still simplistic, is representative and serves to demonstrate what is possible.

3. Parameter Estimation and Association

Consider the method of maximum likelihood for estimation of the parameters T, θ and $\mathbf{s}$ in the simple model. The joint probability density function (p.d.f.) of the observations is

$$f\big(\mathbf{z};\ T, \theta, \sigma^2, \mathbf{s}\big) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\|\mathbf{z} - T\mathbf{s} - \theta\mathbf{1}\|_2^2}{2\sigma^2}\right) \tag{3.1}$$

where $\mathbf{z}$ is a column vector representing the possible values of the Z_i and $\mathbf{s}$ is a column vector representing the pulse indices, which are assumed to be admissible according to (2.1). As usual, $\|\cdot\|_2$ denotes the Euclidean norm.

If the vector of pulse indices, $\mathbf{s}$, is known *a priori* then the problem is simply one of linear regression to estimate T and θ. Recall (*e.g.* Hogg & Craig, 1978) that the maximum likelihood estimates for θ and T are

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\Big(z_i - \hat{T}s_i\Big)$$

and

$$\hat{T} = \frac{n\sum_{i=1}^n z_i s_i - \left(\sum_{i=1}^n z_i\right)\left(\sum_{i=1}^n s_i\right)}{n\sum_{i=1}^n s_i^2 - \left(\sum_{i=1}^n s_i\right)^2}.$$

If we define

$$\mathbf{Q} = \mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}} \tag{3.2}$$

then it can be shown that the estimates $\hat{\theta}$ and $\hat{T}$ can be equivalently expressed as

$$\hat{\theta} = \frac{1}{n}\Big(\mathbf{z} - \hat{T}\mathbf{s}\Big)^T\mathbf{1} \tag{3.3}$$

and

$$\hat{T} = \frac{\mathbf{z}^T\mathbf{Q}\mathbf{s}}{\mathbf{s}^T\mathbf{Q}\mathbf{s}}. \tag{3.4}$$

Notice that $\mathbf{Q}$ is a symmetric PROJECTION MATRIX, by which we mean that $\mathbf{Q}^2 = \mathbf{Q}$. Furthermore, if we write the projections of $\mathbf{s}$ and $\mathbf{z}$ with respect to $\mathbf{Q}$ as

$$\mathbf{x} = \mathbf{Q}\mathbf{s} \qquad \text{and} \qquad \boldsymbol{\zeta} = \mathbf{Q}\mathbf{z} \tag{3.5}$$

then

$$\hat{T} = \frac{\boldsymbol{\zeta}^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

With our likelihood function being the joint p.d.f. in (3.1) and with $\mathbf{z}$ set to the observed values of the Z_i, we find that the likelihood function is maximised with respect to T and θ at $\hat{T}$ and $\hat{\theta}$, respectively. Thus, taking the logarithm of the likelihood function and discarding constants, we find that the likelihood function is maximised with respect to $\mathbf{s}$, T and θ when

$$F(\mathbf{s}) = \boldsymbol{\zeta}^T\boldsymbol{\zeta} - \frac{\big(\boldsymbol{\zeta}^T\mathbf{x}\big)^2}{\mathbf{x}^T\mathbf{x}} \tag{3.6}$$

is minimised, where $\mathbf{x}$ is of course a function of $\mathbf{s}$ as defined by (3.5). Now, $F(\mathbf{s}) \geqslant 0$. We also note from the form of (3.6) that maximisation of the likelihood function is equivalent to minimisation of $\sin^2\phi$ over all admissible $\mathbf{s}$, where ϕ is the angle between the vectors $\boldsymbol{\zeta}$ and $\mathbf{x}$.

We will show in Section 4 that a unique maximum likelihood estimate does not exist in general, so we now consider what can be done using the extended model. In this model, we find that the joint p.d.f. is

$$g\big(\mathbf{z}, \mathbf{s};\ T, \theta, \sigma^2, \lambda\big) = \lambda^{n-1}(1-\lambda)^{s_n - n + 1} f\big(\mathbf{z};\ T, \theta, \sigma^2, \mathbf{s}\big) \tag{3.7}$$

when $\mathbf{s} \in \mathbb{Z}^n$ (now a column vector, the entries of which represent the possible values of the S_i) is admissible according to (2.1) and the joint p.d.f. is 0 otherwise. If we were interested only in finding maximum likelihood estimates for T and θ then we should maximise the likelihood function

$$L(T, \theta) = \sum_{\mathbf{s} \in \mathbb{Z}^n} g\big(\mathbf{z}, \mathbf{s};\ T, \theta, \sigma^2, \lambda\big)$$

with respect to T and θ to obtain our estimates $\hat{T}$ and $\hat{\theta}$. However, suppose we want to simultaneously associate the observations with a set of pulse indices $\hat{\mathbf{s}}$. Then,

in order to make our observations maximally likely, we should simply maximise g over $\mathbf{s}$, T and θ. We call this JOINT MAXIMUM LIKELIHOOD ESTIMATION AND ASSOCIATION (JMLEA).

For any postulated association $\mathbf{s}$, the maximum likelihood estimates for T and θ remain as they were in (3.4) and (3.3), respectively. So, by taking the logarithm of g and discarding constant factors, we find that the JMLEA is obtained with respect to $\mathbf{s}$, T and θ when the function

$$G(\mathbf{s}) = F(\mathbf{s}) + \kappa s_n \tag{3.8}$$

is minimised where

$$\kappa = -2\sigma^2 \log(1-\lambda) > 0. \tag{3.9}$$

Clearly, $\kappa \approx 2\sigma^2\lambda$ when λ is small.

We will show that a JMLEA exists for the extended model, as does the maximum likelihood estimate of T and θ. We will show in the next section that an algorithm for simultaneous Diophantine approximation will usually furnish the JMLEA if the amount of measurement error is sufficiently small.

4. Formulation as a Simultaneous Diophantine Approximation Problem

With the problem of maximum likelihood estimation stated in terms of minimising the function F in (3.6) or G in (3.8), we will show that the problem can be stated as a problem of finding best simultaneous Diophantine approximations in the relative sense.

In Chapter 4, we defined the notion of a system for simultaneous Diophantine approximation (Definition 2.6) and we defined a best approximation in the absolute sense (Definition 2.7) with respect to the radius and height functions of the system. We can define a best approximation in the relative sense in the following way.

DEFINITION 4.1. A lattice point $\mathbf{x}$ in a lattice Ω is a BEST APPROXIMATION IN THE RELATIVE SENSE with respect to a radius function, ρ, and a height function, h, if h is a semi-norm, $h(\mathbf{x}) > 0$ and if, for all $\mathbf{y} \in \Omega$ with $h(\mathbf{y}) > 0$ it is true that

$$\frac{\rho(\mathbf{y})}{h(\mathbf{y})} \leqslant \frac{\rho(\mathbf{x})}{h(\mathbf{x})} \;\Rightarrow\; h(\mathbf{y}) \geqslant h(\mathbf{x})$$

and

$$h(\mathbf{y}) \leqslant h(\mathbf{x}) \;\Rightarrow\; \frac{\rho(\mathbf{y})}{h(\mathbf{y})} \geqslant \frac{\rho(\mathbf{x})}{h(\mathbf{x})}.$$

If we set $\Omega = \mathbb{Z}^2$ and, writing $\mathbf{v} \in \Omega$ as $\mathbf{v} = (p, q)$, we also set $\rho(\mathbf{v}) = |q\alpha - p|$ and $h(\mathbf{v}) = |q|$ then we have a system for simultaneous Diophantine approximation since ρ and h are transverse semi-norms. Moreover, we can see that our definition of a best simultaneous Diophantine approximation in the relative sense can be reduced

to the ordinary definition of a best Diophantine approximation in the relative sense, given in Chapter 2.

The following is a trivial adaptation of a theorem of LAGARIAS (1983).

THEOREM 4.1. *Suppose Ω is a lattice in $\mathbb{R}^m$, ρ is a radius function and h is a height function which together form a simultaneous Diophantine approximation system. If h is a semi-norm then there is a set of all best approximations in the relative sense which can be numbered $\{\mathbf{v}_j\}$ for $b_0 \leqslant j \leqslant b_1$ such that for all $b_0 \leqslant j < k \leqslant b_1$*

$$h(\mathbf{v}_j) \leqslant h(\mathbf{v}_k) \qquad \text{and} \qquad \frac{\rho(\mathbf{v}_j)}{h(\mathbf{v}_j)} \geqslant \frac{\rho(\mathbf{v}_k)}{h(\mathbf{v}_k)}.$$

Furthermore, $b_1 = \infty$ if and only if the restriction of ρ to the real span of Ω is not an extended norm and there exists no $\mathbf{x} \in \Omega$ such that $\rho(\mathbf{x}) = 0$. If $b_1 < \infty$ then $\rho(\mathbf{v}_{b_1}) = \mathbf{0}$. Otherwise,

$$\lim_{j\to\infty} \rho(\mathbf{v}_j) = 0 \qquad \text{and} \qquad \lim_{j\to\infty} h(\mathbf{v}_j) = \infty.$$

We will now set up our estimation and association problems as problems of best simultaneous Diophantine approximation. Consider the simple model in which we hope to find maximum likelihood estimates for T and θ. Let $\mathbf{q}_i$ denote the i^{th} column of $\mathbf{Q}$ as defined in (3.2). Let $\{\mathbf{q}_2, \mathbf{q}_3, \ldots, \mathbf{q}_n\}$ be a basis of a lattice Ω in $\mathbb{R}^n$. Observe that $\mathbf{q}_1 \in \Omega$. Also, any point $\mathbf{x}$ in the lattice Ω can be uniquely expressed $\mathbf{x} = \mathbf{Q}\mathbf{s}$, where $\mathbf{s} \in \mathbb{Z}^n$ with $s_1 = 0$. We define the radius and height functions for $\mathbf{x} \in \mathbb{R}^2$ as

$$\rho(\mathbf{x}) = \|\mathbf{M}\mathbf{x}\|_2 \qquad \text{and} \qquad h(\mathbf{x}) = \|\mathbf{x}\|_2 \tag{4.1}$$

where

$$\mathbf{M} = \mathbf{I}_n - \frac{\boldsymbol{\zeta}\boldsymbol{\zeta}^T}{\boldsymbol{\zeta}^T\boldsymbol{\zeta}}$$

and $\boldsymbol{\zeta}$ is the projection of the TOA measurement vector $\mathbf{z}$ by $\mathbf{Q}$ as described by (3.5). Notice that $\mathbf{M}$ is a symmetric projection matrix like $\mathbf{Q}$. Furthermore, $\mathbf{Q}$ projects along the line $\mathbb{R}\mathbf{1}$ onto an orthogonal hyperplane, $\mathcal{E}$, and $\Omega \subset \mathcal{E}$ and $\boldsymbol{\zeta} \in \mathcal{E}$. $\mathbf{M}$ projects along the line $\mathbb{R}\boldsymbol{\zeta}$. Now, ρ and h are transverse because h is a norm. It should also be clear that the restriction of ρ to $\mathcal{E}$, the real span of Ω, is not a norm.

It is not hard to verify that we can now rewrite the log-likelihood function F from (3.6) as

$$F(\mathbf{s}) = h(\boldsymbol{\zeta})^2 \left[\frac{\rho(\mathbf{x})}{h(\mathbf{x})}\right]^2.$$

where, again, $\mathbf{x} = \mathbf{Q}\mathbf{s}$. Since $\mathbf{x} \in \Omega$, minimisation of F over admissible index vectors is equivalent to minimisation of

$$F^*(\mathbf{v}) = \frac{\rho(\mathbf{v})}{h(\mathbf{v})}$$

over all $\mathbf{v} \in \Omega$ which correspond to admissible index vectors.

Theorem 4.1 implies that either there exists some non-zero lattice point $\mathbf{x}$ such that $F^*(\mathbf{x}) = 0$ or there exists a non-terminating sequence of lattice points $\mathbf{v}_j$ such that

$$\lim_{j\to\infty} F^*(\mathbf{v}_j) = 0 \qquad \text{and} \qquad \lim_{j\to\infty} h(\mathbf{v}_j) = \infty.$$

Since $F^*(\mathbf{x}) = 0$ implies that $F^*(k\mathbf{x}) = 0$ for all $k \in \mathbb{Z}$, we can now see that either there is no non-zero lattice points which minimises $F^*(\cdot)$ or an infinitude. We have neglected to consider the necessity that the lattice points be admissible to our model, as defined by (2.1). We have only taken care to ensure that $s_1 = 0$. However, if the observations are TIME-ORDERED, which is to say that they satisfy

$$z_1 < z_2 < \cdots < z_n, \tag{4.2}$$

then we can conclude that there will be no maximum likelihood estimate. We can expect the observations to be time-ordered with very high probability if the spacing between the indices is large compared with σ. Of course, in practice, it is very likely that observations will be made and recorded in time-order.

Therefore, consider the extended model in which we impose a "cost" on the number of pulses missing from the record. Here, we set out to jointly estimate T and θ and associate the observations with a set of pulse indices, $\mathbf{s}$, so as to maximise the likelihood of our observations. As we described in Section 3, this involves the minimisation of $G(\mathbf{s})$ in (3.8) over all admissible $\mathbf{s}$.

We can show that a JMLEA must exist and there are at most a finite number of them. This is because the number of admissible index vectors $\mathbf{s}$ with $G(\mathbf{s}) \leqslant \nu$ for any $\nu \in \mathbb{R}$ must be finite. In turn, this is because $F(\mathbf{s}) \geqslant 0$, $\kappa s_n > 0$ and the number of admissible index vectors $\mathbf{s}$ with $s_n \leqslant \nu/\kappa$ is finite.

We will state a theorem which suggests that if the radius of the index vector associated with the JMLEA is sufficiently small then it must be a best simultaneous Diophantine approximation in the relative sense. However, we require the following lemmata.

LEMMA 4.1. *Consider a vector* $\mathbf{v} \in \mathbb{R}^n$. *Let* $\mathbf{Q} \in \mathbb{R}^{n\times n}$ *be defined as in* (3.2). *If* $v_1 = 0$ *then*

$$v_n^2 \leqslant 2\,\|\mathbf{Q}\mathbf{v}\|_2^2\,.$$

PROOF. Let $C(\mathbf{v}) = \|\mathbf{Q}\mathbf{v}\|_2^2$. We have

$$C(\mathbf{v}) = \sum_{i=1}^{n} v_i^2 + \frac{1}{n}\left(\sum_{i=1}^{n} v_i\right)^2.$$

Thus,

$$\frac{\partial C}{\partial v_j} = 2v_j - \frac{2}{n}\sum_{i=1}^{n} v_i.$$

For a fixed v_n and $v_1 = 0$, the extremal values of $C(\mathbf{v})$ occur when

$$\frac{\partial C}{\partial v_2} = \frac{\partial C}{\partial v_3} = \cdots = \frac{\partial C}{\partial v_{n-1}} = 0.$$

This is satisfied when

$$v_2 = v_3 = \cdots = v_{n-1} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

which implies that

$$v_2 = v_3 = \cdots = v_{n-1} = \tfrac{1}{2}v_n.$$

The value of C at this point is $C(\mathbf{v}) = \frac{1}{2}v_n^2$. Now, the second order partial differentials are

$$\frac{\partial C}{\partial v_i v_j} = \begin{cases} 2(n-1)/n & \text{if } i = j, \\ -2/n & \text{otherwise.} \end{cases}$$

The matrix of these partial differentials is $2\mathbf{Q}$ and, since $\mathbf{Q}$ is positive semi-definite, we conclude that $C(\mathbf{v}) \geqslant \frac{1}{2}v_n^2$ when $v_1 = 0$ and v_n is fixed. □

LEMMA 4.2. *Suppose $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are two sequences of real numbers for which*

$$x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n, \qquad x_1 + x_2 + \cdots + x_n = 0,$$
$$y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n, \qquad y_1 + y_2 + \cdots + y_n = 0.$$

If $x_1 < 0$ and $y_1 < 0$ then

$$x_1y_1 + x_2y_2 + \cdots + x_ny_n > 0.$$

PROOF. Let

$$d_i = x_i - x_{i-1}.$$

Now, $d_i \geqslant 0$ for each $1 < i \leqslant n$ and there exists at least one value for i such that $d_i > 0$. Let

$$S_i = y_1 + y_2 + \cdots + y_i.$$

We see that $S_i < 0$ for all $1 \leqslant i < n$ and $S_n = 0$. The proof follows from the identity

$$\sum_{i=1}^{n} x_iy_i = x_nS_n - \sum_{i=2}^{n} d_iS_{i-1}.$$

□

Note that the above identity is the obvious adaptation to sequences of the formula for integration by parts.

COROLLARY 4.1. *If* $\mathbf{s}$ *is an admissible index vector in the sense of* (2.1) *and* $\mathbf{z}$ *is a time-ordered set of observations in the sense of* (4.2) *then* $\mathbf{z}^T\mathbf{Q}\mathbf{s} > 0$.

PROOF. The proof follows by noticing that the vectors $\mathbf{Q}\mathbf{s}$ and $\mathbf{Q}\mathbf{z}$ both satisfy the conditions of Lemma 4.2. □

THEOREM 4.2. *Consider the simultaneous Diophantine approximation system consisting of the lattice* $\Omega = \mathbf{Q}\mathbb{Z}^n$ *and the radius function* ρ *and height function* h *as defined by* (4.1). *Suppose* $\hat{\mathbf{s}}$ *is that integer vector with* $s_1 = 0$ *and* $s_n > 0$ *which minimises* $G(\mathbf{s})$ *for the time-ordered observations* $\mathbf{z}$, *where* $G(\mathbf{s})$ *is defined in* (3.8). *If* $\hat{\mathbf{s}}$ *is admissible and*

$$\rho(\hat{\mathbf{x}})^2 < \tfrac{1}{10} \tag{4.3}$$

where $\hat{\mathbf{x}} = \mathbf{Q}\hat{\mathbf{s}}$ *then* $\hat{\mathbf{x}}$ *is a best simultaneous Diophantine approximation in the relative sense for this system.*

PROOF. Firstly, we observe that there exists an integer vector that minimises $G(\mathbf{s})$ with $s_1 = 0$ and $s_n > 0$, for Theorem 4.1 implies that if we continue to decrease the ratio $\rho(\mathbf{x})/h(\mathbf{x})$ then $\rho(\mathbf{x}) \to 0$, where $\mathbf{x} = \mathbf{Q}\mathbf{s}$. But this implies that $s_n \to \infty$. Therefore, we conclude that a minimum must exist.

Suppose $\hat{\mathbf{s}}$ is not a best approximation in the relative sense for our system. Then there exists some $\mathbf{s}^*$ with $s_1^* = 0$ and $s_n^* > 0$ such that

$$\rho(\mathbf{Q}\mathbf{s}^*) \leqslant \rho(\mathbf{Q}\hat{\mathbf{s}}) \qquad \text{and} \qquad h(\mathbf{Q}\mathbf{s}^*) \leqslant h(\mathbf{Q}\hat{\mathbf{s}})$$

but $G(\mathbf{s}^*) > G(\hat{\mathbf{s}})$. From this we must conclude that

$$s_n^* > \hat{s}_n. \tag{4.4}$$

Consider the decomposition of a vector $\mathbf{s} \in \mathbb{R}^n$ into orthogonal components. Let us write

$$\mathbf{s} = \lambda\mathbf{1} + \mu\boldsymbol{\zeta} + \mathbf{c}$$

where

$$\lambda\mathbf{1} = \mathbf{s} - \mathbf{Q}\mathbf{s}, \qquad \mu\boldsymbol{\zeta} = \mathbf{Q}\mathbf{s} - \mathbf{M}\mathbf{Q}\mathbf{s} \qquad \text{and} \qquad \mathbf{c} = \mathbf{M}\mathbf{Q}\mathbf{s}$$

and $\lambda, \mu \in \mathbb{R}$. For such a decomposition, we have $\rho(\mathbf{Q}\mathbf{s}) = \rho(\mathbf{c})$ and $h(\mathbf{Q}\mathbf{s}) = h(\mu\boldsymbol{\zeta} + \mathbf{c})$. Now, we decompose $\hat{\mathbf{s}}$ and $\mathbf{s}^*$ in this manner so that

$$\hat{\mathbf{s}} = \hat{\lambda}\mathbf{1} + \hat{\mu}\boldsymbol{\zeta} + \hat{\mathbf{c}} \qquad \text{and} \qquad \mathbf{s}^* = \lambda^*\mathbf{1} + \mu^*\boldsymbol{\zeta} + \mathbf{c}^*.$$

Corollary 4.1 implies that $\hat{\mu} > 0$ because $\hat{\mathbf{s}}$ is admissible and $\mathbf{z}$ is time-ordered.

Consider how much larger s_n^* can be than $\hat{s}_n$. We have

$$\|\hat{\mathbf{c}}\|_2 = \rho(\mathbf{Q}\hat{\mathbf{s}}) \geqslant \rho(\mathbf{Q}\mathbf{s}^*) = \|\mathbf{c}^*\|_2 = \|(\mathbf{c}^* - \hat{\mathbf{c}}) + \hat{\mathbf{c}}\|_2 \geqslant \|\mathbf{c}^* - \hat{\mathbf{c}}\|_2 - \|\hat{\mathbf{c}}\|_2 .$$

Therefore,

$$\|\mathbf{c}^* - \hat{\mathbf{c}}\|_2 \leqslant 2\,\|\hat{\mathbf{c}}\|_2 = 2\rho(\mathbf{Q}\hat{\mathbf{s}}). \tag{4.5}$$

By making use of the orthogonality of $\boldsymbol{\zeta}$ and $\mathbf{c}^*$ we find that

$$h(\mathbf{Q}\hat{\mathbf{s}})^2 \geqslant h(\mathbf{Q}\mathbf{s}^*)^2 = \|\mu^*\boldsymbol{\zeta} + \hat{\mathbf{c}}\|_2^2 = \|\mu^*\boldsymbol{\zeta}\|_2^2 + \|\hat{\mathbf{c}}\|_2^2 \geqslant \|\mu^*\boldsymbol{\zeta}\|_2^2 \tag{4.6}$$

Suppose $\mu^* > \hat{\mu}$. Then

$$\|\mu^*\boldsymbol{\zeta}\|_2^2 = [\|(\mu^* - \hat{\mu})\boldsymbol{\zeta}\|_2 + \|\hat{\mu}\boldsymbol{\zeta}\|_2]^2 \geqslant \|(\mu^* - \hat{\mu})\boldsymbol{\zeta}\|_2^2 + \|\hat{\mu}\boldsymbol{\zeta}\|_2^2 . \tag{4.7}$$

Now $\|\hat{\mu}\boldsymbol{\zeta}\|_2^2 = h(\mathbf{Q}\hat{\mathbf{s}})^2 - \rho(\mathbf{Q}\hat{\mathbf{s}})^2$. Therefore, (4.6) and (4.7) imply that

$$\|(\mu^* - \hat{\mu})\boldsymbol{\zeta}\|_2^2 \leqslant \rho(\mathbf{Q}\hat{\mathbf{s}})^2. \tag{4.8}$$

Now, from Lemma 4.1, we know that

$$\begin{aligned}(s_n^* - \hat{s}_n)^2 &\leqslant 2\,\|\mathbf{Q}(\mathbf{s}^* - \hat{\mathbf{s}})\|_2^2 \\ &= 2\,\|(\mu^* - \hat{\mu})\boldsymbol{\zeta} + \mathbf{c}^* - \hat{\mathbf{c}}\|_2^2 \\ &= 2\,\|(\mu^* - \hat{\mu})\boldsymbol{\zeta}\|_2^2 + 2\,\|\mathbf{c}^* - \hat{\mathbf{c}}\|_2^2\end{aligned}$$

If $\mu^* > \hat{\mu}$ then we can use (4.5) and (4.8) to show that

$$(s_n^* - \hat{s}_n)^2 \leqslant 10\rho(\mathbf{Q}\hat{\mathbf{s}})^2. \tag{4.9}$$

Thus, if $\rho(\mathbf{Q}\hat{\mathbf{s}})^2 < \frac{1}{10}$ then the difference between the indices must be less than one. Since the indices are integers, they must be equal, contradicting (4.4).

Suppose instead that $\mu^* \leqslant \hat{\mu}$. Noticing that $\zeta_n > 0$ because $\mathbf{z}$ is time-ordered, we find that

$$s_n^* - \hat{s}_n = (\mu^* - \hat{\mu})\zeta_n + c_n^* - \hat{c}_n \leqslant c_n^* - \hat{c}_n.$$

In this case, we find that the coefficient on the right hand side of (4.9) can be reduced from 10 to 8, again furnishing a contradiction with (4.4) if $\rho(\mathbf{Q}\hat{\mathbf{s}})^2 < \frac{1}{10}$. □

We remark that this theorem is not as strong as we would like. It does not state that the JMLEA must be a best simultaneous Diophantine approximation for our system, even if it satisfies (4.3). This will be true only if it also happens to be that index vector which minimises $G(\mathbf{s})$ when the condition that $\mathbf{s}$ is admissible is replaced by the slightly weaker conditions that $s_1 = 0$ and $s_n > 0$. Computationally, it appears difficult to verify that a candidate value of $\mathbf{s}$ truly minimises $G(\mathbf{s})$ with respect to these conditions. Nevertheless, it seems reasonable to believe that, for sufficiently small noise variance, the conditions of the theorem statement will be met very frequently and the JMLEA will be a best approximation.

5. An Algorithm For Estimation and Association

We have shown in the previous section that, when the noise variance is small, the JMLEA in the extended model is likely to be a best approximation in the relative sense with respect to the simultaneous Diophantine approximation system consisting of the lattice $\Omega = \mathbf{Q}\mathbb{Z}^n$, the radius function ρ and the height function

h, defined in (4.1). As we discussed in Chapter 4, no computationally efficient algorithm is known which can be guaranteed of finding an uninterrupted sequence of best approximations for lattices of rank higher than two. Therefore, if we are to obtain estimates and associations in a reasonable amount of time, we must turn to sub-optimal methods, and hope that these methods are good enough to find the particular best approximation we require most of the time, and that these approximations correspond to the JMLEA.

Underlying the method is the LLL algorithm of Lenstra *et al.* for lattice reduction, which we discussed in Chapter 3. In Chapter 4, we discussed several variants of the LLL algorithm for simultaneous Diophantine approximation, including the HJLS algorithm and the PSLQ algorithm. However, we have chosen to implement our own variation, largely because of the convenience and reliability of the implementations of the LLL algorithm in the LiDIA library (LiDIA Group, 1995).

Our aim is to calculate a basis of Ω which contains small vectors with respect to a certain norm on each iteration. The norm is adjusted on each iteration in order to obtain a sequence of good approximations, with the hope that this sequence includes the lattice point associated with the JMLEA. On the k^{th} iteration, we use the norm

$$\|\mathbf{v}\|^{(k)} = \big(1-\gamma^k\big)\rho(\mathbf{v}) + \gamma^k h(\mathbf{v}) \tag{5.1}$$

where $0 < \gamma < 1$ is an adjustable constant which, as γ is reduced, increases the speed of the algorithm at the expense of missing some good approximations (which may include the JMLEA). If we can be assured that the vectors in our basis are within some constant factor of the smallest vectors with respect to the norm $\|\cdot\|^{(k)}$ then, with each successive iteration of our algorithm, the basis will tend to contain vectors with smaller and smaller radius, but with larger and larger height. If we are "lucky" enough that the basis we calculate contains the shortest vector with respect to $\|\cdot\|^{(k)}$ then we have found a best approximation in the relative sense. From Lagarias (1982), we know that the size of successive best approximations in the absolute sense grow exponentially in height. Although we don't know this for best approximations in the relative sense, it seems reasonable to adopt the exponential form of (5.1).

We now discuss how the LLL algorithm is modified for our purpose. Notice that $\|\mathbf{v}\|^{(k)}$ can be equivalently expressed as

$$\|\mathbf{v}\|^{(k)} = \big\|\mathbf{L}^k\mathbf{v}\big\|_2$$

where

$$\mathbf{L} = \gamma\mathbf{I} + (1-\gamma)\mathbf{M}.$$

Now, $\mathbf{L}$ is invertible if $\gamma > 0$ and the inverse is

$$\mathbf{L}^{-1} = \gamma^{-1}\mathbf{I} + \big(1-\gamma^{-1}\big)\mathbf{M},$$

as can be readily checked. Therefore, if we Lovász-reduce a basis of the lattice $\mathbf{L}^k\mathbf{Q}\mathbb{Z}^n$ then, by applying the inverse transformation $\mathbf{L}^{-k}$ to the reduced basis matrix, we have a basis of Ω for which we can be sure that each element is within a factor of $2^{(n-1)/2}$ of the size of the corresponding element in a Minkowski-reduced basis with respect to $\|\cdot\|^{(k)}$. In this sense, each of the basis vectors is within a constant factor of the shortest possible. This is a consequence of (6.16) in Theorem 6.5 from Chapter 3.

To minimise the amount of computational work required for reduction, we use the basis matrix $\mathbf{B}^{(k-1)}$ obtained from the reduction on the $(k-1)^{\text{th}}$ iteration as the input to the reduction step on the k^{th} iteration, after premultiplication by $\mathbf{L}$. In this way, we maintain a basis of $\mathbf{L}^k\mathbf{Q}\mathbb{Z}^n$.

We can now state an algorithm for attempting to discover the JMLEA. Before we do, a few minor points need to be clarified. The procedure *LLLreduce* is essentially Algorithm 7.1 from Chapter 3. It takes as its input a basis matrix $\mathbf{B}$ and outputs a Lovász-reduced basis matrix $\mathbf{B}'$. However, it differs in that it also outputs the unimodular transformation $\mathbf{U}$ such that $\mathbf{B}' = \mathbf{B}\mathbf{U}$. We find this useful for updating the matrix of index vectors, $\mathbf{S}^{(k)}$. Initially, $\mathbf{S}^{(0)}$ is set to the identity matrix without the first column (the $\mathbf{e}_i$ on line 3 represent the i^{th} column of $\mathbf{I}_n$) which is a basis of all index vectors with $s_1 = 0$. The initial basis matrix is set to $\mathbf{Q}\mathbf{S}^{(0)}$. Also, since *LLLreduce* takes no care to ensure that the corresponding index vectors in the Lovász-reduced basis matrix have $s_n > 0$, we must do this separately (at line 11).

ALGORITHM 5.1.

```
1  begin
2    B^(0) := (q_2, q_3, ..., q_n);
3    S^(0) := (e_2, e_3, ..., e_n);
4    ŝ := arg min_{v∈{e_2,e_3,...,e_n}} {G(v)};
5    k := 0;
6    while min_{1≤i≤n} {κ s^(k)_{n,i}} < β G(ŝ) do
7          k := k + 1;
8          (B^(k), U) := LLLreduce(LB^(k-1));
9          S^(k) := S^(k-1) U;
10         for i := 1 to n − 1 do
11             if s^(k)_{n,i} < 0 then s^(k)_i := −s^(k)_i fi;
12             if G(s^(k)_i) < G(ŝ) then ŝ := s^(k)_i fi;
13         od
14   od
15 end.
```

We have already described most of what Algorithm 5.1 is doing in trying to discover the JMLEA. To summarise, it is calculating a sequence of good approximations and returning that index vector which minimises G from the sequence of index

vectors corresponding to those approximations. It remains to briefly describe the stopping criterion for our algorithm. The algorithm terminates when the right-hand term in the expression for G in (3.8) has become so large for all the index vectors corresponding to approximations from the most recent reduction that, regardless of the left-hand term, we cannot better that of $\hat{\mathbf{s}}$. Because further reductions would most likely only increase the right-hand term further, it is appropriate to end the search. However, we have included a "safety factor" $\beta \geqslant 1$ to allow the search to continue a little longer, just in case.

Notice that Algorithm 5.1 only produces an association $\hat{\mathbf{s}}$. The estimates $\hat{T}$ and $\hat{\theta}$ are then obtained by substitution into (3.4) and (3.3), respectively.

It is unclear exactly how much computation is required by our algorithm in total. However, from our knowledge of the computational efficiency of the LLL algorithm, there is very good reason to be optimistic that the amount of computation required should be considerably less than a "brute force" search over index vectors. Numerical results obtained for Section 7 indicate that, usually, only a small number of iterations are required.

6. A Related Trigonometric Sum

In this section, we point out the relationship between the simultaneous Diophantine approximation problem discussed in Section 4 and the maximisation (in magnitude) of the trigonometric sum

$$A(\omega) = \sum_{j=1}^{n} e^{-iz_j\omega}. \tag{6.1}$$

It is natural to wonder if a relationship of some kind exists, because the magnitude of (6.1) can be thought of as the periodogram of the function

$$u(t) = \sum_{j=1}^{n} \delta(t - z_j),$$

a train of impulses (Dirac delta functions) at the measured TOAs. It seems intuitively obvious that a good candidate for the PRI of the observed pulse train is the inverse of a frequency which maximises $|A(\omega)|$. Furthermore, in Example 6.2 of Section 6 of Chapter 2, we showed that the correspondence of successive maxima of a periodogram with $n = 3$ and the best Diophantine approximations of the ratio $\alpha = (z_2 - z_1)/(z_3 - z_1)$ is one-to-one, in a certain sense.

We will now show a relationship between $|A(\omega)|$ and $\rho(\mathbf{x})$ as defined in (4.1) for $n \geqslant 3$. Given a vector of pulse indices, $\mathbf{s}$, we define the function

$$\omega(\mathbf{s}) = \frac{2\pi}{\hat{T}(\mathbf{s})} = 2\pi\frac{\mathbf{x}^T\mathbf{x}}{\boldsymbol{\zeta}^T\mathbf{x}}$$

where, as usual, $\mathbf{x} = \mathbf{Q}\mathbf{s}$ and $\boldsymbol{\zeta} = \mathbf{Q}\mathbf{z}$. Let

$$\boldsymbol{\epsilon} = \omega(\mathbf{s})\boldsymbol{\zeta} - 2\pi\mathbf{x}.$$

That is,

$$\epsilon_i = \omega(\mathbf{s})(z_i - \overline{z}) + 2\pi(s_i - \overline{s}),$$

where $\overline{z}$ and $\overline{s}$ denote the arithmetic means of the z_i and s_i, respectively. We note that

$$\sum_{i=1}^{n} \epsilon_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \epsilon_i^2 = 4\pi^2\rho(\mathbf{x})^2.$$

We can employ these identities to find that

$$\begin{aligned}
|A(\omega(\mathbf{s}))| &= \left|e^{-i[\omega(\mathbf{s})\overline{z} - 2\pi\overline{s}]}A(\omega)\right| \\
&= \Big[\Big(\sum_{j=1}^{n} \cos\epsilon_j\Big)^2 + \Big(\sum_{j=1}^{n} \sin\epsilon_j\Big)^2\Big]^{1/2} \\
&= \Big[\Big(\sum_{j=1}^{n} 1 - \tfrac{1}{2}\epsilon_j^2 + O\big(\epsilon_j^4\big)\Big)^2 + \Big(\sum_{j=1}^{n} \epsilon_j + O\big(\epsilon_j^3\big)\Big)^2\Big]^{1/2} \\
&= \Big[\big(n - 2\pi^2\rho(\mathbf{x})^2 + O(\rho(\mathbf{x})^4)\big)^2 + O(\rho(\mathbf{x})^3)^2\Big]^{1/2} \\
&= n - 2\pi^2\rho(\mathbf{x})^2 + O(\rho(\mathbf{x})^4).
\end{aligned}$$

Thus, we can see that there is indeed a link between the maximisation in magnitude of the trigonometric sum $A(\omega)$ (the periodogram) with the radius function ρ used in formulating the simultaneous Diophantine problem. It also provides us with a way of visualising the behaviour of Algorithm 5.1 in the frequency domain, which we will make use of in the next section.

7. Numerical Results

We now assess the performance of Algorithm 5.1 with some numerical tests. The algorithm was coded in C++. The LiDIA library (LiDIA Group, 1995) was used for its implementations of variants of the LLL algorithm. The constants $\beta = 2$ and $\gamma = 0.5$ were used for all numerical testing.

The first test performed was of the ability of the algorithm to find the correct association with the true pulse indices for various noise levels and various numbers of observed pulses under the condition $\lambda = 0.001$ and $T = 1$. Indeed, these settings for λ and T are used throughout this section. Figure 2 shows the results obtained. Each data point plotted is the average frequency of successful association from 500 trials. We can see that a success rate better than 99% is achieved even for a small number of observed pulses ($n > 8$) up to and exceeding a noise level of 1% of the

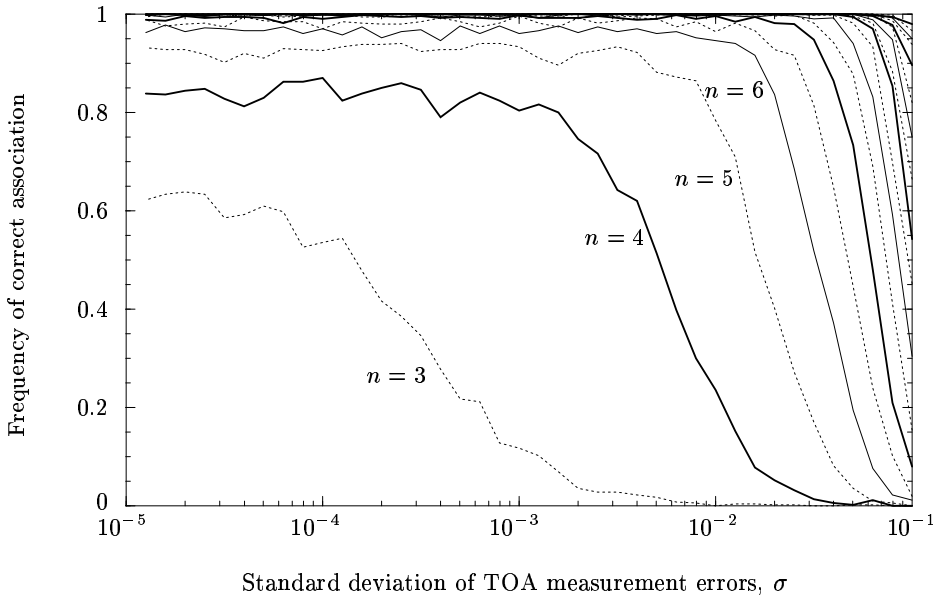


FIGURE 2. The experimental frequency of correct association of the pulse indices as a function of the standard deviation of the measurement errors, σ, for various numbers of pulses, n, with $T = 1$ and $\lambda = 10^{-3}$.

PRI. Of course, given that the correct association is made with the pulse indices, the variance of $\hat{T}$ is that which is obtained by linear regression. That is,

$$\operatorname{var} \hat{T} = \frac{\sigma^2}{\|\mathbf{Q}\hat{\mathbf{s}}\|_2^2}.$$

We can also see that, for any fixed n, the probability of correct association appears to approach an upper bound depending on n as σ is decreased. The upper bound represents the probability that the true pulse indices are coprime. The limiting probability is related to the Riemann zeta function. See CASEY & SADLER (1996) for a discussion of its relationship with the PRI estimation problem. The probability that n numbers chosen at random are coprime asymptotically approaches 1 very quickly as n increases, and we witness this in Figure 2.

The graphs of Figure 2 were all computed under the assumption that the constant κ, defined in (3.9), was known exactly. In practice, we expect that there would be some uncertainty as to the exact value of κ. Figure 3 shows the effect of uncertainty in the true value of κ to the probability of correct association, for the case where $n = 10$ and $T = 1$. The value κ' is used in place of κ in Algorithm 5.1. The graph for $\kappa' = \kappa$ is of course the same as that for $n = 10$ in Figure 2. For $\kappa' = 10^3\kappa$ and $\kappa' = 10^{-3}\kappa$ we notice a degradation in performance, which is to be expected. We also notice that the performance is more severely affected for $\kappa' = 10^{-3}\kappa$. This, and experimental evidence for other values of κ' (not presented here), leads us to believe that it is better to overestimate κ than to underestimate it. In any case, we can see that the estimation and association algorithm is fairly robust to uncertainty in κ,

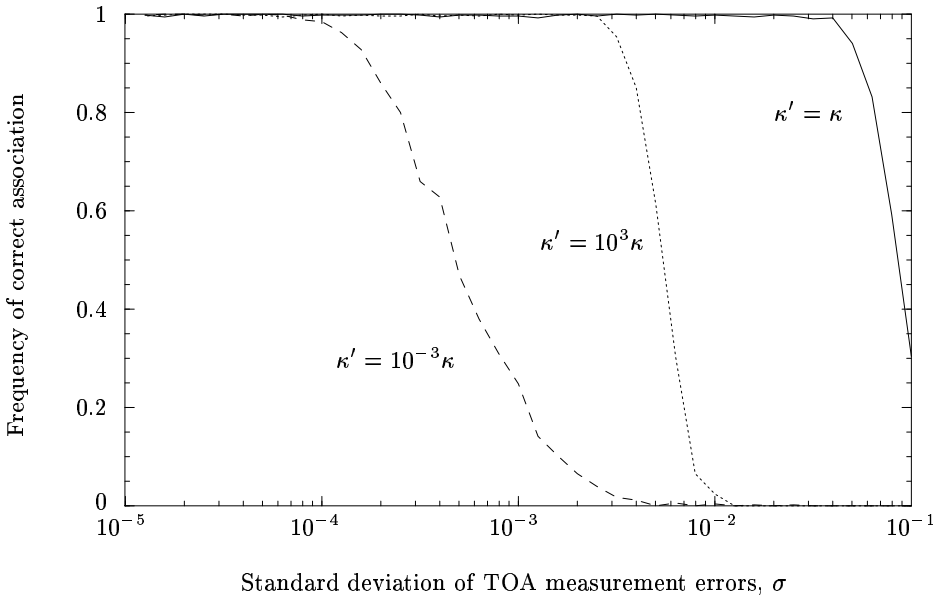


FIGURE 3. The experimental frequency of correct association of the pulse indices for $n = 10$ and $T = 1$ for erroneous estimates of κ.

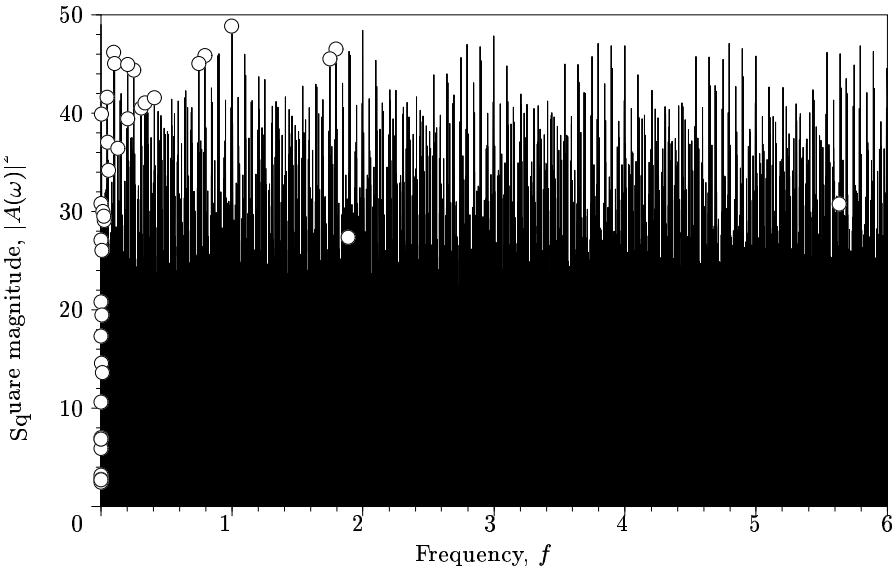


FIGURE 4. Interpretation of the outputs of the algorithm as a maximisation of the periodogram.

with $> 98\%$ probability of correct association when $\sigma < 10^{-4}$, even though κ' is out by three orders of magnitude.

Finally, we consider an interpretation of the behaviour of the lattice points produced by Algorithm 5.1 in the frequency domain. Figure 4 shows the square of the magnitude of $A(f)$, $f = 2\pi\omega$, as defined in (6.1) for a particular set of observations with $n = 7$, $T = 1$, $\lambda = 10^{-3}$ and $\sigma = 10^{-2}$. As witnessed in Figure 2, we have obtained over 95% experimental frequency of correct association for these parameters. The circles represent peaks at frequencies $\omega\big(\mathbf{s}_j^{(k)}\big)/(2\pi)$ for each of the matrices of index vectors $\mathbf{S}^{(k)}$ considered by Algorithm 5.1. Our algorithm, then, considers a

proportionally large number of low frequencies before "accelerating" away into the higher frequencies. In fact, the rate of increase in frequencies appears to be exponential. If a logarithmic scale had been used along the frequency axis, we would see that the frequency samples considered by the algorithm are roughly equally distributed over the range. For the set of observations used to generate Figure 4, the algorithm was able to correctly associate indices. Observe that the frequency corresponding to these indices (at $f = 0.999\,997$) is the largest peak in the periodogram for the range plotted.

Note that the number of different frequency points considered by the algorithm is 41. In contrast, the number of samples calculated to plot the full periodogram in Figure 4 was 600 000.

CHAPTER 7

CONCLUSIONS

The aim of this thesis has been to explore the topic of approximation of linear forms by lattice points, with emphasis on some problems of signal processing and, in particular, electronic support measures. The presentation has focussed on algorithms for Diophantine approximation, simultaneous Diophantine approximation and lattice reduction. The application has been to problems involving periodic pulse trains. Periodic pulse trains are a common feature of many ESM problems and many other physical problems. The thesis has demonstrated the applicability and effectiveness of Diophantine approximation to the solution of probability of intercept problems between periodic pulse trains and to the estimation of the period of a pulse train of which only a few sparse and noisy observations exist.

In Chapter 2, we introduced Diophantine approximation of a single real number. The intention was to explore methods of calculating best Diophantine approximations, both homogeneous and inhomogeneous, in both the absolute and relative senses, and to examine their relationship with other mathematical objects such as simple continued fractions, diagonal functions and Farey series. We began with homogeneous Diophantine approximation and showed how the best approximations in the absolute sense and relative sense could be calculated efficiently using Euclid's algorithm. We then showed the link between the sequence of best homogeneous Diophantine approximations of a number in the absolute sense and the convergents of its simple continued fraction expansion. The relationship was shown to be almost one-to-one. For inhomogeneous Diophantine approximation, we proved that the best approximations can be obtained efficiently using Cassels' algorithm. We then showed that the successive maxima of diagonal functions satisfying certain conditions enjoy an almost one-to-one correspondence with best homogeneous Diophantine approximations. We demonstrated that a periodogram of three samples with positive amplitudes is an example of such a diagonal function. We concluded the chapter with a discussion of the relationship between Farey series and best homogeneous Diophantine approximations. We review some of the basic properties of Farey series and showed how best homogeneous approximations with a prescribed approximation error can be located in a Farey series of the appropriate order.

We reviewed the theory of the geometry of numbers in Chapter 3. We introduced point lattices and introduced Minkowski's first (fundamental) and second theorems. We discussed the problem of finding short vectors in a lattice and developed a "brute force" algorithm for finding the shortest vector. We reviewed various notions

of lattice reduction, namely those of Gauss, Minkowski, Hermite, Korkin & Zolotarev and Lovász. We demonstrated the relationship between Gaussian reduction of a lattice and the centred continued fraction expansion of a complex number. Given the apparent computational infeasibility of the shortest lattice vector problem, we highlighted the importance of reduction in the sense of Lovász. It is important because an algorithm exists — the LLL algorithm — that can produce a Lovász-reduced basis from an arbitrary basis in an amount of time which is bounded above by a polynomial in the size of the input. At the same time, a Lovász-reduced basis contains vectors which are within a constant factor of the shortest possible vectors, in a certain sense. The behaviour of the LLL algorithm is analysed in the last section of Chapter 3.

In Chapter 4, we discussed simultaneous Diophantine approximation. The first part of the chapter described a theory for simultaneous Diophantine approximation using (ρ, h)-minimal sets. These sets can be employed in algorithms for producing the best simultaneous Diophantine approximations according to a quite general definition. The second part of the chapter discusses computationally efficient algorithms for producing good simultaneous Diophantine approximations.

The chapter began by discussing the theory of (ρ, h)-minimal sets. We used this theory to construct algorithms which find best approximations for lattices of rank two and three. The algorithms have a fairly simple, additive form and are guaranteed to find all the best approximations (or equivalent lattice points) of a given system with radius less than or equal to the minimum radius of the input lattice basis vectors. For lattices of rank two, we show that the algorithms are equivalent to additive forms of Euclid's algorithm or Gauss' algorithm (discussed in Chapter 2 and Chapter 3, respectively), depending on the nature of the simultaneous Diophantine approximation system it is applied to. For lattices of rank three, we were able to improve the computational efficiency of the algorithm. This resulted in an algorithm — the "accelerated" algorithm — which is very similar in structure to an algorithm of Furtwängler but which is capable of producing best approximations according to our more general definition. We provided a number of numerical examples to demonstrate the ability of the algorithms for lattices of rank three to find sequences of best approximations. We demonstrated the operation of the algorithm for finding best approximations of a line by lattice points — "traditional" simultaneous Diophantine approximation — with respect to various norms including the Euclidean norm and sup-norm. We also demonstrated its operation for finding best approximations of a plane by lattice points, *i.e.* best approximate integer relations.

However, the implementation of these algorithms depends on geometric properties which disappear in higher dimensions. Furthermore, it is likely that the problem of finding best approximations is computationally infeasible for lattices of arbitrary rank. For this reason, Chapter 4 continues by discussing approaches to simultaneous

Diophantine approximation which aim to find good, rather than best, approximations in lattices of arbitrary rank with a modest amount of computational effort. Brun's algorithm was discussed in this context. We found that the algorithm is attractive in its simplicity and has a natural geometrical interpretation. However, it is known that the algorithm does not always produce very good approximations. We then reviewed some of the more recent algorithms for simultaneous Diophantine approximation which are based on the LLL algorithm. As a typical example of this class of algorithms, we chose to examine the HJLS algorithm of Hastad *et al.* in detail. We also discussed the similar but independently-developed PSLQ algorithm of Ferguson & Bailey as well as algorithms of Just and Rössner & Schnorr. We concluded the chapter by presenting a numerical example which compared the performance of our accelerated algorithm, Brun's algorithm and the HJLS algorithm on lattices of rank three. For that example (Example 6.1), we found that both Brun's algorithm and the HJLS algorithm were not particularly good at finding best approximations, discovering only three and four, respectively, of the seven best approximations discovered by the accelerated algorithm.

Our discourse then turned to the application of this theory to some problems in signal processing. In Chapter 5, problems of determining intercept times or probabilities between two or more periodic pulse trains was discussed. Calculation of these quantities is important in the design and analysis of ESM equipment such as radar warning receivers. The chapter began with a discussion of the problem for two pulse trains. When only two pulse trains are involved, we found that the calculation of the intercept time for *in phase* initial conditions, which is to say the condition in which both pulse trains have the same phase, is a Diophantine approximation problem. As such, the pulse indices at which the first intercept occurs correspond to a best homogeneous Diophantine approximation in the absolute sense to the ratio of the PRIs of the pulse trains involved. Therefore, the intercept time can be efficiently computed using Euclid's algorithm, and the pulse indices correspond to the numerator and denominator of a convergent in the s.c.f. expansion of the PRI ratio. For *arbitrary phase* initial conditions, we found that the pulse indices corresponding to the first intercept are a best inhomogeneous Diophantine approximation. Thus, they can be efficiently found using Cassels' algorithm. We also found that, given one coincidence between the two pulse trains, all future coincidences can be found by means of a recurrence relation.

We then considered the probability of intercept of two pulse trains, under the assumptions that one or both of the phases is a random variable, uniformly distributed over the range of the associated PRI. Where one phase is random and the other known — the *discrete time* case — the probability of at least one coincidence after N pulses from the pulse train with known phase has a piecewise linear form in N, consisting of four segments. The boundaries between the segments are found to be determined by the denominators of convergents and intermediate fractions of the

s.c.f. expansion of the PRI ratio. For the problem where both phases are random — the *continuous time* case — we gave an expression for the probability of at least one intercept occurring within an observation interval of length t. We found that the expression for the probability of intercept as a function of t again consists of four linear segments but is complicated by the addition of an extra quadratic segment between two of the linear segments. However, we showed that the expression for discrete time case can be made to serve as a good approximation for the continuous time expression. For two pulse trains, we were able to make use of the theory relating best Diophantine approximations to the Farey series that we developed in Chapter 2 to give an expression for the discrete time probability of intercept as function of the PRI ratio. The expression was given in terms of neighbouring points and their *parents* (see Definition 5.1) in a Farey series of the appropriate order. A recursive procedure can then be devised which allows an average or representative probability of intercept to be calculated if the value of the PRI ratio is not known precisely.

For intercept problems involving more than two pulse trains, we found that the theory was much less well-developed. We were able to express the intercept time problems as simultaneous Diophantine approximation problems: as a homogeneous problem for in phase initial conditions or as an inhomogeneous problem otherwise. For three pulse trains, we found that the intercept time could therefore be calculated using either the additive or accelerated algorithm from Chapter 4. We were not able to derive expressions for the probability of intercept. However, we were able to give an exact expression for the continuous time probability over short time intervals and we were able to prove the negative result that the expression for the probability of intercept does not consist of a fixed number of linear segments for more than two pulse trains. To conclude the chapter, we reviewed some other approaches which have appear in the literature. In particular, we noted the similarity between our expression for the continuous time probability of intercept with that derived by Self & Smith. We warned of the dangers of using either expression as a source of information about the probability of intercept over a long time interval.

In Chapter 6, we examined the problem of estimating the period of a pulse train from which only a few, sparse and noisy measurements of TOAs have been made. This problem arises in an ESM setting where a scanning receiver infrequently observes the portion of parameter space in which a periodic emitter is operating. We proposed two statistical models for the observation process: a simple model and an extended model. In the simple model, we assumed that the measurement errors were independent, identically-distributed (i.i.d.) normal random variables but we assumed very little about the way in which pulse went missing from the record. In the extended model, we made the assumption that the differences between consecutive observed pulse indices were i.i.d. random variables from a geometric distribution. We formulated the maximum likelihood estimation problem for both models for

estimation of the PRI and phase of the pulse train. If the pulse indices of the observations are known, we found that the problem was simply one of linear regression. As the pulse indices are unknown, we found that maximum likelihood estimation of the parameters for the simple model was equivalent to a simultaneous Diophantine approximation problem in which we seek the best simultaneous Diophantine approximation in the relative sense which has the least approximation error. Because of this, we concluded that there were either no maximum likelihood estimates of the parameters or there were an infinitude.

For the extended model, we proposed a method of *joint maximum likelihood estimation and association* (JMLEA) of the PRI, phase and pulse indices. We found that, with sufficiently small measurement noise, the JMLEA is likely to be associated with a best simultaneous Diophantine approximation in a certain system. For this reason, we proposed a simultaneous Diophantine approximation algorithm for attempting to find the JMLEA. The algorithm we proposed is based on the LLL algorithm. We were also able to show a strong connection between the particular simultaneous Diophantine approximation system required for this problem and the maximisation of the periodogram of TOA data, further extending the results of this nature in Chapter 2. Finally, we presented some numerical simulations that show that our algorithm is able to correctly associate pulse indices, and thereby obtain statistically efficient estimates of the PRI and phase, with an experimental frequency exceeding 99% even for records of 9 pulses in which the expected number of missing pulses is 99.9% and the measurement error, as a proportion of PRI, was as high as 0.01. Moreover, we found that the algorithm appears to be quite robust with respect to parameters which are assumed to be known *a priori*, namely the measurement noise variance and the expected number of missing pulses.

In the author's opinion, the important original contributions of this thesis are:

- the enunciation and proof of the conditions under which the auxiliary convergents and intermediate auxiliary convergents of Cassels' algorithm are best inhomogeneous Diophantine approximations,
- the demonstration of the relationship between Diophantine approximation and certain diagonal functions and, in particular, the periodogram,
- the development of the theory of (ρ, h)-minimal sets leading to the derivation of algorithms for best simultaneous Diophantine approximation for lattices of rank three,
- the elucidation of the relationship between intercept time problems and the theory of Diophantine approximation, leading to a unified treatment of a number of intercept time problems and
- the application of a variant of the LLL algorithm to joint maximum likelihood estimation and association of PRI, phase and pulse indices for short, sparse and noisy TOA records of a periodic pulse train, thereby obtaining results which improve the state-of-the-art.

We conclude this thesis with a short discussion about further research which might follow from this work. The geometry of numbers and simultaneous Diophantine approximation are areas of mathematics that contain many open problems and are of great interest to the mathematical community. The development of algorithms is an important part of the theoretical development. The original algorithms presented here for simultaneous Diophantine approximation do not seem well-suited for generalisation to systems involving lattices of higher rank while retaining the freedom of choice of radius and height functions. However, the author intends to explore other avenues in this field.

The obvious area for improvement in the results concerning intercept time problems is in the problem of coincidence of more than two pulse trains. However, as we have already noted, the theoretical obstacles appear to be formidable at this time. Some progress might be made in the application of the additive and accelerated algorithms of Chapter 4 to the formulation of the probability of intercept between three pulse trains, but the practical interest in such a result may not be commensurate with the amount of effort required to obtain it. However, it may be easier to extend the results for coincidence of two pulse trains to problems where the two pulse trains involved are not strictly periodic but rather have some closely-related modulation such as stagger or jitter. A STAGGERED pulse train is one which is created from the superposition of several periodic pulse trains with the same PRI but different phases. A JITTERED is a periodic pulse train to which some noise has been added to the TOAs.

The results we presented that relate simultaneous Diophantine approximation to maximisation of the periodogram do not appear to be the best possible. A more thorough characterisation of the relationship with best simultaneous Diophantine approximations seems possible and desirable. While we have demonstrated an algorithm that can be interpreted as finding successive peaks in a periodogram, a more thorough exploitation of the properties of the periodogram should yield improved performance and application to a wider range of problems.

However, it is the application of simultaneous Diophantine approximation to other problems in signal processing which holds the most interest for the author. Applications which the author has identified include stochastic resonance, the calculation of coefficients for FIR filters and for beamforming and signal and image compression. The work presented in Chapter 6 for estimation of the period of an imperfectly observed pulse train can be generalised not only to different statistical models for pulse trains, such as an incoherent model where errors accumulate, but also to point processes in higher dimensions. For example, the method could be generalised to estimating the basis of a point lattice of which only sparse and noisy observations exist. In conjunction with a better understanding of its relationship with the periodogram, the theory could be useful in irregular sampling. These are application areas which the author intends to study at the earliest opportunity.

BIBLIOGRAPHY

Bergman, G. M. (1980). Notes on Ferguson and Forcade's generalized Euclidean algorithm. Unpublished report, University of California, Berkeley.

Brentjes, A. J. (1981). *Multi-Dimensional Continued Fraction Algorithms*, vol. 145 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam.

Brun, V. (1919). En generalisation av kjedebrøken I. *Skr. Vidensk. Selsk. Kristiana, mat. nat. kl.*, (6), 1–29.

Brun, V. (1920). En generalisation av kjedebrøken II. *Skr. Vidensk. Selsk. Kristiana, mat. nat. kl.*, (6), 1–24.

Casey, S. D. & Sadler, B. M. (1995). A modified Euclidean algorithm for isolating periodicities from a sparse set of noisy measurements. In *Proceedings of ICASSP '95*, vol. 3, 1764–1767.

Casey, S. D. & Sadler, B. M. (1996). Modifications of the Euclidean algorithm for isolating periodicities from a sparse set of noisy measurements. *IEEE Trans. Signal Process.*, **44** (9), 2260–2272.

Cassels, J. W. S. (1954). Über $\underline{\lim}_{x\to+\infty} x|x\vartheta+\alpha-y|$. *Math. Ann.*, **127**, 288–304.

Cassels, J. W. S. (1957). *An Introduction to Diophantine Approximation*. Cambridge University Press.

Cassels, J. W. S. (1971). *An Introduction to the Geometry of Numbers*. Springer-Verlag, Berlin.

Clarkson, I. V. L., Perkins, J. E. & Mareels, I. M. Y. (1996). Number theoretic solutions to intercept time problems. *IEEE Trans. Inform. Theory*, **42** (3), 959–971.

Cohen, H. (1993). *A Course in Computational Algebraic Number Theory*. Springer-Verlag, Berlin.

Daudé, H., Flajolet, P. & Vallée, B. (1994). An analysis of the Gaussian algorithm for lattice reduction. In Adleman, L. M. & Huang, M.-D. (eds.), *Algorithmic Number Theory*, no. 877 in Lecture Notes in Computer Science, 144–158. Springer-Verlag, Berlin.

Delone, B. N. & Faddeev, D. K. (1964). *The Theory of Irrationalities of the Third Degree*, vol. 10 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island.

Descombes, R. (1956). Sur la répartition des sommets d'une ligne polygonale régulière non fermée. *Ann. Sci. École Norm. Sup. (3)*, **73**, 283–355.

Dickson, L. E. (1919). *History of the Theory of Numbers*, vol. 1 (Divisibility & Primality). Carnegie Institution of Washington.

Dziech, A. (1993). *Random Pulse Streams and their Applications*, vol. 44 of *Studies in Electrical and Electronic Engineering*. Elsevier, Amsterdam.

Ferguson, H. R. P. & Bailey, D. H. (1991). A polynomial time, numerically stable integer relation algorithm. Tech. Rep. RNR–91–032, NASA Ames Research Center, Mail Stop T27A–1, Moffett Field, CA 94035–1000, USA.

Ferguson, H. R. P., Bailey, D. H. & Arno, S. (1996). Analysis of PSLQ, an integer relation finding algorithm. Submitted for publication.

Ferguson, H. R. P. & Forcade, R. W. (1979). Generalization of the Euclidean algorithm for real numbers to all dimensions higher than two. *Bull. Am. Math. Soc. (N. S.)*, **1**, 912–914.

Ferguson, H. R. P. & Forcade, R. W. (1982). Multidimensional Euclidean algorithms. *J. Reine Angew. Math.*, **334**, 171–181.

Fincke, U. & Pohst, M. (1985). Improved methods for calculating vectors of short length in a lattice, including a complexity analysis. *Math. Comp.*, **44** (170), 463–471.

Friedman, H. D. (1954). Coincidence of pulse trains. *J. Appl. Phys.*, **25** (8), 1001–1005.

Furtwängler, Ph. (1927). Über die simultane Approximation von Irrationalzahlen. *Math. Ann.*, **99**, 71–83.

Gauss, C. F. (1801). *Disquisitiones Arithmeticae*. Fleischer, Leipzig. English transl. of 2nd ed. (1870) by Arthur A. Clarke, Yale University Press, 1966.

Golub, G. H. & van Loan, C. F. (1989). *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2nd edition.

Greiter, G. (1977). *Mehrdimensionale Kettenbrüche*. Ph.D. thesis, Technische Universität München.

Hardy, G. H. & Wright, E. M. (1979). *An Introduction to the Theory of Numbers*. Oxford University Press, 5th edition.

Hastad, J., Just, B., Lagarias, J. C. & Schnorr, C. P. (1989). Polynomial time algorithms for finding integer relations among real numbers. *SIAM J. Comput.*, **18** (5), 859–881.

Hawkes, R. M. (1983). The analysis of interception. Unpublished research report, Defence Science and Technology Organisation, P.O. Box 1500, Salisbury, 5108, South Australia.

Heath, T. L. (1908). *The Thirteen Books of Euclid's Elements*. Cambridge University Press.

Hermite, Ch. (1850). Extraits de lettres de M. Ch. Hermite à M. Jacobi sur différents objets de la théorie des nombres, Deuxième lettre. *J. Reine Angew. Math.*, **40**, 279–290.

Hogg, R. V. & Craig, A. T. (1978). *Introduction to Mathematical Statistics.* Macmillan Publishing, New York, 4th edition.

Jacobi, C. G. J. (1868). Allgemeine Theorie der kettenbruchähnlichen Algorithmen. *J. Reine Angew. Math.*, **69**, 29–64.

Just, B. (1992). Generalizing the continued fraction algorithm to arbitrary dimensions. *SIAM J. Comput.*, **21** (5), 909–926.

Kaib, M. (1994). *Gitterbasenreduktion für beliebige Normen.* Ph.D. thesis, Johann Wolfgang Goethe-Universität, Frankfurt am Main.

Kelly, S. W., Perkins, J. E. & Noone, G. P. (1996). The effects of synchronisation on the probability of pulse train interception. *IEEE Trans. Aerospace Elec. Systems*, **32** (1), 213–220.

Khinchin, A. Ya. (1964). *Continued Fractions.* University of Chicago Press, 3rd edition.

Knuth, D. E. (1981). *The Art of Computer Programming*, vol. 2 (Seminumerical algorithms). Addison-Wesley, Reading, Massachusetts, 2nd edition.

Lagarias, J. C. (1980). Worst-case complexity bounds in the theory of integral quadratic forms. *J. Algorithms*, **1**, 142–186.

Lagarias, J. C. (1982). The computational complexity of simultaneous Diophantine approximation problems. In *23rd Annual Symposium on the Foundations of Computer Science*, 32–39. IEEE Computer Society.

Lagarias, J. C. (1983). Best Diophantine approximations to a set of linear forms. *J. Austral. Math. Soc. Ser. A*, **34**, 114–122.

Lagarias, J. C. (1994). Geodesic multidimensional continued fractions. *Proc. London Math. Soc. (3)*, **69**, 464–488.

Lagarias, J. C., Lenstra, Jr., H. W. & Schnorr, C. P. (1990). Korkin-Zolotarev bases and successive minima of a lattice and its reciprocal lattice. *Combinatorica*, **10** (4), 333–348.

Lagrange, J. L. (1773). Recherches d'arithmétique. *Nouv. Mém. Acad. Berlin*, 265–312. *Œuvres III*, 695–795.

Lenstra, A. K., Lenstra, Jr., H. W. & Lovász, L. (1982). Factoring polynomials with rational coefficients. *Math. Ann.*, **261**, 515–534.

Levitan, B. M. & Zhikov, V. V. (1982). *Almost Periodic Functions and Differential Equations.* Cambridge University Press.

Lewis, H. R. & Papadimitriou, C. H. (1981). *Elements of the Theory of Computation.* Prentice-Hall, Englewood Cliffs, New Jersey.

LiDIA Group (1995). *LiDIA — A Library for Computational Number Theory.* Universität des Saarlandes.

Miller, K. S. & Schwarz, R. J. (1953). On the interference of pulse trains. *J. Appl. Phys.*, **24** (8), 1032–1036.

Minkowski, H. (1896*a*). Généralisation de la théorie des fractions continues. *Ann. Sci. École Norm. Sup. (3)*, **13**, 41–60.

Minkowski, H. (1896*b*). *Geometrie der Zahlen*. Teubner, Leipzig.

Niven, I. & Zuckerman, H. S. (1980). *An Introduction to the Theory of Numbers*. John Wiley & Sons, New York.

Perron, O. (1907). Grundlagen für eine Theorie des Jacobischen Kettenbruchalgorithmus. *Math. Ann.*, **64**, 1–76.

Richards, P. I. (1948). Probability of coincidence for two periodically recurring events. *Ann. Math. Stat.*, **19** (1), 16–29.

Rössner, C. & Schnorr, C. P. (1996). An optimal, stable continued fraction algorithm for arbitrary dimension. In *Proceedings of the 5th IPCO Conference on Integer Programming and Combinatorial Optimization*, Lecture Notes in Computer Science. Springer-Verlag. Preprint.

Scharlau, W. & Opolka, H. (1985). *From Fermat to Minkowski: Lectures on the Theory of Numbers and Its Historical Development*. Springer-Verlag, Berlin.

Schnorr, C. P. & Euchner, M. (1994). Lattice basis reduction: Improved practical algorithms and solving subset sum problems. *Math. Programming*, **66**, 181–199.

Self, A. G. & Smith, B. G. (1985). Intercept time and its prediction. *IEE Proc.*, **132F** (4), 215–222.

Siegel, C. L. (1989). *Lectures on the Geometry of Numbers*. Springer-Verlag, Berlin.

Slocumb, B. J. (1993). A number theory approach for evaluating the probability of coincidence of two periodic pulse trains. Unpublished technical report, Cooperative Research Centre for Robust and Adaptive Systems, 71–Labs, DSTO Salisbury, P.O. Box 1500, Salisbury, 5108, South Australia.

Stein, S. & Johansen, D. (1958). A statistical description of coincidences among random pulse trains. *Proc. IRE*, **46** (5), 827–830.

Vallée, B. (1991). Gauss' algorithm revisited. *J. Algorithms*, **12**, 556–572.

van Emde Boas, P. (1981). Another NP-complete partition problem and the complexity of computing short vectors in a lattice. Tech. Rep. 81–04, Mathematisch Instituut, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands.

Voronoĭ, G. (1896). *On a Generalisation of the Algorithm for Continued Fractions*. Ph.D. thesis, Warsaw. (Russian).

Williams, H. C., Cormack, G. & Seah, E. (1980). Calculation of the regulator of a pure cubic field. *Math. Comp.*, **34** (150), 567–611.

INDEX

ERRATA

p. v	l. 2↓	The text "geometry numbers" should read "geometry *of* numbers."
p. xi	l. 2↓	The following sentence should be appended: "We use $\mathbb{C}$ to denote the complex numbers."
p. xi		An additional item of notation should be described. We should add: "In reference to Boolean expressions, the operations $\wedge$, $\vee$ and $\neg$ denote *and*, *or* & *not*. In accordance with the usual convention, $\neg$ takes precedence over $\wedge$ which in turn takes precedence over $\vee$."
p. 8	l. 19↓	The word "fist" should read "first."
p. 29	l. 10↓	The occurrence of q_{N-1} in the denominator should be replaced by q_{N-3}.
p. 47	l. 7↑	The occurrence of A_2^2 on the left hand side should be replaced by A_3^2.
p. 54	l. 10,12↓	The word "lesser" should be replaced by "smaller."
p. 59	l. 10↓	The qualification "$0 \leqslant c_1, c_2 < 1$" is unnecessarily strong and should be replaced by "c_1, c_2 are not both integers."
p. 61	l. 3↓	The word "mapping" should read "*a* mapping."
p. 62	l. 2↑	After the text "there exists a unique upper triangular matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$" there should follow the qualification "*with positive diagonal elements.*"
p. 63	l. 11↑	The occurrence of $\|\mathbf{x} + \mathbf{x}\|$ should be replaced by $\|\mathbf{x} + \mathbf{y}\|$.
p. 65	l. 11↑	The reference to "Figure 16" should be to "Figure 2."
p. 73	l. 1↓	The occurrence of $\mathbf{Q}$ should be replaced by $\mathbf{Q}'$.
p. 73	l. 5↑	Between the sentence which ends "...when $i > 0$" and the sentence which begins "We call an expansion of this type...," we should insert the following clarification: "If the fraction is continued only n times then the final term is ϵ_{n-1}/ξ_n, where $\xi_n \in \mathbb{C}$, $0 \leqslant \Re\{\xi_n^{-1}\} \leqslant \frac{1}{2}$ and $\|\xi_n\| \leqslant 1$."
p. 75	l. 8↑	The word "bound" should read "bounded."
p. 84	l. 4↓	The word "constant" should read "*a* constant."
p. 92	l. 1↑	The word "are" should read "is."
p. 140	l. 4↓	The word "parallelepiped" should read "parallelepipeds."

p. 146 l. 17↑ The text "so that..." should be replaced by "on which the exchange is performed satisfies..."

p. 147 l. 9,10↓ The index j should be replaced by i throughout.

p. 147 l. 3↑ The word "integers" should read "integer."

p. 148 l. 2↓ The occurrence of α should be replaced by α_i.

p. 174 l. 6↑ The word "is" should read "are."

p. 175 l. 7↓ The word "or" should read "and."

p. 188 l. 15↑ The text "an adaptation LLL algorithm" should read "an adaptation *of the* LLL algorithm."

p. 189 l. 5↑ The word "are" should read "is."

p. 193 l. 11↓ The second last sentence of Theorem 4.1 should be deleted. The last sentence should begin "In this case, ..." instead of "Otherwise, ..."

p. 194 l. 6↓ The word "points" should read "point."

p. 194 l. 3↑ The occurrence of $+$ should be replaced by $-$.

p. 195 l. 9↓ The occurrence of ∂C should be replaced by $\partial^2 C$.

p. 199 l. 13↑ On line 6 of Algorithm 5.1, "$\min_{1\leqslant i\leqslant n}$" should read "$\min_{1\leqslant i\leqslant n-1}$."

p. 203 l. 5↑ The text "magnitude of $A(f)$, $f = 2\pi\omega$" should read "magnitude of $A(\omega)$, $\omega = 2\pi f$."