STRUCTURAL PROPERTIES OF MUSICAL SCALES

NORMAN CAREY AND DAVID CLAMPITT

Contents

Preface

This unpublished paper was our first and rather fulsome attempt to treat the topic of "well-formed scales." We wrote it in the early-mid 1980s, the results of which were later synthesized and greatly abbreviated to form the basis of our first published work on the topic.[1] Whereas this paper situates the topic in Pythagorean pitch space ($2^a \times 3^b$), the 1989 paper factors out the role of octaves (2^a) and focuses on pitch classes generated by fifths (3^b). We returned to the methodology developed here in a 1996 paper where we discuss the pervasive presence of "regions" in medieval tonal spaces.[2] The linear ordering of Pythagorean pitch space described on page 7ff became the basis for the study of "Lambda words."[3]

---

[1]Carey, N. and D. Clampitt, Aspects of well-formed scales. 1989. *Music Theory Spectrum*. 11(2), 187-206.

[2]Carey, N. and D. Clampitt, Regions: A theory of tonal spaces in early medieval treatises. 1996. *Journal of Music Theory*. 40(1), 113-147.

[3]Carey, N. Lambda words: A class of rich words over an infinite alphabet. 2013. *Journal of Integer Sequences*. Article 13.3.4.

1. Introduction

There exists a correspondence, hitherto unacknowledged, linking many significant musical scales to the elements of a certain class of mathematical structures. These structures are interesting in themselves for their purely mathematical properties, involving topics from the theory of numbers such as continued fractions, Diophantine equations, and the Chinese Remainder Theorem, among others. The scales which share this common structure include certain pentatonic, diatonic, and chromatic scales, as well as an Arabic scale system of 17 tones to the octave and a Chinese system of 53 tones to the octave, and also such important musical relationships as tonic-subdominant-dominant and the octave itself.

Moreover, the existence of such a relationship raises a number of questions, discussed in a brief interpretation which follows the mathematical argument, concerning the way the mind hears music. For example, the mathematical description will suggest a basis for the sensation many listeners have of hearing a scale such as the diatonic scale as a succession of equal steps like steps of a ladder, in spite of the fact that the diatonic scale is composed of unequal scale-step intervals. On the other hand, the whole-tone scale, which *is* composed of equally-spaced whole tones when played on the piano, does not evoke this response from most listeners.

The paradigmatic status of the diatonic scale in this and many other cultures is another large question on which the mathematical correspondence has some bearing. There is, again at least for many in this culture, a sense of completeness in the diatonic scale, an intuitive sense of closure in a pattern of growth leading to this scale, that is reflected in the remarkable passage of the *Timaeus* where Plato describes the creation of the World-Soul according to the proportions of the diatonic scale. Our analysis, deriving from the same Pythagorean tradition as this passage, will locate the diatonic scale in a hierarchy of scales of similar structure, and suggest some reasons for its pre-eminence.

The theory, however, is primarily descriptive. We demonstrate that all of the scales mentioned above belong to the set of *well-formed Pythagorean scales*, all of which are generated by the formula

$$R_{(A,B)} = \left\{ 2^{(a_k(-1)^k(B-n) \bmod A} \cdot 3^{(b_k(-1)^k(n-A) \bmod B} \;\middle|\; 0 < n < A+B \right\}$$

where A/B and a_k/b_k are members of a certain sequence of rational approximations to $\log_2 3$.

In what follows, the procedure will be to consider the set of tones expressed in terms of the intervals octave and twelfth, which we will see is equivalent to considering the set P where $P = \{2^a 3^b \mid a, b \in Z\}$. (Throughout, Z refers to the set of all integers.) This is known as *Pythagorean tuning*, and strictly speaking the scales we will be discussing are all Pythagorean scales. Thus, for example, when we speak of the diatonic scale we are referring to the Pythagorean diatonic scale. However, after developing the theory in terms of Pythagorean tuning we will indicate how it may be generalized to apply to other tuning systems, such as equal temperament or mean-tone temperament.

We will show in an informal way that we can translate back and forth between the set P and the notes of ordinary musical notation, but this is merely a convenience. We make no assumptions based on ordinary musical notation, and as we will discuss below the elements of P are for us the symbols representing the tones and intervals of musical space. A general Pythagorean scale will be represented by the ordered sequence of P numbers $2^{x_1}3^{y_1} < 2^{x_2}3^{y_2} < \cdots < 2^{x_1+1}3^{y_1}$, where the last value is the double of the first, just as the last tone of the scale spanning an octave is the duplication of the first.

Once this mapping between tone and number is set in place we will develop a working definition of the mathematical entity we call a *region*, denoted by $R_{(A,B)}$. The concept of a region allows us to define a *well-formed Pythagorean scale* and a *modal system* of well-formed Pythagorean scales. The reader, even one without a thorough mathematical background, will be able to use this operational definition to see that each of the scales mentioned in the introduction is a well-formed Pythagorean scale.

Following this ad hoc derivation of the notion of a region, the exposition moves to a much higher level of generality to delineate the algebraic structure underlying all regions and hence all well-formed scales. The theorems we will prove will give a complete account of the mathematical relationships which characterize the notion of region in general. The development will lead us to results which can be summarized by referring to the formula

$$R_{(A,B)} = \left\{ \left. 2^{(a_k(-1)^k(B-n)) \bmod A} \cdot 3^{(b_k(-1)^k(n-A)) \bmod B} \;\right|\; 0 < n < A+B \right\}.$$

Assuming A/B and a_k/b_k belong to the required sequence, the tones of the well-formed Pythagorean scale associated with (A,B) correspond to the elements of $R_{(A,B)}$. As n ranges from 1 to $B+1$ one octave of the well-formed scale is obtained, and each sequence of $B+1$ consecutive integers within the sequence from 1 to $A+B-1$ produces what a musician would call a mode or modal variety of that scale over the range of one octave. The striking fact is that counting each modal variety as a separate scale form exactly 100 different scale forms from world music are seen to be well-formed scales. The roundness of this number is doubtless a coincidence, but it suggests the variety of forms that are manifestations of a single underlying phenomenon according to this theory.

2. Pythagorean Pitch Space and the Notion of *Region*

2.1. **Some Preliminary Mathematics.** For the mathematician we enumerate the technical results which characterize the notion of a region and the well-formed scales which it contains: We will be considering a subset of P,

$$\hat{P} = \left\{2^a 3^b \mid a, b \in Z; a, b \geq 0\right\}.$$

Thus $\hat{P} \subset Z$ and its elements can be ordered in sequence, $1, 2, 3, 4, 6, 8, 9, 12$, etc. We make use of the notion of a *connected set*. We will give a technical definition of connectedness, but for the time being it is enough to think of a connected subset of $\hat{P}$ as an unbroken section of the above sequence. Likewise, a connected subset of Z would be a set of consecutive integers. We show that each region is a connected subset of $\hat{P}$. Further we show that the region $R_{(A,B)}$ is a connected and symmetrical subset of a finite group associated with the ordered pair (A,B) which we denote by $G_{(A,B)}$ and which is isomorphic to the cyclic group Z_{AB}.

We show that the mapping $n \to R_{(A,B)}(n)$ is strictly increasing for n from 1 to $A+B-1$, and we show that each region $R_{(A,B)}$ is the image of the last $A+B-1$

elements of Z_{AB} under an isomorphism $\nu : Z_{AB} \to G_{(A,B)}$. If we consider the elements of Z_{AB} to be ordered as ordinary integers, the isomorphism ν preserves both order and connectedness as it maps these $A+B-1$ elements of Z_{AB} onto the elements of $R_{(A,B)}$.

We prove also that in general well-formed Pythagorean scales are made up of two scale steps. The two intervals are distinctly different, but from the point of view of group theory they will be seen to represent a single element.

Along the way we will prove a little-known theorem of elementary number theory, involving the solutions in non-negative integers of the Diophantine equation $ax + by = c$. As Davenport writes, "The question of when this equation is soluble in natural numbers is a more difficult one, and one that cannot well be completely answered in any simple way."[4]

In discussions of art and mathematics there are frequent references to the celebrated Golden number and the Fibonacci sequence associated with it. The theory of well-formed Pythagorean scales reveals the basis for the frequent speculation on the role of the Fibonacci sequence in the formation of scales.

2.2. **Musical Scales.** Since we will be considering musical notions in a very abstract setting, it may be helpful to consider some reflections on the nature of musical scales. In the most open sense a scale is understood to be an ordered set of pitches spanning an octave. We would like our notion of scale to be as unrestricted as possible to start with, but we note immediately the role of the octave as the frame for the scale. This is because tones at the interval of one or more octaves are universally considered to be "the same" except for register: The octave is the equivalence relation among tones. The scale may be experienced, then, as the movement outward and away from an initial tone, with a sense of return to the tone in its representation an octave higher. This sense of movement is the melodic sense of the scale, which carries with it the implcation of a tonal context, that is, a musical situation in which there is a central tone to which all other tones relate. Without offering an explicit definition of tonality, we may at least say that its existence and operation depend upon the polarity between tonic and dominant, that is, between the central tone and the tone a perfect fifth above it.

The scale also has a more static aspect as a repertoire of available tones. In practice, some scales are very much melodic structures in themselves: certatinly the diatonic and pentatonic scales are, while in the case of the chromatic scale the melodic sense is present but attenuated. In the Western world though, the chromatic is pre-eminently the repository of available tones.

In either of these two contexts, however, the spatial analogy is useful for thinking about scales. The octave is the measure of musical space, a yardstick marking off equal portions of space, and a scale is a particular way of filling up that musical space. Once the scale is determined through one octave it is completely determined: a C major scale transposed an octave higher is still a C major scale. Not only is the scale independent of register, but the type of scale is determined not by particular pitches but by a sequence of intervals. Thus the invariant characteristic of the major mode of the diatonic scale is the sequence of intervals *whole step*, *whole step*, *half step*; *whole step*; *whole step*, *whole step*, *half step*. Similarly, even the most complex musical edifice is essentially determined by its interval structure.

[4] H. Davenport *The Higher Arithmetic* p. 32 Dover 1983.

2.3. **Pitch and Number.** Pythagoras is credited with the knowledge, probably available to the ancient Babylonians and Egyptians as well, that musical intervals correspond to simple proportions. The octave is determined by the ratio 2 : 1, the fifth by 3 : 2, and the fourth by 4 : 3. Evidently an octave plus a fifth, or a twelfth, corresponds to the ratio 3 : 1, and the double octave to 4 : 1. Pythagoras thought in terms of string lengths, which are inversely proportional to pitch, while today we are more inclined to consider frequencies, which are directly proportional to pitch. This relationship between pitch and frequency is the fundamental correspondence between tone and number.

We may assign, then, a positive real number to every pitch, and, in principle, every positive real number determines a pitch. A musical interval corresponds to a ratio between two frequencies, with the value 1 assigned to unison, while any number greater than 1 represents an interval upward and a number strictly between 0 and 1 representing a downward interval. That is, there exists a one-to-one correspondence between the set of positive real numbers R_0^+ and all possible musical intervals. Furthermore, two intervals with frequency-ratios r_1 and r_2 may be combined to form another interval, and the frequency-ratio of the resultant interval equals $r_1 r_2$, the product of the two frequency ratios. Given $r \in R_0^+$, corresponding to a given interval, adding unison to the interval leaves the interval unchanged, which corresponds to the fact that $1r = r$, and $1/r \in R_0^+$ with $r\frac{1}{r} = 1$, so for any interval there is another interval which combined with it results in unison.

The mathematical reader will recognize that the above can be stated concisely by saying that the set of all musical intervals is a group isomorphic to the group of positive real numbers . We can think of this isomorphism as providing a precise translation between intervals and numbers, where adding intervals corresponds to multiplying frequency-ratios.[5]

R_0^+ represents the universal set of pitches, and R_0^+ with ordinary multiplication the universal group of intervals. This set is too big and chaotic, however, to be useful as a musical universe of discourse. We consider instead a subset of the universal group, the set of intervals which can be expressed as combinations of octaves and twelfths. Since the octave (upwards) corresponds to the ratio 2/1 and the twelfth (upwards) to the ratio 3/1, and since the ratio corresponding to the combination of two intervals is the product of their respective frequency ratios, this set is $P = \left\{2^a 3^b \mid a, b \in Z\right\}$. Evidently, if two elements $p_1 = 2^a 3^b$ and $p_2 = 2^c 3^d$ are in P their product, $p_1 p_2 = 2^{a+c} 3^{b+d}$ is in P, and if $p = 2^a 3^b$ is an element of P, $(2^a 3^b)(2^{-a} 3^{-b}) = 1$, so every element has an inverse which is in P. Therefore, P is a subgroup of R_0^+, the subgroup generated by the elements 2 and 3.

Of course, P can also represent tones, and we can assign note names in ordinary musical notation to elements of P in the following way: We assign a tone F_0 to the number 1. The subscript on a note refers to the octave register, with the octave

[5]A *group* G is a non-empty set together with a rule for combining any pair of elements in G, called an operation or a product, such that if $g_1, g_2 \in G$, $g_1 g_2 \in G$, and such that the operation is associative, i.e., $(g_1 g_2) g_3 = g_1 (g_2 g_3)$ and containing an element e such that $ge = eg = g$ for all $g \in G$, and containing an element $\dot{g}$ for every $g \in G$ such that $g\dot{g} = e = \dot{g}g$.The element e is called the *identity element*, and the element $\dot{g}$ is called the *inverse* of g. It is easy to prove that e is unique, and given $g \in G$, $\dot{g}$ is unique.

A subset of G which is a group is of course a *subgroup* of G

Two groups G and H with operations $\cdot$ and $\circ$, respectively, are said to be *isomorphic* if there exists a one-to-one mapping $i : G \to H$ such that for all $g_1, g_2 \in G$, $i(g_1 \cdot g_2) = i(g_1) \circ i(g_2)$.

containing F_0 as a reference point. The choice of "F_0" is for notational convenience and need not refer to any specific pitch. The note-name associated with 2^a3^b is the name of the note a octaves and b twelfths from F_0, upwards or downwards depending on the signs on a and b.

It is in fact the case that there is a one-to-one correspondence between the elements of P and all the notes of ordinary notation, which is not surprising given the long history of Pythagorean tuning in musical theory. Since we offer this "dictionary" only as a convenience for translation into the more familiar language of musical notation we do not go into this one-to-one correspondence rigorously, only remark that, first of all, the note name given by a letter and a certain number of sharps or flats can be associated with a single integer by considering the familiar cycle of fifths (or twelfths):

...F♭	C♭	G♭	D♭	A♭	E♭	B♭	F	C	G	D	A	E	B	F♯	C♯	G♯	D♯...
...−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10...

and secondly, the register of a note can be associated with the number of octaves it is from some reference note.[6]

For us, the preferred symbol for representing a tone or interval will be an element of P, unencumbered by the diatonic bias of ordinary notation, with its seven letters corresponding to the seven tones of the diatonic scale system. As we saw above, the system of letters and sharps or flats obscures the more fundamental relationship which is the number of fifths (or twelfths) from some reference point. Therefore we will make free use of the elements of P to represent tones as well as intervals, and the distinction between tones and intervals should always be clear from the context.

There is a further simplification of P which we will make much use of, and which will help make the distinction between tones and intervals clear in practice. We define $\hat{P} = \left\{2^a3^b \mid a, b \in Z, a \geq 0, b \geq 0\right\}$, and in fact we can say that $\hat{P}$ is the restriction of P to the elements of P which are integers. Any element of P can be represented (in many ways) by a ratio of two $\hat{P}$ elements. The relationship between $\hat{P}$ and P is similar to that between the positive integers and the positive rationals. *We will usually think of the elements of $\hat{P}$ as representing tones, which we can name in reference to the note F_0.* Then any interval can be represented as the ratio between two $\hat{P}$ numbers, that is, as an element of P. Likewise, any sequence of intervals can be represented by a sequence of ratios of $\hat{P}$ numbers.

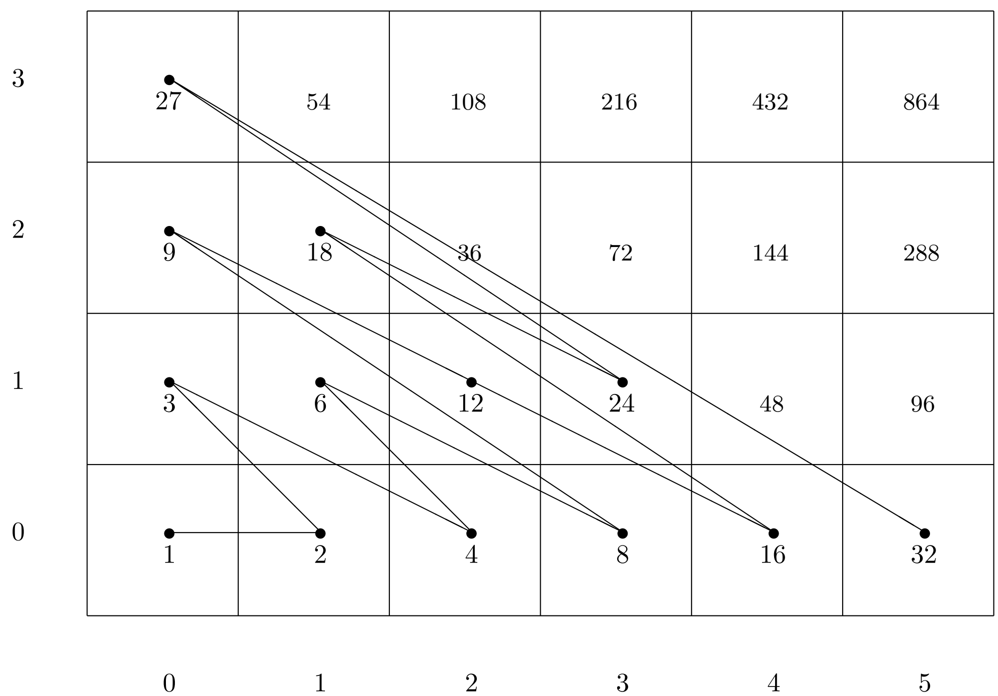
Next we remark that there is an isomorphism of groups $Z \times Z \cong P$, given by the mapping $\ell : Z \times Z \to P : (z_1, z_2) \to 2^a3^b$ where the operation in $Z \times Z$ is addition defined by $(a, b) + (c, d) = (a + c, b + d)$ for all ordered pairs of integers. Clearly ℓ is one-to-one and onto P, and $\ell\left((a, b) + (c, d)\right) = \ell(a, b) \cdot \ell(c, d)$: $\ell(a, b) \cdot \ell(c, d) = (2^a3^b)(2^c3^d) = 2^{a+c}3^{b+d} = \ell(a + c, b + d) = \ell\left((a, b) + (c, d)\right)$. Therefore $Z \times Z$ is an additive group isomorphic to the multiplicative group P. The identity element in $Z \times Z$ is of course $(0, 0)$ and the inverse of (z_1, z_2) is $(-z_1, -z_2)$.

The lattice points which $Z \times Z$ embodies enables us to think geometrically about Pythagorean tones and intervals. However, $Z \times Z$ and P are not identical since we can order the elements of P, that is, given any two elements of P we can decide which is greater, according to the ordinary notion of order in arithmetic.

---

[6]Cf. Eric Regener *Pitch Notation and Equal Temperament* University of California 1973. Regener gives a treatment of musical notation in a mathematical setting.

Similarly we can consider $\hat{P}$ in this two-dimensional setting, represented by the positive quadrant of the lattice $Z \times Z$. This two-dimensional version of $\hat{P}$ was known in antiquity and is referred to as the Nicomachus Triangle. Of course, $\hat{P}$ can also be ordered in a linear sequence. The interplay between the linear (one-dimensional) sequence of $\hat{P}$ and the two-dimensional display of $\hat{P}$ provides the basis for much of what follows. The accompanying diagrams show $\hat{P}$ in a two-dimensional array and in a linear sequence. The reader will note that every pure power of 3 presents a new tone, while every pure power of 2 re-presents the origin F_0 in different registers.



Here the elements of $\hat{P}$ are arranged linearly, each element together with its corresponding note and pair of exponents:

$\hat{P}$:	1	2	3	4	6	8	9	12	16	18
$Z \times Z$:	(0, 0)	(1, 0)	(0, 1)	(2, 0)	(1, 1)	(3, 0)	(0, 2)	(2, 1)	(4, 0)	(1, 2)
Tone:	F_0	F_1	C_1	F_2	C_2	F_3	G_3	C_3	F_4	G_4
24	27	32	36	48	54	64	72	81	96	108
(3,1)	(0, 3)	(5, 0)	(2, 2)	(4, 1)	(1.3)	(6, 0)	(3, 2)	(0, 4)	(5, 1)	(2, 3)
C_4	D_4	F_5	G_5	C_5	D_5	F_6	G_6	A_6	C_6	D_6
128	144	162	192	216	243	256	288	324	384	432
(7,0)	(4, 2)	(1, 4)	(6, 1)	(3, 3)	(0, 5)	(8, 0)	(5, 2)	(2, 4)	(7, 1)	(4, 3)
F_7	G_7	A_7	C_7	D_7	E_7	F_8	G_8	A_8	C_8	D_8

486	512	576	648	729	768	864	972	1,024	1,152	1,296
(1,5)	$(9,0)$	$(6,2)$	$(3,4)$	$(0,6)$	$(8,1)$	$(5,3)$	$(2,5)$	$(10,0)$	$(7,2)$	$(4,4)$
E$_{8}$	F$_{9}$	G$_{9}$	A$_{9}$	B$_{9}$	C$_{9}$	D$_{9}$	E$_{9}$	F$_{10}$	G$_{10}$	A$_{10}$

1,458	1,536	1,728	1,944	2,048	2,187	2,304	2,592	2,916	3,072	3,456
(1,6)	$(9,1)$	$(6,3)$	$(3,5)$	$(11,0)$	$(0,7)$	$(8,2)$	$(5,4)$	$(2,6)$	$(10,1)$	$(7,3)$
B$_{10}$	C$_{10}$	D$_{10}$	E$_{10}$	F$_{11}$	F$\sharp_{11}$	G$_{11}$	A$_{11}$	B$_{11}$	C$_{11}$	D$_{11}$

3,888	4,096	4,374	4,608	5,184	5,832	6,144	6,561	6,912	7,776	8,192
(4,5)	$(12,0)$	$(1,7)$	$(9,2)$	$(6,4)$	$(3,6)$	$(11,1)$	$(0,8)$	$(8,3)$	$(5,5)$	$(13,0)$
E$_{11}$	F$_{12}$	F$\sharp_{12}$	G$_{12}$	A$_{12}$	B$_{12}$	C$_{12}$	C$\sharp_{12}$	D$_{12}$	E$_{12}$	F$_{13}$

8,748	9,216	10,368	11,664	12,288	13,122	13,824	15,552	16,384	17,496
$(2,7)$	$(10,2)$	$(7,4)$	$(4,6)$	$(12,1)$	$(1,8)$	$(9,3)$	$(6,5)$	$(14,0)$	(3,7)
F$\sharp_{13}$	G$_{13}$	A$_{13}$	B$_{13}$	C$_{13}$	C$\sharp_{13}$	D$_{13}$	E$_{13}$	F$_{14}$	F$\sharp_{14}$

18,432	19,683	20,736	23,328	24,576	26,244	27,648	31,104	32,768
$(11,2)$	$(0,9)$	$(8,4)$	$(5,6)$	$(13,1)$	$(2,8)$	$(10,3)$	$(7,5)$	$(15,0)$
G$_{14}$	G$\sharp_{14}$	A$_{14}$	B$_{14}$	C$_{14}$	C$\sharp_{14}$	D$_{14}$	E$_{14}$	F$_{15}$

The Linear Sequence of $\hat{P}$

2.4. **Definition of the Region** $R_{(A,B)}$**.** At this point the framework for the theory of well-formed Pythagorean scales is in place. We proceed to offer a definition of the *region* $R_{(A,B)}$, an operational definition in the sense that it gives a procedure for constructing all regions, making use of the set $\hat{P}$ and two properties, *connectedness* and *complementarity*.

Let S be a subset of $\hat{P}$ and let $\hat{p} \in \hat{P}$. S is a *connected subset of* $\hat{P}$ if, whenever $s_a < \hat{p} < s_b$, where s_a and $s_b \in S$, it is true that $\hat{p} \in S$. The connected subsets of $\hat{P}$ may be regarded simply as uninterrupted sections of the linear sequence of $\hat{P}$. A pair of elements that are adjacent in the linear sequence will be called *connected elements*.

Suppose that 2^A and 3^B are connected elements in $\hat{P}$. By inspection we know that connected pairs of pure powers exist. (For example, $3 = 3^1$ and $4 = 2^2$ are connected, and so are 27 and 32, i.e., 3^5 and 2^5.) Let $S_{(A,B)} = \{\hat{p} \mid \hat{p} < 2^A \text{ and } \hat{p} < 3^B\}$. $S_{(A,B)}$ is then the connected subset of $\hat{P}$ which contains all of $\hat{P}$ up to, but not including 2^A or 3^B. It follows that for any $s = 2^a 3^b \in S_{(A,B)}$, $a < A$ and $b < B$, i.e., $a \in Z_A$ and $b \in Z_B$, and $S_{(A,B)}$ is a subset of $G_{(A,B)} = \{2^a 3^b \mid a \in Z_A \text{ and } b \in Z_B\}$. $G_{(A,B)}$ is the set of all factors of the number $2^{A-1}3^{B-1}$. The element $2^{A-1}3^{B-1} \in G_{(A,B)}$ and is the least common multiple not only of the elements in $G_{(A,B)}$ but also of those in $S_{(A,B)}$, because 2^{A-1} and 3^{B-1} both belong to $S_{(A,B)}$ and have no common factors. In general, if g and $g\prime$ belong to $G_{(A,B)}$, and $gg\prime = 2^{A-1}3^{B-1}$ we say that g and $g\prime$ are *complementary* factors of $2^{A-1}3^{B-1}$, or that they are *complements* of each other.

The elements of $G_{(A,B)}$ are ordered from 1 to $2^{A-1}3^{B-1}$ in the following way:

$$g_1(= 1) < g_2 < \cdots < g_{AB}(= 2^{A-1}3^{B-1})$$

The complements are ordered symmetrically, as follows:

$$g\prime_1(= 2^{A-1}3^{B-1}) > g\prime_2 > \cdots > g\prime_{AB}(= 1)$$

This property is of course very general, in that the factors of any positive integer are ordered symmetrically in this way. Suppose n_i and n_j are any two factors of some positive integer N, with $n_j > n_i$. Then $n_j/n_1 > 1$. If $n_i\prime$ and $n_j\prime$ are the respective complementary factors of N, $n_i n_i\prime = n_j n_j\prime = N$. Then $n_i\prime/n_j\prime = n_j/n_i$. Therefore $n_i\prime/n_j\prime > 1$, or $n_i\prime > n_j\prime$ so the order of the complements is the opposite of the order of the factors. In our example,

$$g_1 < \cdots < g_{AB} = g_1\prime > \cdots > g_{AB}\prime$$

We define a *set of complements* in $G_{(A,B)}$ to be a subset R of $G_{(A,B)}$ such that, if $r \in R$, then $r\prime \in R$. $G_{(A,B)}$ itself of course satisfies this definition.

The *region* $R_{(A,B)}$ is defined to be the largest connected set of complements in $G_{(A,B)}$. $S_{(A,B)}$ is the largest connected subset of $G_{(A,B)}$, so

$$R_{(A,B)} \subset S_{(A,B)} \subset G_{(A,B)}.$$

Since $R_{(A,B)}$ is a set of complements it displays the same symmetry as does all of $G_{(A,B)}$.

It is important to keep in mind that this definition of $R_{(A,B)}$ starts from the premise that 2^A and 3^B are connected. It would have been possible to define sets G, S, and R for arbitrary positive integers A and B, but for $R_{(A,B)}$ to have the structure of a region we require that to begin with 2^A and 3^B be connected. We refer to the ratio of these connected pure powers as the *prime ratio* $\rho(A, B)$, where $\rho(A, B) = \frac{2^A}{3^B}$ or $\frac{3^B}{2^A}$, whichever is greater than 1. In the case $2^A > 3^B$ we say $\rho(A, B)$ is positive, and in the case $3^B > 2^A$, $\rho(A, B)$ is negative. If $\rho(A, B)$ is positive, it is the P element $2^A 3^{-B}$; if negative, the P element $2^{-A}3^B$. Since $\rho(A, B)$ is an element of P, it represents an interval, so each region is determined by an interval.

In the case $(A, B) = (1, 0)$, $2^1 = 2$ and $3^0 = 1$ are connected in $\hat{P}$, but $R_{(A,B)} = S_{(A,B)} = G_{(A,B)} = \emptyset$. To avoid trivial complications later we will add the further restriction that $R_{(A,B)}$ be non-empty. As long as $B > 1$, $R_{(A,B)} \neq \emptyset$: Certainly $s = 2^0 3^{B-1}$ and $s = 2^{A-1}3^0$ are in $S_{(A,B)}$ as noted above. Then at least these elements (or this element, if $A = B = 1$) belong to $R_{(A,B)}$. If $B > 1$, then there are elements in $G_{(A,B)}$ that are not in $S_{(A,B)}$, hence elements in $S_{(A,B)}$ not in $R_{(A,B)}$.

Assuming 2^A and 3^B are connected in $\hat{P}$, the region $R_{(A,B)}$ is defined to be the largest non-empty connected set of complements in $G_{(A,B)}$. Table 2.4 gives the situation for the first four prime ratios:

(A, B)	$\rho(A, B)$	Interval	$G_{(A,B)}$	$S_{(A,B)}$	$R_{(A,B)}$
$(1, 0)$	$2^1 3^0 = 2/1$	Octave	$\emptyset$	$\emptyset$	$\emptyset$
$(1, 1)$	$2^{-1}3^1 = 3/2$	Fifth	$\{1\}$	$\{1\}$	$\{1\}$
$(2, 1)$	$2^2 3^{-1} = 4/3$	Fourth	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$
$(3, 2)$	$2^{-3}3^2 = 9/8$	Whole Step	$\{1, 2, 3, 4, 6, 12\}$	$\{1, 2, 3, 4, 6\}$	$\{2, 3, 4, 6\}$

Table 2.4

In order to show the musical significance of regions we develop the next two examples, the Structural Region and the Pentatonic Region.

The next connected pair of pure powers, that is, the next prime ratio, is $\rho(5,3) = 32/27$, which corresponds to the Pythagorean minor third.

$$G_{(5,3)} = \left\{2^a 3^b \mid a \in Z_5 \text{ and } b \in Z_3\right\}$$

9	18	36	72	144
3	6	12	24	48
1	2	4	8	16

$S_{(5,3)} \subset G_{(5,3)}$

$G_{(5,3)}$ consists of the 15 factors of $144 = 2^4 3^2$ and contains the 11 elements of $S_{(5,3)}$ plus the four elements 144 ($2^4 3^2$), 72 ($2^3 3^2$), 48 ($2^4 3^1$), and 36 ($2^2 3^2$). The complements of these four are the first four elements of $S_{(5,3)}$, 1 ($= 2^0 3^0$), 2 ($= 2^1 3^0$), 3 ($= 2^0 3^1$), and 4 ($= 2^2 3^0$). The remaining 7 elements form $R_{(5,3)}$, the largest connected subset of complements in $G_{(5,3)}$.

9	18	36	72	144
3	6	12	24	48
1	2	4	8	16

$R_{(5,3)} \subset G_{(5,3)}$

6	8	9	12	16	18	24
$(1,1)$	$(3,0)$	$(0,2)$	$(2,1)$	$(4,0)$	$(1,2)$	$(3,1)$
C_2	F_3	G_3	C_3	F_4	G_4	C_4

This region embraces the key proportion of Greek musical theory $6 : 8 :: 9 : 12$ taken over two octaves. This proportion has the property that 8 and 9 are the harmonic and arithmetic means, respectively, between 6 and 12. Musically, of course, this proportion corresponds to the structure, Tonic, Sub-Dominant, Dominant, and Tonic, which forms the framework of fixed tones for many scales throughout the world. All of this leads us to call $R_{(5,3)}$ the *Structural Region* and the elements of $R_{(5,3)}$ the *structurals*.

Ancient Greek theory estabished the modes within the "dis-diapason," a distance of two octaves, and there is a tradition which maintains a two octave span for this proportion.[7] The fact that the span of the Structural Region $R_{(5,3)}$ is exactly two octaves is a feature peculiar to this region, as we will see, but it provides $R_{(5,3)}$ with a particularly high degree of symmetry, neatly dividing it into two "well-formed" symmetrical scales.

[7]Cf. *Schola Enchiriades*

Inspection of the linear sequence of $\hat{P}$ shows that the next prime ratio is $\rho(8,5) = 2^8 3^5 = 256/243$, the *limma* or diatonic half-step. The figure below shows the 12 elements of the region $R_{(8,5)}$ as a subset of $G_{(8,5)}$:

81	162	324	648	1296	2592	5184	10368
27	54	108	216	432	864	1728	3456
9	18	36	72	144	288	576	1152
3	6	12	24	48	96	192	324
1	2	4	8	16	32	64	128

$R_{(8,5)} \subset G_{(8,5)}$

The 12 Elements of the Region $R_{(8,5)}$

48	54	64	72	81	96	108	128	144	162	192	216
$(4,1)$	(1.3)	$(6,0)$	$(3,2)$	$(0,4)$	$(5,1)$	$(2,3)$	(7,0)	$(4,2)$	$(1,4)$	$(6,1)$	$(3,3)$
C_5	D_5	F_6	G_6	A_6	C_6	D_6	F_7	G_7	A_7	C_7	D_7

Giving note-names to the first six elements of $R_{(8,5)}$ we see the region begins C_5, D_5, F_6, G_6, A_6, C_6: the pentatonic scale, or more precisely, one of the modes of the anhemitonic pentatonic scale, which is an important scale in the music of many cultures. $R_{(8,5)}$ contains all modal varieties of the anhemitonic pentatonic scale, and we refer to it as the *Pentatonic Region*.

The Pentatonic Region motivates our definition of a *well-formed Pythagorean scale*. We recall that in general a scale may be represented by a sequence of tones spanning one octave, and so a general scale in Pythagorean tuning may be represented by a sequence of P numbers $2^{x_1}3^{y_1} < 2^{x_2}3^{y_2} < \cdots < 2^{(x_1+1)}3^{y_1}$. Again, since scale-step intervals determine the type of scale, we lose nothing by translating or transposing from P to $\hat{P}$. (Simply multiply each $2^{x_i}3^{y_i}$ by the lowest common denominator of all of the elements of the scale to obtain a new sequence of $\hat{P}$ elements $\hat{p}_1 < \hat{p}_2 < \cdots < \hat{p}_n = 2\hat{p}_1$. If the greatest common factor (GCF) of the integers $\hat{p}_1, \hat{p}_2, \cdots < \hat{p}_n$ is 1, we say the Pythagorean scale is *reduced*. If a scale is not reduced, it can be reduced by dividing each of its elements by the GCF. We say that a Pythagorean scale is *well-formed* if it is connected in $\hat{P}$ and is a subset of a region, or if when reduced it is a connected subset of a region. The different well-formed Pythagorean scales associated with a region $R_{(A,B)}$ comprise a *modal system* of well-formed Pythagorean scales.

These definitions and the example of the Pentatonic Region lead us to ask if there exists a Diatonic Region containing all modal systems of well-formed diatonic scales. In fact, $2^{11} = 2{,}048$ and $3^7 = 2{,}187$ are connected in $\hat{P}$, with $\rho(11,7) = 2^{-11}3^7$ representing the interval from F to F♯, the chromatic half-step. The region associated with this interval, $R_{(11,7)}$ is such a Diatonic Region. It contains all possible modes of the diatonic scale, since it spans two octaves and a major third of the diatonic scale from C_8 to E_{10}:

384	432	486	512	576	648	729	768	864	972	1024	1152	1296	1458	1536	1728	1944
C_8	D_8	E_8	F_9	G_9	A_9	B_9	C_9	D_9	E_9	F_{10}	G_{10}	A_{10}	B_{10}	C_{10}	D_{10}	E_{10}

The Diatonic Region $R_{(11,7)}$

The next example of a system of well-formed scales is the set of chromatic scales associated with the Chromatic Region $R_{(19,12)}$. However, we have just about reached the limit of the usefulness of our ad hoc approach to the determination of regions since the numbers are getting very large and the sequence of $\hat{P}$ numbers very long. In this case, $2^{19} = 524{,}288$ and $3^{12} = 531{,}441$ are connected and $\rho(19,12) = 2^{-19}3^{12}$ corresponds to the interval from F to $E_\sharp$, the Pythagorean comma.

The next connected pair is 3^{17} and 2^{27} $\rho(27,17) = 2^{27}3^{-17} = \frac{134{,}217{,}728}{129{,}140{,}163}$. However, this region also gives rise to an important system, the Arabic division of the octave into 17 parts.[8]

This is a remarkable correlation, to say the least. We hope the musician will find it interesting to reflect on the pattern of growth displayed in the formation of these scales, as the octave is progressively filled in, or, from a slightly different point of view, on the hierarchical arrangement of these scales.

So far we know very little about regions beyond what is contained in the definition, and, as we have remarked, the practical usefulness of the operational definition even for computing higher order regions is limited. Moreover, the correlation between regions and significant musical scales together with the mathematical properties, namely connectedness and symmetry, which regions are known to possess, motivate a mathematical treatment of general regions. The aim of the next section will be to prove theorems which will yield a deeper understanding of regions and well-formed scales.

---

3. Regions Formalized

The next central section is highly technical. In fact, it can be read as a work of pure mathematics. Nonetheless, when the scaffolding of logical argument and abstract symbolism is cleared away, what remains is a new way of seeing and thinking about the scales we have been considering. As with any abstract mathematical text, the best way to get past the symbols to the essential ideas is to refer to particular examples or counterexamples. For instance, the reader can conveniently test many of the theorems with the example of the Structural Region, $R_{(5,3)}$.

To begin the mathematical investigation of the set of regions $R_{(A,B)}$ it will be useful to develop some preliminary notions from elementary group theory and the theory of continued fractions.

3.1. **Group Theory.** First of all, the set $G_{(A,B)}$ may be made into a group by defining an appropriate multiplication. (For the purposes of this discussion it is not necessary to put any special conditions on the ordered pair (A, B). Thus $G_{(A,B)}$

[8] Curt Sachs *The Rise of Music in the Ancient World: East and West* pp. 279-280. W.W. Norton 1943

is the set of all factors of 2^{A-1} and 3^{B-1} where A and B are any two positive integers.)

To see how $G_{(A,B)}$ becomes a group first consider $Z_N = \{0, 1, \ldots N-1\}$ where N is any positive integer. Z_N may be provided with an operation "addition modulo N" as follows: Let $z_1, z_2 \in Z_N$. Then simple division, sometimes called Euclid's Theorem, says that $z_1 + z_2 = Nq + r$ for unique integers q and r where $0 \leq r < N$, i.e., $r \in Z_N$. Then we say $(z_1 + z_2)_{\text{ mod } N} \equiv r$, or, if the context is clear, $z_1 + z_2 = r$, where the symbol "+" is understood to represent addition modulo N. In this way a closed operation is defined for Z_N and the reader can readily determine that Z_N provided with this operation is a finite group with N elements. Z_N is called the *cyclic* group modulo N because the whole group can be generated by a single element. Any element k relatively prime to N repeatedly added to itself will generate he whole group, returning to the element k after one cycle. The group may be visualized as a clock with 0 at 12 o'clock.

Then Z_A and Z_B are groups and $Z_A \times Z_B = \{(a,b) \mid a \in Z_A, b \in Z_B\}$ can be considered a group with addition defined as follows:

Let (a,b) and $(c,d) \in Z_A \times Z_B$.

$$(a,b) + (c,d) = ((a+c)_{\text{ mod } A}, (b+d)_{\text{ mod } B}).$$

Again, clearly $Z_A \times Z_B$ is a group with the operation defined.

There is a natural one-to-one correspondence between $G_{(A,B)}$ and $Z_A \times Z_B$ defined by $\lambda : Z_A \times Z_B \to G_{(A,B)} : (a,b) \to 2^a 3^b$ and $\lambda^{-1} : G_{(A,B)} :\to Z_A \times Z_B : 2^a 3^b \to (a,b)$. $G_{(A,B)}$ can be made into a group and λ into an isomorphism of groups by defining multiplication in $G_{(A,B)}$ by $(2^a 3^b)(2^c 3^d) = (2^{(a+c)_{\text{ mod } A}} 3^{(b+d)_{\text{ mod } B}})$. Then $\lambda((a,b) + (c,d)) = \lambda(a,b) \cdot \lambda(c,d)$ for all $(a,b), (c,d) \in Z_A \times Z_B$:

$$\begin{aligned} \lambda((a,b) + (c,d)) &= \lambda((a+c)_{\text{ mod } A}, (b+d)_{\text{ mod } B}) \\ = 2^{(a+c)_{\text{mod} A}} 3^{(b+d)_{\text{mod} B}} &= (2^a 3^b)(2^c 3^d) = \lambda(a,b) \cdot \lambda(c,d) \end{aligned}$$

The group structure of $Z_A \times Z_B$ is then carried over to $G_{(A,B)}$ and λ is an isomorphism between the two groups.

If we add the condition that A and B are relatively prime it is also true that $Z_A \times Z_B$ and Z_{AB} are isomorphic, where Z_{AB} is the cyclic group modulo AB.

To show that $Z_A \times Z_B \cong Z_{AB}$ when A and B are relatively prime we make use of a lemma which will be useful later on so we label it

Lemma 1.

$$(A(y_{\text{ mod } B}) + B(x_{\text{ mod } A})) \equiv (Ay + Bx)_{\text{ mod } AB}$$

Proof. We will show that both sides of the equation are congruent to $Ay\prime + Bx\prime_{\text{ mod } AB}$ where $x = x\prime + As$ and $y = y\prime + Bt$ for some integers s and t and such that $0 \leq x\prime < A$, $0 \leq y\prime < B$:

$$\begin{aligned} (A(y_{\text{ mod } B}) + B((x_{\text{ mod } A})) &\equiv (A((y\prime + Bt)_{\text{ mod } B}) + B(((x\prime + As)_{\text{ mod } A}))_{\text{ mod } AB} \\ &\equiv (Ay\prime + Bx\prime)_{\text{ mod } AB} \end{aligned}$$

Also,

$$\begin{aligned} (Ay + Bx)_{\text{ mod } AB} &\equiv (A(y\prime + Bt) + B(x\prime + As))_{\text{ mod } AB} \\ \equiv (Ay\prime + Bx\prime + AB(t+s))_{\text{ mod } AB} &\equiv (Ay\prime + Bx\prime)_{\text{ mod } AB} \end{aligned}$$

Therefore $(A(y_{\text{ mod } B}) + B((x_{\text{ mod } A})) \equiv (Ay + Bx)_{\text{ mod } AB}$.

□

Now consider the mapping $\mu : Z_A \times Z_B \to Z_{AB} : (x, y) \to (Ay + Bx)_{\bmod AB}$, where it is understood that addition on the left-hand side is in $Z_A \times Z_B$, and addition on the right-hand side is in Z_{AB}, i.e., addition modulo AB.

Let $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ be elements of $Z_A \times Z_B$. Then

$$\begin{aligned}
\mu(z_1 + z_2) &= \mu\left((x_1, y_1) + (x_2, y_2)\right) = \mu\left((x_1 + x_2)_{\bmod A}, (y_1 + y_2)_{\bmod B}\right) \\
&= (A((y_1 + y_2)_{\bmod B}) + B((x_1 + y_1)_{\bmod A}))_{\bmod AB} \\
&\equiv (A(y_1 + y_2) + B(x_1 + y_1))_{\bmod AB} \text{ (by Lemma 1)} \\
&\equiv ((Ay_1 + Bx_1) + (Ay_2 + Bx_2))_{\bmod AB} \\
&\equiv ((Ay_1 + Bx_1)_{\bmod AB} + (Ay_2 + Bx_2)_{\bmod AB})_{\bmod AB} \\
&= (\mu(z_1 + z_2))_{\bmod AB} = \mu(z_1) + \mu(z_2)
\end{aligned}$$

Therefore $\mu(z_1 + z_2) \equiv \mu(z_1) + \mu(z_2)$ and since both sides of the equation are elements of the set Z_{AB}, in fact we have $\mu(z_1 + z_2) = \mu(z_1) + \mu(z_2)$, bearing in mind the different meanings attached to the symbol '+' on the right and left sides of the equation.

The condition that A and B are relatively prime comes into play in demonstrating that μ is one-to-one and onto, which is what remains to be shown for μ to be an isomorphism.

For μ to be one-to-one means that if $z_1 \neq z_2 \in Z_A \times Z_B$, then $\mu(z_1) \neq \mu(z_2)$. Suppose on the contrary that there exist two distinct elements of $Z_A \times Z_B$, $z_1 \neq z_2$, such that $\mu(z_1) = \mu(z_2)$. If $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$, then either $(x_1 - x_2)_{\bmod A} \neq 0$, or $(y_1 - y_2)_{\bmod A} \neq 0$ since $z_1 \neq z_2$. Let $s = (x_1 - x_2)_{\bmod A}$, $t = (y_1 - y_2)_{\bmod B}$. Then $0 \leq s < A$, $0 \leq t < B$, and at least one of them is not zero.

Now $\mu(z_1) - \mu(z_2) = 0$, so $\mu(z_1 - z_2) = 0$ and $z_1 - z_2 = ((x_1 - x_1)_{\bmod A}, (y_1 - y_2)_{\bmod B}) = (s, t)$. Thus $\mu(s, t) = 0$, i.e., $(At + Bs)_{\bmod AB} = 0$, or $At + Bs = ABk$ for some integer k. But then $\frac{At}{B} + s = Ak$ and $t + \frac{Bs}{A} = Bk$, meaning that both $\frac{At}{B}$ and $\frac{Bs}{A}$ are integers. But $t < B$ and $s < A$, so it is impossible for both to be integers. Therefore $\mu(z_1) \neq \mu(z_2)$ whenever $z_1 \neq z_2$, and μ is a one-to-one map.

Ontoness of μ is comparatively easy: Both $Z_A \times Z_B$ and Z_{AB} have AB elements, so it is enough to know that μ is one-to-one to also know that it is onto Z_{AB}.

Putting all of the above together we see that μ satisfies the requirements for an isomorphism of groups, and so for A and B relatively prime, $Z_A \times Z_B \cong Z_{AB}$.

This result will be of such importance that we spell out what it means, at the risk of belaboring the subject. That μ is one-to-one and onto Z_{AB} means that every element of Z_{AB} can be expressed uniquely in the form $Ay + Bx$ for some $(x, y) \in Z_A \times Z_B$, and there exists an inverse mapping $\mu^{-1} : Z_{AB} \to Z_A \times Z_B$.[9] Finally, since $Z_A \times Z_B \cong Z_{AB}$, Z_{AB} is a cyclic group when A and B are relatively prime.

Now restricting (A, B) to values such that $\rho(A, B)$ is defined, i.e., where 2^A and 3^B are connected, we ask if necessarily A and B are relatively prime. Certainly this has been true in the examples we have encountered so far.

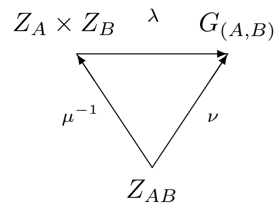
Let $2^A > 3^B$. Suppose 2^A and 3^B are connected but A and B are not relatively prime i.e., there is some integer $n > 1$ such that $A = an$ and $B = bn$ for integers

[9] This fact is closely related to the celebrated Chinese Remainder Theorem of number theory.

A and B. Then $2^a > 3^b$ since $3^b > 2^a$ implies $3^B = 3^{bn} > 2^{an} = 2^A$. Therefore, $2^A = 2^{an} > 2^{a(n-1)}3^b \geq 2^a 3^{b(n-1)} > 3^{bn} = 3^B$, since $n > 1$ and substituting 3^b for 2^a decreases the value of the $\hat{P}$ number. So 2^A and 3^B could not be connected after all. The same argument obtains in the case $3^B > 2^A$.

Therefore, for 2^A and 3^B to be connected, a necessary condition is that A and B be relatively prime. (It is not a sufficient condition, of course. For example, $3^4 = 81$ and $2^7 = 128$ have between them the $\hat{P}$ numbers $2^5 3^1 = 96$ and $2^2 3^3 = 108$.) Putting all of the preceding together, when $\rho(A, B)$ is defined, $G_{(A,B)} \cong Z_A \times Z_B \cong Z_{AB}$ and we can define the isomorphism $\nu = \lambda\mu^{-1}$, where

$$\nu : Z_{AB} \to G_{(A,B)} : z \xrightarrow{\mu^{-1}} (x, y) \xrightarrow{\lambda} 2^x 3^y$$



In the principal theorem we will prove characterizing the region we will examine the effect of these isomorphisms on the elements of $R_{(A,B)}$, as they are mapped into the various versions of the cyclic group.

3.2. **Continued Fractions.** The region $R_{(A,B)}$ is defined in terms of the connected pair of elements 2^A and 3^B. So far the only information we have about these numbers is that A and B must be relatively prime. We do not know anything about the ordered pairs as such, how to compute them, even whether or not the number of such pairs is infinite. The branch of mathematics which provides the tools for answering these questions is the branch of number theory referred to as the theory of continued fractions. Therefore, we will give a brief treatment of continued fractions, referring the reader unfamiliar with the topic to introductory texts in number theory for proofs and more detailed discussion.[10]

A finite continued fraction is a number of the form

$$t_0 + \cfrac{1}{t_1 + \cfrac{1}{t_2 + \cfrac{1}{t_3 + \cfrac{1}{\ddots + \cfrac{1}{t_n}}}}} \tag{3.1}$$

where t_1, $t_1 ... t_n$ are positive.

For example,

[10]C.D. Olds *Continued Fractions* Random House 1963; Hardy and Wright *An Introduction to the Theory of Numbers* Oxford University Press 1979.

$$\frac{45}{16} = 2 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{3}}} \tag{3.2}$$

For brevity, the continued fraction is notated $[2; 1, 4, 3]$, or generally, $[t_0; t_1, t_2, \ldots, t_n]$. The t_i are called the *partial quotients* of the continued fraction.

We develop some aspects of the continued fraction, beginning with a way to conveniently calculate its value.

Given a continued fraction $[t_0; t_1, t_2, \ldots, t_n]$, consider the sequence of continued fractions $c_0 = [t_0]$, $c_1 = [t_0; t_1]$, $c_2 = [t_0; t_1, t_2]$, $\ldots$, $c_n = [t_0; t_1, t_2, \ldots, t_n]$. The c_i are called the *convergents* of the continued fraction $[t_0; t_1, t_2, \ldots, t_n]$.

Theorem 1. *Continued Fraction Theorem 1*

Let $a_{-2} = 0$, $b_{-2} = 1$, $a_{-1} = 1$, $b_{-1} = 0$ *and define recursively* $a_i = t_i a_{i-1} + a_{i-2}$ *and* $b_i = t_i a_{i-1} + b_{i-2}$. *Then* $c_i = \frac{a_i}{b_i}$ *for* $i = 0, 1, \ldots n$.

Proof. The proof is by induction. For $i = 0$, $c_0 = [t_0] = t_0$ and

$$\begin{aligned} a_0 &= t_0 \cdot 1 + 0 = t_0 \\ b_0 &= t_0 \cdot 0 + 1 = 1 \end{aligned}$$

so $c_0 = \dfrac{a_0}{b_0}$ and the theorem holds in this case.

For $i = 1$, $c_1 = [t_0; t_1] = t_0 + \dfrac{1}{t_1} = \dfrac{t_0 t_1 + 1}{t_1}$ Now

$$\begin{aligned} a_1 &= t_1 a_0 + a_{-1} &&= t_0 t_1 + 1 \\ b_1 &= t_1 b_0 + b_{-1} = t_1 \cdot 1 + 0 &&= t_1 \end{aligned}$$

and again the theorem holds. Now we know the formula holds in the first two cases. We assume it holds for the first k cases, and we show it then must hold in the $(k+1)$th case:

Note that c_k and c_{k+1} differ only in that in c_{k+1} the term t_k is replaced by $t_k + \dfrac{1}{t_{k+1}}$

$$\begin{aligned} c_k &= [t_0, t_1, t_2, \ldots, t_k] \\ c_{k+1} &= [t_0, t_1, t_2, \ldots, (t_k + \frac{1}{t_{k+1}})] \end{aligned}$$

So c_{k+1} can be written as a continued fraction with k terms and the induction hypothesis can be applied. (Changing t_k to t_{k+1} has no effect on the value of the previous c_{k-1} and c_{k-2}, which depend only on t_0, t_1,$\ldots$ t_{k-1}.) Then

$$c_{k+1} = \frac{\left(t_k + \frac{1}{t_{k+1}}\right) a_{k-1} + a_{k-2}}{\left(t_k + \frac{1}{t_{k+1}}\right) b_{k-1} + b_{k-2}} \quad \text{By the induction hypothesis}$$

$$= \frac{t_{k+1}\left(t_k a_{k-1} + a_{k-2}\right) + a_{k-1}}{t_{k+1}\left(t_k b_{k-1} + b_{k-2}\right) + b_{k-1}} \quad \text{Multiplying numerator and denominator by } t_{k+1} \text{ and rearranging terms}$$

$$= \frac{t_{k+1} a_k + a_{k-1}}{t_{k+1} b_k + b_{k-1}} = \frac{a_{k+1}}{b_{k+1}} \quad \text{and the theorem is proven by induction}$$

□

In particular, if t_n is the last partial quotient, the value of the continued fraction is

$$c_n = \frac{a_n}{b_n} = \frac{t_n a_{n-1} + a_{n-2}}{t_n b_{n-1} + b_{n-2}}$$

The next result, also proved by induction, is a fundamental property of the convergents of a continued fraction.

Theorem 2. *Continued Fraction Theorem 2*
Let a_k, b_k and $[t_0; t_1 \ldots t_n]$ be defined as before. Then $a_{k+1}b_k - a_k b_{k+1} = (-1)^k$

Proof. The theorem is trivially true for $k = -2$, no matter what the particular fraction $[t_0; t_1 \ldots t_n]$ is, and for $k = -1$, $(t_0 \cdot 0) - (1 \cdot 1) = -1 = (-1)^{-1}$ and the theorem holds again. In the case of $k = 0$, $a_1 b_0 - a_0 b_1 = ((t_0 t_1 + 1) \cdot 1) - (t_0 t_1) = 1 = (-1)^0$.

Next we assume that the theorem holds for some $k - 1$ where $k < n - 1$, and prove that it then holds for k: By hypothesis $a_k b_{k-1} - a_{k-1} b_k = (-1)^{k-1}$. Since, by Theorem 1 $a_{k+1} = t_{k+1} a_k + a_{k-1}$ and $b_{k+1} = t_{k+1} b_k + b_{k-1}$, we can write

$$\begin{aligned} a_{k+1}b_k - a_k b_{k+1} &= \left(t_{k+1} a_k + a_{k-1}\right) b_k - a_k (t_{k+1} b_k + b_{k-1}) \\ &= t_{k+1} a_k b_k + a_{k-1} b_k - t_{k+1} a_k b_k - a_k b_{k-1} \\ &= a_{k-1} b_k - a_k b_{k-1} \\ &= (-1)(a_k b_{k-1} - a_{k-1} b_k) \\ &= (-1)(-1)^{k-1} \\ &= (-1)^k \end{aligned}$$

and Theorem 2 is proven by induction for k up to and including $n - 1$. □

So far the only condition we have placed on the partial quotients $t_0, t_1, \ldots t_n$ is that $t_1, \ldots t_n$ are positive numbers. From now on, we will assume that the partial quotients $t_0, t_1, \ldots t_n$ are all integers, with $t_1, \ldots t_n$ all strictly positive integers. Then $[t_0; t_1, \ldots t_n]$ is called a *simple* finite continued fraction, but since in the sequel we will only only be concerned with simple continued fractions, we will just use the term “continued fraction.” Naturally, the Continued Fraction Theorems 1 and 2 still hold.

Since now the partial quotients t_i are integers, it is also true that a_i and b_i are integers and Theorem 2 has the following important corollary:

Theorem 3. *Continued Fraction Theorem 3*
For any continued fraction $c_i = [t_0; t_1, \dots t_i]$, $c_i = a_i/b_i$ *is in lowest terms, i.e.,* a_i *and* b_i *are relatively prime integers.*

Proof. $c_i = a_i/b_i$ according to Theorem 1. Now, by Theorem 2, $a_{i+1}b_i - a_i b_{i+1} = (-1)^i$ so any integer which is a factor of b_i and a_i must be a factor of $(-1)^i$. Therefore, the greatest common factor of a_i and b_i is 1, that is, a_i and b_i are relatively prime. □

Another corollary of Theorem 2 is a statement about the differences between two successive convergents.

Theorem 4. *Continued Fraction Theorem 4*
$c_{k+1} - c_k = \dfrac{(-1)^k}{b_{k+1}b_k}$ *for* $k = 0, 1, \dots n-1$, *and* $|c_{k+1} - c_k| < \dfrac{1}{(b_k)^2}$

Proof. By Theorem 2, $a_{k+1}b_k - a_k b_{k+1} = (-1)^k$ and when c_k is defined, b_k and b_{k+1} are non-zero, so dividing through by $b_{k+1}b_k$ we have

$$\frac{a_{k+1}}{b_{k+1}} - \frac{a_k}{b_k} = \frac{(-1)^k}{b_{k+1}b_k}$$

i.e., $c_{k+1} - c_k = \dfrac{(-1)^k}{b_{k+1}b_k}$

Note that b_k is positive and since $b_0 = 1$, $b_1 = t_1 \geq 1$, and if $k > 1$, $b_k = t_k b_{k-1} + b_{k-2} \geq b_{k-1}$. (Observe that $t_k \geq 1$, $b_{k-2} \geq 1$ if $k > 1$.) So $b_k > b_{k-1}$ if $k > 1$, or $b_{k+1} > b_k$ if $k > 0$ so we have $|c_{k+1} - c_k| < \dfrac{1}{(b_k)^2}$.

□

This will be important when we consider infinite continued fractions.

Similarly, we have a result about the even convergents and the odd convergents:

Theorem 5. *Continued Fraction Theorem 5*
$c_{k+2} - c_k = \dfrac{t_k(-1)^k}{b_{k+2}b_k}$ *for* $k = 0, 1, 2, \dots n-2$.

That is, if k is even, $c_{k+2} > c_k$, and, if k is odd, $c_{k+2} < c_k$.

Proof.

$$c_{k+2} - c_k = \frac{a_{k+2}}{b_{k+2}} - \frac{a_k}{b_k} = \frac{a_{k+2}b_k - a_k b_{k+2}}{b_{k+2}b_k}$$

Then substituting for a_{k+2} and b_{k+2} by

$$a_{k+2} = t_{k+2}a_{k+1} + a_k$$
$$b_{k+2} = t_{k+2}b_{k+1} + b_k$$

we have

$$\begin{aligned} c_{k+2} - c_k &= \frac{(t_{k+2}a_{k+1} + a_k)b_k - a_k(t_{k+2}b_{k+1} + b_k)}{b_{k+2}b_k} \\ &= \frac{t_{k+2}(a_{k+1}b_k - a_k b_{k+1}) + a_k b_k - a_k b_k}{b_{k+2}b_k} \\ &= \frac{t_{k+2}(-1)^k}{b_{k+2}b_k} \end{aligned}$$

□

Thus if k is even, $c_{k+2} - c_k$ is positive, i.e., $c_{k+2} > c_k$, and if k is odd, $c_{k+2} - c_k$ is negative, i.e., $c_{k+2} < c_k$. Putting the results of Theorems 4 and 5 together we can see that the even convergents are strictly increasing, the odd convergents are strictly decreasing, and every convergent c_k where $k \geq 2$ lies between the previous two convergents. Thus $c_0 < c_2 < c_1$; $c_2 < c_3 < c_1$; $c_2 < c_4 < c_3$; $c_4 < c_5 < c_3$. So $c_0 < c_2 < c_4 < \cdots < c_5 < c_3 < c_1$, and c_n is the greatest of the even convergents or the least of the odd convergents.

We remark that any finite continued fraction may be written with either an odd or an even number of terms:

if $t_n > 1$, $\dfrac{1}{t_n} = \dfrac{1}{(t_n - 1) + \dfrac{1}{1}}$ so $[t_0; t_1, \dots t_n] = [t_0; t_1, \dots (t_n - 1), 1]$,

and, if $t_n = 1$, $\dfrac{1}{(t_{n-1}) + \dfrac{1}{t_n}} = \dfrac{1}{(t_{n-1}) + \dfrac{1}{1}} = \dfrac{1}{(t_{n-1}) + 1}$

so $[t_0; t_1, \dots t_n] = [t_0; t_1, \dots (t_{n-1} + 1)]$.

Any finite continued fraction clearly represents a *rational* number h/k where $h, k \in \mathbb{Z}$: By Theorem 1, the continued fraction has the value $c_n = \dfrac{a_n}{b_n}$ and a_n and b_n are integers when the partial quotients are integers. It is also true that any rational number may be represented by a finite continued fraction which is unique except for the possible modification of the last term just mentioned. Since this result is not essential for our purposes we will not prove it, but merely illustrate the *continued fraction algorithm* with an example. A proof can be had by examining this algorithm and noting its close relation to the Euclidean algorithm, the repeated application of Euclid's theorem. For example, to express 45/16 as a continued fraction we proceed as follows:

$$\frac{45}{16} = 2 + \frac{13}{16},\ \frac{16}{13} = 1 + \frac{3}{13},\ \frac{13}{3} = 4 + \frac{1}{3},\ \frac{3}{1} = 3.$$

Therefore, $\frac{45}{16} = 2 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{3}}}$.

The continued fraction algorithm illustrated above may be applied to an irrational number, but in this case the process continues indefinitely, generating an infinite sequence of integers. In fact, though again we do not prove it, the continued fraction algorithm sets up a one-to-one correspondence between the set of all irrationals and the set of all infinite sequences of integers, $t_0, t_1, \dots t_n \dots$ where $t_1, t_2, \dots$ are positive.

We do show that any infinite sequence of integers $t_0, t_1, \dots t_n \dots$ with $t_1, t_2, \dots$ positive defines an infinite continued fraction $[t_0; t_1, \dots]$: If c_k is defined as before, the question is whether c_k approaches a limit as k increases without bound, and we appeal to the fundamental theorem of analysis that a bounded, monotonically increasing or decreasing sequences converges to a limit. Theorems 1 through 5 may be applied to the c_k, so we immediately have the result that the even convergents are strictly increasing and the odd convergents are strictly decreasing. Moreover, the even convergents are all less than any odd convergent, and vice versa, so both sequences are bounded, so both approach limits. The only thing to see is that they both approach the same limit. Let L_E be the limit of the even convergents and L_O be the limit of the odd convergents. If the limits were unequal, $L_O - L_E > 0$. Applying Theorem 4, we have $|c_{k+1} - c_k| < \frac{1}{{b_k}^2}$ for $k > 0$. Now one of $k, k+1$ is odd, one is even, so $|c_{k+1} - c_k| > L_O - L_E > 0$. Also by Theorem 4, $b_{k+1} > b_k$ if $k > 0$, and in fact, it is easy to see that b_k increases faster than k, so $\lim_{k\to\infty} \frac{1}{(b_{k+1})^2} = 0$.[11] That is, we may choose k large enough that the difference between $[c_{k+1}$ and $c_k]$ is less in absolute value than any real number, however small, and in particular large enough so that $|c_{k+1} - c_k| < L_O - L_E > 0$, contradicting the previous statement. Therefore, $L_O = L_E = x$, and we may write

Theorem 6. *Continued Fraction Theorem 6*
$x = [t_0; t_1, \ldots t_k \ldots]$.

We have just proved that the sequence of convergents c_k for an infinite continued fraction $[t_0; t_1, \ldots]$ approaches a limit x, which we can define to be the value of the continued fraction. So far we know that the subsequence of even convergents approaches x from below, and the subsequence of odd convergents approaches x from above. In fact, each convergent is closer to x than the one before. Again, since we do not need this result we will state it without proof.

A result which we *will* appeal to later is the following:

Theorem 7. *Continued Fraction Theorem 7*
If a_k/b_k is a convergent to the continued fraction whose value is x, there is no rational number p/q with $q \leq b_k$ that lies between x and a_k/b_k.

That is, if $a_k/b_k < p/q < x$, or $x < p/q < a_k/b_k$, then $q > b_k$. (A stronger result is possible, namely that any p/q which is *closer* to x than a_k/b_k must have denominator $q > b_k$.)

The proof requires a LEMMA: if $\frac{m}{n} < \frac{s}{t} < \frac{m'}{n'}$ and $mn' - m'n = -1$, where all letters represent positive integers, then $t > n$ and $t > n'$.

Proof. Proof of LEMMA:
$0 < \frac{m'}{n'} - \frac{s}{t} < \frac{m'}{n'} - \frac{m}{n}$.
So $0 < \frac{m't - n's}{n't} < \frac{m'n - mn}{n'n} = \frac{1}{n'n}$.
Thus $n'(m't - n's) < t$, $m't - n's$ is a positive integer, so we must have $n < t$.
Likewise, $0 < \frac{s}{t} - \frac{m}{n} < \frac{m'}{n'} - \frac{m}{n}$ so $0 < \frac{sn - mt}{tn} < \frac{m'n - n'm}{n'n} = \frac{1}{n'n}$.
Thus $n'(sn - mt) < t$, $(sn - mt)$ is a positive integer, and so $n' < t$. □

Proof. Proof of CF Theorem 7:
Assume p/q lies between x and a_k/b_k. The two cases are k is even or k is odd. If k is even,

$$\frac{a_k}{b_k} < x < \frac{a_{k-1}}{b_{k-1}} \text{ and } \frac{a_k}{b_k} < \frac{p}{q} < x$$

[11] $\lim_{k\to\infty} \frac{1}{{b_k}^2} = 0$ - NC

so
$$\frac{a_k}{b_k} < \frac{p}{q} < \frac{a_{k-1}}{b_{k-1}}$$
and by Theorem 2, $a_k b_{k-1} - a_{k-1} b_k = (-1)^{k-1} = -1$ so we may apply the lemma and $q > b_k$.

If k is odd, we have
$$\frac{a_{k-1}}{b_{k-1}} < x < \frac{a_k}{b_k}$$
and $a_{k-1} b_k - a_k b_{k-1} = -(a_k b_{k-1} - a_{k-1} b_k) = (-1)^{k-1} = -1$, and we again may apply the lemma, thus $q > b_k$.

□

3.2.1. *Semi-convergents.* If c_k is a convergent to a continued fraction, finite or infinite, and $t_k > 1$, we can form *semi-convergents* of the continued fraction by substituting in place of the term t_k the values $1, 2, \ldots (t_k - 1)$:
$$[t_0; t_1, \ldots t_{k-1}, 1], [t_0; t_1, \ldots t_{k-1}, 2], \ldots [t_0; t_1, \ldots t_{k-1}, (t_k - 1)].$$
These are defined to be the semi-convergents and by Theorem 1 they have the values $\frac{1 \cdot a_{i-1} + a_{i-2}}{1 \cdot b_{i-1} + b_{i-2}}$, $\frac{2 \cdot a_{i-1} + a_{i-2}}{2 \cdot b_{i-1} + b_{i-2}}$, $\frac{(t_k - 1) \cdot a_{i-1} + a_{i-2}}{(t_k - 1) \cdot b_{i-1} + b_{i-2}}$, respectively. For example, $[2; 1, 4]$ is a convergent to $45/16 = [2; 1, 4, 3]$ so $[2; 1, 1]$, $[2; 1, 2]$ and $[2; 1, 3]$ are semi-convergents to $45/16$. The complete list of convergents and semi-convergents for $45/16$, arranged into families associated with each partial quotient, would be

$[1]$	$[\underline{2}]$		
$[2; \underline{1}]$			
$[2; 1, 1]$	$[2; 1, 2]$	$[2; 1, 3]$	$[2; 1, \underline{4}]$
$[2; 1, 4, 1]$	$[2; 1, 4, 2]$	$[2; 1, 4, \underline{3}]$	

Theorem 7 can be extended to apply to semi-convergents as well. Since we will require this result and since semi-convergents are seldom encountered in elementary treatments of continued fractions we prove this final theorem.

Theorem 8. *Continued Fraction Theorem 8*

Consider $\frac{A}{B} = \frac{t \cdot a_k + a_{k-1}}{t \cdot b_k + b_{k-1}}$, $0 < t < t_{k+1}$. *Then A/B is a semi-convergent of the continued fraction $x = [t_0; t_1, \ldots]$. If p/q is a rational number that lies between A/B and x, $q > B$.*

Proof. We give a proof for the case when k is odd. Then $\frac{a_{k-1}}{b_{k-1}} < x < \frac{a_k}{b_k}$. $\frac{A}{B} = [t_0; t_1, \ldots t_k, t]$ is a continued fraction in its own right, an extension of $c_{k-1} = [t_0; t_1, \ldots t_{k-1}]$ and $c_k = [t_0; t_1, \ldots t_k]$, so it lies between c_{k-1} and c_k: $\frac{a_{k-1}}{b_{k-1}} < \frac{A}{B} < \frac{a_k}{b_k}$. c_{k+1} is an even convergent to x, so we have $\frac{a_{k+1}}{b_{k+1}} < x < \frac{a_k}{b_k}$. Applying the lemma of Theorem 7 we show that $\frac{A}{B}$ does not lie between c_{k+1} and c_k: $c_{k+1} = \frac{a_{k+1}}{b_{k+1}}$ is an even convergent to x, so we have $\frac{a_{k+1}}{b_{k+1}} < x < \frac{a_k}{b_k}$, and $a_{k+1} b_k - a_k b_{k+1} = (-1)^k = -1$ since k is assumed to be odd. Then we can apply the lemma, and if $\frac{a_{k+1}}{b_{k+1}} < \frac{A}{B} < \frac{a_k}{b_k}$, we would have $B > b_{k+1}$, but $b_{k+1} = t_{k+1} b_k + b_{k-1}$, $B =$

$tb_k + b_{k-1}$, $t < t_{k+1}$, so $B < b_{k+1}$, and therefore $\frac{A}{B}$ is forced between $\frac{a_{k-1}}{b_{k-1}}$ and $\frac{a_{k+1}}{b_{k+1}}$, thus $\frac{A}{B} < x$: $\frac{a_{k-1}}{b_{k-1}} < \frac{A}{B} < \frac{a_{k+1}}{b_{k+1}} < x$. Therefore $\frac{A}{B} < x < \frac{a_k}{b_k}$. Again, $\frac{A}{B} = [t_0; t_1, \dots t_k, t]$ is a continued fraction with the continued fraction preceeding it being $\frac{a_k}{b_k}$, so by Theorem 2, $Ab_k - Ba_k = (-1)^k = -1$, so we can apply the lemma of Theorem 7 again, and we must have $q > B$. The argument is the same if k is even, and we leave it to the reader. □

3.2.2. *The Golden Number.* The continued fraction $[1; 1, 1, 1, \dots]$ where $t_i = 1$ for all i has no semi-convergents, nor has $[0; 1, 1, 1, \dots]$. These are in a sense the simplest infinite continued fractions. We know that $[1; 1, 1, \dots]$ has a well-defined value x by Theorem 6, so $x - 1 = [0; 1, 1, 1, \dots]$. Then

$$x = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}, \quad x - 1 = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}$$

so x satisfies $x - 1 = 1/x$, that is x is a solution of $x^2 - x - 1 = 0$. Solving for x we have $x = \frac{\sqrt{5}+1}{2}$ and $x - 1 = \frac{1}{x} = \frac{\sqrt{5}-1}{2}$

The convergents to x are the quotients of successive numbers of the Fibonacci sequence: $1, 1, 2, 3, 5, 8, \dots \mathrm{f}_n = \mathrm{f}_{n-1} + \mathrm{f}_{n-2}$. Then the convergents to x are $\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \dots$. $\frac{\sqrt{5}+1}{2}$ is called the “golden number”, which appears in many contexts, both in pure mathematics and in discussions of beauty in art and nature. The proportion which gives rise to x, the golden section, has been used as a principle of construction by artists from Praxiteles to Stradivarius to Béla Bartók.

Recalling that the first five regions $R_{(A,B)}$ are associated with the ordered pairs, $(1, 1), (2, 1), (3, 2), (5, 3), (8, 5)$, we might be tempted to conclude that the regions correspond to convergents of the golden number. However, the Fibonacci sequence continues, $13, 21, 34, 55$, etc., and after $R_{(8,5)}$ the next regions are $R_{(11,7)}$, $R_{(19,12)}$, and $R_{(27,17)}$. This conjecture proved false, but perhaps such a correspondence exists with another continued fraction. The reader may have guessed that the value A/B approximates is $\log_2 3$, saying that 2^A and 3^B are close together, or $A/B \sim \log_2 3$. Therefore we would like to be able to compute the continued fraction expansion of the value of $\log_2 3$, assuming there is one.

3.2.3. *Calculating the Continued Fraction of a Logarithm.* We start with a very general case. Let M_0 and M_1 be real numbers, where $1 < M_1 < M_0$. if M_0 and M_1 are rational powers of each other, then $\log_{M_1}(M_0)$ is a rational number which may be represented by a finite continued fraction. If M_0 and M_1 are not rational powers of each other, $\log_{M_1}(M_0)$ is necessarily irrational, so the continued fraction for $\log_{M_1}(M_0)$ (if it exists: on the basis of results proved here we do not know this *a priori*) must be infinite.

Determine a sequence of integers n_i as follows:

1: n_1 such that $(M_1)^{n_1} < M_0 < (M_1)^{n_1+1}$ and let $M_2 = \left(\frac{M_0}{M_1}\right)^{n_1}$.

2: n_2 such that $(M_2)^{n_2} < M_1 < (M_2)^{n_2+1}$ and let $M_3 = \left(\frac{M_1}{M_2}\right)^{n_2}$.

. . . .

k: n_k such that $(M_k)^{n_k} < M_{k-1} < (M_k)^{n_k+1}$ and let $M_{k+1} = \left(\frac{M_{k-1}}{M_k}\right)^{n_k}$.

Line 1 shows that $M_0 = (M_1)^{n_1+(\frac{1}{x_1})}$ for some $x_1 > 1$.

Line 2 shows that $M_1 = (M_2)^{n_2+(\frac{1}{x_2})}$ for some $x_2 > 1$.

. . . .

Line k shows that $M_{k-1} = (M_k)^{n_k+(\frac{1}{x_k})}$ for some $x_k > 1$, etc.

We see that $M_2 = \left(\frac{M_0}{M_1}\right)^{n_1} = \frac{{M_1}^{\frac{n_1+1}{x_1}}}{{M_1}^{n_1}} = {M_1}^{\frac{1}{x_1}}$. That is $M_1 = {M_2}^{x_1}$. But we also have $M_1 = {M_2}^{\frac{n_2+1}{x_2}}$, so $x_1 = n_2 + \frac{1}{x_2}$. Similarly we can show that $x_k = n_{k+1} + \frac{1}{x_{k+1}}$, and we have

$$\begin{aligned} M_0 &= {M_1}^{\frac{n_1+1}{x_1}} \\ &= {M_1}^{(n_1+1)/\frac{n_2+1}{x_2}} \\ &= M_1^{\left(n_1 + \cfrac{1}{n_2 + \cfrac{1}{n_3 + \cfrac{1}{n_4 + \ddots}}} \right)} \end{aligned}$$

By Theorem 6 we know that $[n_1; n_2, n_3 \ldots]$ is a well-defined real number, and so by the definition of a logarithm, $\log_{M_1} M_0 = [n_1; n_2, n_3 \ldots]$

Applying this algorithm to compute the continued fraction of $\log_2 3$ we have:

$$\begin{array}{rll}
2^{n_1} < 3 < 2^{n_1+1} & n_1 = 1. & \text{Then } M_2 = \frac{3}{2}^1 = \frac{3}{2} \quad (= 2^{-1}3^1) \\
\left(\frac{3}{2}\right)^{n_2} < 2 < \left(\frac{3}{2}\right)^{n_2+1} & n_2 = 1. & \text{Then } M_3 = \frac{2}{3/2}^1 = \frac{4}{3} \quad (= 2^2 3^{-1}) \\
\left(\frac{4}{3}\right)^{n_3} < \frac{3}{2} < \left(\frac{4}{3}\right)^{n_3+1} & n_3 = 1. & \text{Then } M_4 = \frac{3/2}{(2^2/3)^1} = \frac{3^2}{2^3} \quad (= 2^{-3}3^2) \\
\left(\frac{9}{8}\right)^{n_4} < \frac{4}{3} < \left(\frac{9}{8}\right)^{n_4+1} & n_4 = 2. & \text{Then } M_5 = \frac{2^2/3^1}{(3^2/2^3)^2} = \frac{2^8}{3^5} \quad (= 2^8 3^{-5}) \\
\left(\frac{256}{243}\right)^{n_5} < \frac{9}{8} < \left(\frac{256}{243}\right)^{n_5+1} & n_5 = 2. & \text{Then } M_6 = \frac{3^2/2^3}{(2^8/3^5)^2} = \frac{3^{12}}{2^{19}} \quad (= 2^{-19}3^{12}) \\
\left(\frac{3^12}{2^19}\right)^{n_6} < \frac{256}{243} < \left(\frac{3^12}{2^19}\right)^{n_6+1} & n_6 = 3. & \text{Then } M_7 = \frac{2^8/3^5}{(3^12/2^19)^3} = \frac{2^{65}}{3^{41}} \quad (= 2^{65}3^{-41})
\end{array}$$

Continuing in this way it is possible to compute the continued fraction expansion of $\log_2 3 = [1; 1, 1, 2, 2, 3, 1, 5, 2, 23, 2, \ldots]$.

The Musician's Algorithm. Each of the M_i is a P number, a ratio between $\hat{P}$ values, and thus M_i represents an interval. It is interesting to note that a musician can rapidly compute the first five partial quotients for $\log_2 3$ by translating the algorithm into musical terms: Instead of computing the largest power of 2 which is less than 3, the musician would consider the number of octaves in a twelfth, which is one ($n_1 = 1$) leaving a remainder of a perfect fifth ($M_2 = \frac{3}{2}$). The fifth goes into an octave once ($n_2 = 1$) leaving a perfect fourth ($M_3 = \frac{4}{3}$), and the fourth goes into the fifth once ($n_3 = 1$) leaving a major second ($M_4 = \frac{9}{8}$). The major second in turn goes into the fourth twice ($n_4 = 2$) leaving a minor second or half-step ($M_5 = \frac{256}{243}$). At this point, if the musician is thinking in terms of equal temperament, he will say that the half-step goes into the major second or whole step twice ($n_5 = 2$) exactly and the process stops. However, even the musician entirely innocent of tuning theory can quickly do these calculations.

3.2.4. *The Importance of Semi-Convergents and a Theorem About Them.* The first five convergents to $\log_2 3$ are $\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}$ and the first five M_i are all prime ratios $\rho(A/B)$. Missing from the list of M_i are $2^5 3^3$ and $2^{-11} 3^7$, though $3^3 = 27$ and $2^5 = 32$ are connected in $\hat{P}$, as are $2^{11} = 2048$ and $3^7 = 2187$.

However, $5/3 = [1; 1, 1, 1]$ and $11/7 = [1; 1, 1, 2, 1]$ *are* included among the semi-convergents to $\log_2 3$. The obvious conjecture is that A and B determine a connected pair of pure powers of 2 and 3 when (and only when) A/B is a convergent or semi-convergent of $\log_2 3$. We will prove that this is true, allowing us to assert, for example, that 2^{50508} and 3^{31867} are connected in $\hat{P}$. The table below lists the first 43 values for A and B when A/B is a convergent or semi-convergent of $\log_2 3$. The sign "+" or "-" refers to the polarity of the prime ratio $\rho(A/B)$, "+" for odd convergents "-" for even. After each full convergent the value of t_k is written.[12]

A	B	+/-	t_k
1	1	-	1
2	1	+	1
3	2	-	1
5	3	+	
8	5	+	2
11	7	-	
19	12	-	2
27	17	+	
46	29	+	
65	41	+	3
84	53	-	1
149	94	+	
233	147	+	
317	200	+	
401	253	+	
485	306	+	5
569	359	-	
1054	665	-	2
1539	971	+	
2593	1636	+	
3647	2301	+	
4701	2966	+	
5755	3631	+	
6809	4296	+	
7863	4961	+	
8917	5626	+	
9971	6291	+	
11025	6956	+	
12079	7621	+	
13133	8286	+	
14187	8951	+	
15241	9616	+	
16295	10281	+	
17349	10946	+	
18403	11611	+	
19457	12276	+	
20511	12941	+	
21565	13606	+	
22619	14271	+	
23673	14936	+	
24727	15601	+	23
25781	16266	-	
50508	31867	-	2

TABLE 3.2.4. PRIME RATIOS

Table 3.2.4 points to the importance of the semi-convergents to $\log_2 3$. We will show that the correspondence conjectured above holds true, but to do this we will need one more result to characterize semi-convergents in general. This motivates a closer study of Continued Fraction Theorem 8, to reinforce this result and to attempt to prove a converse theorem.

It will be sufficient to consider an infinite continued fraction, that is, a continued fraction $[t_0; t_1, \dots]$ which represents an irrational number x. The combinations of Theorems 7 and 8 says that if A/B is convergent or semi-convergent to x and p/q is a rational number between A/B and x, then $q > B$. The converse of this statement is that

Theorem 9. *Continued Fraction Theorem [9] (Converse Theorem). If a positive fraction A/B, $B > 1$, is neither a convergent nor semi-convergent to x, there does exist a fraction p/q between A/B and x, with $q < B$.*

Proof. We consider the case where $A/B < x$, the argument being the same in the other case. Let $c_n = \dfrac{a_n}{b_n} = [t_0; t_1, \dots t_n]$ as before. Fix b_k to be the largest denominator strictly less than B: $b_k < B \leq b_{k+1}$. (For any continued fraction, $b_0 =$

[12] We are reverting to our previous notation t_k, $k = 0, 1, 2, \dots$ in place of n_k, $k = 1, 2,$ etc.

1, so such a b_k exists.) If k is even, it must be $A/B < a_k/b_k < x$, since $A/B > a_k/b_k$ violates the lemma of Theorem 7: $a_k/b_k < A/B < a_{k+1}/b_{k+1}$ implies $B > b_{k+1}$. In that case, [i.e., $A/B < a_k/b_k < x$. - NC] a_k/b_k is the required p/q since $b_k < B$. If on the other hand k is odd, $A/B < x < a_k/b_k$. If $A/B < a_{k-1}/b_{k-1} < x$ there is nothing to prove since a_{k-1}/b_{k-1} is the required p/q, so assume $a_{k-1}/b_{k-1} < A/B < a_{k+1}/b_{k+1} < x$. Now consider the semi-convergents $\dfrac{1\cdot a_k + a_{k-1}}{1\cdot b_k + b_{k-1}}$, $\dfrac{2\cdot a_k + a_{k-1}}{2\cdot b_k + b_{k-1}}$, $\dfrac{(t_{k+1}-1)\cdot a_k + a_{k-1}}{(t_{k+1}-1)\cdot b_k + b_{k-1}}$. By the argument in Theorem 8, each $\dfrac{ta_k + a_{k-1}}{tb_k + b_{k-1}}$ with $0 < t < t_{k+1}$ is between c_{k-1} and c_{k+1}, and as the denominators increase the value of the semi-convergent increases: Since $[t_0; \dots t_k, t+1] = [t_0; \dots t_k, t, 1]$ by the remark on page 19, we have $[t_0; \dots t_k, t]$ an even continued fraction and $[t_0; \dots t_k, t, 1]$ a subsequent odd fraction so we have $[t_0; \dots t_k, t] < [t_0; \dots t_k, t+1]$ for any t from 1 to $t_{k+1} - 1$, thus it follows that

$$\frac{a_{k-1}}{b_{k-1}} < \frac{a_k + a_{k-1}}{b_k + b_{k-1}} < \frac{2a_k + a_{k-1}}{2b_k + b_{k-1}} < \dots < \frac{(t_{k+1}-1)a_k + a_{k-1}}{(t_{k+1}-1)b_k + b_{k-1}}$$

$$< \frac{(t_{k+1})a_k + a_{k-1}}{(t_{k+1})b_k + b_{k-1}} = \frac{a_{k+1}}{b_{k+1}} < x$$

So A/B must lie between

$$[t_0; t_1, \dots t_k, t] = \frac{ta_k + a_{k-1}}{tb_k + b_{k-1}} \text{ and } [t_0; t_1, \dots t+1] = \frac{(t+1)a_k + a_{k-1}}{(t+1)b_k + b_{k-1}}$$

for some t from 0 to $t_{k+1}-1$. Remembering that $[t_0; t_1, \dots t_k, t+1] = [t_0; t_1, \dots t_k, t, 1]$ is a continued fraction itself, whose penultimate convergent is $[t_0; t_1, \dots t_k, t]$, we can see that the two fractions satisfy the condition of Theorem 7 lemma, and B is therefore greater than $(t+1)b_k + b_{k-1}$. Then $\frac{(t+1)a_k + a_{k-1}}{(t+1)b_k + b_{k-1}}$ satisfies the requirements for p/q in the theorem. Therefore, in all cases if A/B is not a semi-convergent or convergent to x, there is a fraction p/q which is a better approximation to x with a smaller denominator. □

3.3. **The Characterization Theorem.** We are now in a position to return to the main line of development and to state and prove the principal theorem characterizing regions. [13]

Theorem 10. *CHARACTERIZATION THEOREM: If, and only if, A/B is a convergent or semi-convergent of $\log_2 3$, the ordered pair (A, B) determines a region $R_{(A,B)}$, and*

$$R_{(A,B)} = \left\{ 2^{(a_k(-1)^k(B-n) \bmod A} \cdot 3^{(b_k(-1)^k(n-A) \bmod B} \;\middle|\; 0 < n < A + B \right\}$$

where a_k/b_k is the full convergent immediately preceding A/B in the continued fraction representation of $\log_2 3$. The mapping

$$n \to R_{(A,B)}(n) : n \to 2^{(a_k(-1)^k(B-n) \bmod A} \cdot 3^{(b_k(-1)^k(n-A) \bmod B}$$

[13] We have defined regions in the context of $\hat{P}$, the positive integers of the form $2^a 3^b$. It will be evident that the entire theory could be generazlized by replacing 2 and 3 by real numbers M_1 and M_0, and later we will extend the theory to cover other tunings by maintaining $M_1 = 2$ but replacing 3 with a value M_0 which is an approximation to 3.

is strictly increasing with n *for* $n = 1, 2, \ldots A+B-1$*, and the isomorphism* ν *maps the last* $A+B-1$ *elements of* $Z_{(AB)}$ *from* $AB-(A+B)+1$ *to* $AB-1$ *in order onto the elements of* $R_{(A,B)}$.

This main result will be the goal of the development which follows, though the theorems and corollaries along the way have their own interest.

Proof.

3.3.1. *Proof I: Connectedness and Approximations.* We prove the first part of the principal theorem by proving the equivalence between connectedness and convergence to $\log_2 3$ conjectured above. Then a necessary and sufficient condition for the existence of the region $R_{(A,B)}$ will be that A/B is a convergent or semi-convergent to $\log_2 3$.

Theorem 11. 2^A *and* 3^B *are connected in* $\hat{P}$ *when* $B \geq 1$ *if and only if* A/B *is a convergent or semi-convergent in the continued fraction expansion of* $\log_2 3$*.*

Proof. In the case $B = 0$, 2^1 and 3^0 are connected, but A/B is not defined. For the case $B = 1$ we can determine by inspection that the theorem holds, [Not true if $\theta > 2$ replaces $\log_2 3$ and $B = 1$. Also, of $1/0$ and $0/1$, one is always connected the other not (and note that with $0/1$, $B = 1$ and A/B *is* defined). - NC] so we may assume $B > 1$. We will also assume throughout that $2^A > 3^B$, that is, $A/B > \log_2 3$, since the same arguments may be made in the case $3^B > 2^A$. We will make use of Continued Fraction Theorems 7 and 8 and the Converse Theorem and also the following lemma:

Lemma 2. *If* a, b, c, d *are positive real numbers where* $\dfrac{a}{b} < \dfrac{c}{d}$*, then* $\dfrac{a}{b} < \dfrac{a+c}{b+d} < \dfrac{c}{d}$

Proof of Lemma:

$\frac{a}{b} < \frac{c}{d} \to ad < bc \to (ad+ab) < (bc+ab)$, so $a(d+b) < b(c+a)$, so $\frac{a}{b} < \frac{a+c}{b+d}$, and likewise, $ad < bc \to (ad+cd) < (bc+cd)$ so $d(a+c) < c(b+d)$, $\frac{c}{d} < \frac{a+c}{b+d}$.

Part 1 of the proof shows that if A/B is a semi-convergent or convergent of $\log_2 3$, then 2^A and 3^B are connected in $\hat{P}$:

Assume 2^A and 3^B are not connected, and it will follow that A/B cannot be a semi-convergent or convergent. That is, we assume there exist integers $x \geq 0$ and $y \geq 0$ such that $2^A > 2^x 3^y > 3^B$.

Clearly $x < A$. If $y \geq B$, we have $2^A > 2^x 3^y \geq 2^x 3^B > 3^B$ and $x \neq 0$ in the case $y = B$. Then [if $y = B$ –NC] $\frac{2^A}{3^B} > 2^x \geq 2$. Then $A - B\log_2 3 > 1$, that is, $\frac{A-1}{B} > \log_2 3$. Since $A > A-1$, $\frac{A}{B} > \frac{A-1}{B} > \log_2 3$, so by theorems 7 and 8 either A/B is not a semi-convergent or convergent of $\log_2 3$ or $y < B$.

Therefore we take $y < B$. Then $\frac{2^A}{3^B} > \frac{2^x}{3^{B-y}}$ and $\frac{2^A}{3^B} > \frac{2^{A-x}}{3^y}$ and $\frac{x}{B-y} > \log_2 3$ and $\frac{A-x}{y} > \log_2 3$ all follow from $2^A > 2^x 3^y > 3^B$. If $\frac{x}{B-y} = \frac{A-x}{y}$, $xy = AB-(Ay+Bx)+xy$ or $0 = AB-(Ay+Bx)$ in which case A and B are not relatively prime, hence A/B is not a convergent or semi-convergent and the theorem is proved. Next we consider the cases $\frac{x}{B-y} > \frac{A-x}{y}$ or $\frac{x}{B-y} < \frac{A-x}{y}$. Since $x < A$ and $y < B$, the numerators and denominators of both are positive integers and we may apply the lemma. Thus, either $\frac{A-x}{y} > \frac{A}{B} > \frac{x}{B-y}$ or $\frac{x}{B-y} > \frac{A}{B} > \frac{A-x}{y}$. But $\frac{x}{B-y} > \log_2 3$ and $\frac{A-x}{y} > \log_2 3$, so either we have $A/B > \frac{x}{B-y} > \log_2 3$ or else $A/B > \frac{A-x}{y} > \log_2 3$, and both y and $B - y$ are less than B, so by Continued Fraction Theorems 7 and 8, A/B cannot be a convergent or semi-convergent of $\log_2 3$.

We have proved that if A/B is a convergent or semi-convergent of $\log_2 3$, then A and B are connected in $\hat{P}$.

Part 2 of the proof shows that if 2^A and 3^B are connected, then A/B is a convergent or semi-convergent of $\log_2 3$.

We will prove again by contradiction, assuming A/B is not a convergent [or semi-convergent] and showing that 2^A and 3^B are not connected. We continue to assume the case where $2^A > 3^B$, so $\frac{A}{B} > \log_2 3$. If A/B is not a semi-convergent or convergent to $\log_2 3$, by the converse theorem there exists p/q, $q < B$, such that $\frac{A}{B} > \frac{p}{q} > \log_2 3$. Then $p < A$, and $2^{A-p}3^q < 2^A$: $2^p > 3^q$, so $\frac{3^q}{2^p} < 1$, thus $2^A > 2^A \frac{3^q}{2^p}$. Suppose $2^{A-p}3^q < 3^B$. Then $\frac{2^p}{3^q} > \frac{2^A}{3^B}$, which implies that $p - q\log_2 3 > A - B\log_2 3$. Then $\left(\frac{p}{q} - \log_2 3\right) > \frac{(A - B\log_2 3)}{q} > \frac{(A - B\log_2 3)}{B}$ since $q < B$, so we have $\left(\frac{p}{q} - \log_2 3\right) > \left(\frac{A}{B} - \log_2 3\right)$ or $p/q > A/B$, contrary to assumption, so we must have $2^A > 2^{A-p}3^q > 3^B$, that is, 2^A and 3^B are not connected. In the argument above, we could just as well have taken the element $2^p 3^{B-q}$, shown that it must be greater than 3^B, and derived a contradiction by assuming it to be greater than 2^A. Therefore, if 2^A and 3^B are connected, A/B must be a convergent or semi-convergent of $\log_2 3$.

Finally, putting the two parts of the proof together, we have 2^A and 3^B with $B \geq 1$ are connected in $\hat{P}$ if and only if $\frac{A}{B}$ is at least a semi-convergent of $\log_2 3$. □

An immediate corollary of this theorem is that the ratio between any two connected $\hat{P}$ elements, reduced to lowest terms, is a prime ratio:

Corollary 12. *If $\hat{p}_1$ and $\hat{p}_2$ are connected in $\hat{P}$, $\hat{p}_1 > \hat{p}_2$, then $\frac{\hat{p}_1}{\hat{p}_2} = \frac{2^A}{3^B}$ or $\frac{3^B}{2^A}$ where $\frac{A}{B}$ is a convergent or semi-convergent of $\log_2 3$ except in the one case $\hat{p}_1 = 2$, $\hat{p}_2 = 1$.*

Proof. Suppose $n = 2^a 3^b$ is the highest common factor of $\hat{p}_1$ and $\hat{p}_2$. Then $\dfrac{\hat{p}_1}{\hat{p}_2} = \dfrac{2^A n}{3^B n}$ or $\dfrac{3^B n}{2^A n}$. (One of the exponents must be negative, since if $\frac{\hat{p}_1}{\hat{p}_2} > 2$, $\hat{p}_1 > 2\hat{p}_2 > \hat{p}_2$, contradicting the assumption that $\hat{p}_1$ and $\hat{p}_2$ are connected. If $\frac{\hat{p}_1}{\hat{p}_2} = 2$, then either $\hat{p}_1 = 2$ and $\hat{p}_2 = 1$, or they are not connected.)

Then suppose $\dfrac{A}{B}$ is not a convergent or semi-convergent, that is, 2^A and 3^B are not connected. $2^A > 2^x 3^y > 3^B$ or $3^B > 2^x 3^y > 2^A$. Then it is also true that $2^A \cdot n > (2^x 3^y) \cdot n > 3^B \cdot n$ or $3^B \cdot n > (2^x 3^y) \cdot n > 2^A \cdot n$, and $\hat{p}_1$ and $\hat{p}_2$ are not connected. □

3.3.2. *Proof II: Strictly increasing functions: The isomorphism $\boldsymbol{\nu}$.* At this point we turn from $\hat{P}$ to the isomorphic groups Z_{AB} and $Z_A \times Z_B$, where A and B are relatively prime. Of course, A and B must be relatively prime if A/B is a convergent or semi-convergent.

Recall that we exhibited an isomorphism

$$\mu : Z_A \times Z_B \to Z_{AB} : (x, y) \to (Ay + Bx)_{\text{ mod } AB}.$$

This guarantees the existence of an inverse mapping which is also an isomorphism

$$\mu^{-1} : Z_{AB} \to Z_A \times Z_B : (Ay + Bx)_{\text{ mod } AB} \to (x, y).$$

This inverse isomorphism composed with the isomorphism λ yields a mapping ν from Z_{AB} onto $G_{(A,B)}$: μ^{-1} maps Z_{AB} onto $Z_A \times Z_B$ and λ maps $Z_A \times Z_B$ onto $G_{(A,B)}$. Since μ^{-1} and λ are isomorphisms, ν is an isomorphism.

We will be interested in the situation where $R_{(A,B)}$ is defined, and we would like to determine the inverse image under ν of $R_{(A,B)}$, that is, the subset of Z_{AB} which ν maps onto the elements of $R_{(A,B)}$. This will enable us to give explicitly the elements of $R_{(A,B)}$ in the general case.

We start with the relationship between Z_{AB} and $Z_A \times Z_B$. Consider the equation $Ay+Bx = z$, where the operation is ordinary addition. If the equation can be solved in non-negative integers x and y we will say that z is *solvable*, and x and y *solve* for z or are *solutions*. It is not hard to show that any $z \geq AB$ is solvable in this sense, as long as A and B are relatively prime. For positive integers less than AB, i.e., for elements of Z_{AB}, there are in general elements which are not solvable. We give a somewhat lower bound for solvability by proving a rather obscure theorem of number theory:

Theorem 13. *Assume A and B are relatively prime with $A > B > 1$. Then the $A + B - 1$ integers from $AB - (A + B) + 1$ to $AB - 1$ are all solvable in positive integers x, y where $0 \leq x < A$ and $0 \leq y < B$. Of the remaining elements, half of them, numbering $\frac{(A-1)(B-1)}{2}$, are solvable and the other $\frac{(A-1)(B-1)}{2}$ are not solvable. The element $AB-(A+B)$ is not solvable, so $AB-(A+B)$ is the smallest integer above which every integer is solvable.*

Proof. If $Ay + Bx = z$ with strict equality for $z \in Z_{AB}$ then z is solvable, by definition. If $z \in Z_{AB}$ is not solvable, then it must be that $AB + z = Ay + Bx$ for some pair $(x, y) \in Z_A \times Z_B$, and conversely, if $Ay + Bx = AB + z$, then z is not solvable, since μ is one-to-one and onto Z_{AB}. Then consider the ordered pair $(A-1, B-1)$. $A(B-1)+B(A-1) = 2AB-(A+B) = AB+(AB-(A+B))$ and since $A > B > 1$, $AB-(A+B)$ is not solvable. But this is the highest value the form $Ay+Bx$ can take on when $0 \leq x < A$, $0 \leq y < B$, i.e. when $(x, y) \in Z_A \times Z_B$, and since μ is onto Z_{AB}, the elements $AB-(A+B)+1$ to $AB-1$ are solvable for $(x, y) \in Z_A \times Z_B$. That leaves the elements from 0 to $AB - (A + B)$. There are $(A - 1)(B - 1)$ of them since $AB - (A + B) + 1 = (A - 1)(B - 1)$. Suppose an element $(x\prime, y\prime) \in Z_A \times Z_B$ is a solution for $n\prime$, where $0 \leq n\prime < (A-1)(B-1)$. Then the element $((A-1)-x\prime, (B-1)-y\prime)$ is not a solution for any $z \in Z_{AB}$:

$$\begin{aligned} &A((B-1) - y\prime) + B(A-1) - x\prime) \\ &= 2AB - (A+B) - (Ay\prime + Bx\prime) \\ &= 2AB - (A+B) - n\prime > AB \end{aligned}$$

Conversely, suppose $(x\prime\prime, y\prime\prime)$ is not a solution for any element of Z_{AB}. Then $A(x\prime\prime+By\prime\prime) > AB$. Then $((A-1)-x\prime\prime, (B-1)-y\prime\prime)$ is a solution for some element $n\prime\prime$ such that $0 \leq n\prime\prime < (A-1)(B-1) = AB - (A+B) + 1$:

$$\begin{aligned} A((B-1)-y'') + B(A-1)-x'') \\ = 2AB-(A+B)-(Ay''+Bx'') \\ < AB+(A+B) \end{aligned}$$

Therefore $\frac{(A-1)(B-1)}{2}$ of the elements of Z_{AB} from 0 to $AB-(A+B)$ are solvable, and $\frac{(A-1)(B-1)}{2}$ not.

The element $((A-1)-x,(B-1)-y)$ used in the latter part of the above proof is reminiscent of the notion of complementarity which was defined in $G_{(A,B)}$. The property of complementarity is a group property, that is, it is defined in terms of the group operation. In $G_{(A,B)}$, g and $g\prime$ were defined to be complements of each other if $gg\prime = 2^{A-1}3^{B-1}$. Although the group operation in $G_{(A,B)}$ is slightly different from ordinary multiplication, clearly it remains true that if $g = 2^a3^b$, $g\prime = 2^{(A-1)-a}3^{(B-1)-b}$ is the complement of g.

Then we can define complements in $Z_A \times Z_B$ and Z_{AB} so that complementarity is preserved by the various isomorphisms: In $Z_A \times Z_B$, given $x = (a,b)$, $x\prime = ((A-1)-a,(B-1)-b)$. Then it is true that $\lambda(x\prime) = (\lambda(x))\prime$. Since $\mu(A-1,B-1) = AB-(A+B)$ as we saw above, if $z \in Z_{AB}$, $z\prime \equiv (AB-(A+B)-z)_{\text{ mod } AB}$. Thus if $z \leq AB-(A+B)$, $z\prime \leq AB-(A+B)$ and if $z > AB-(A+B)$, $z\prime > AB-(A+B)$, ordering the elements of Z_{AB} as ordinary integers.

□

Looking back at the proof of the theorem, [Theorem 13] evidently half of the elements of Z_{AB} from 0 to $AB-(A+B)$ are solvable, and their complements are not. Furthermore, if, and only if, an element and its complement are both solvable, they both belong to what we may call the *abstract region* from $AB-(A+B)+1$ to $AB-1$.

Recalling our definition of a set of complements, we can see that both of the subsets 0 to $AB-(A+B)$ and $AB-(A+B)+1$ to $AB-1$ are sets of complements. Also, both subsets are connected, if we carry over the definition of connectedness to Z_{AB}. However, what we have called the abstract region has the additional property that all of its elements are solvable, so this is our candidate for the inverse image under ν of the region $R_{(A,B)}$.

First we determine the solutions $(x,y) \in Z_A \times Z_B$ for the elements of the abstract region.

Theorem 14. *The values* $(a_k(B-n)(-1)^k_{\text{mod } A}$ *and* $(b_k(n-A)(-1)^k)_{\text{ mod } B}$*, where* $\frac{a_k}{b_k}$ *is the penultimate convergent in the continued fraction representing* A/B[14] *are solutions* x *and* y *respectively in positive integers to the equation* $Ay + Bx = AB-(A+B)+n$ *for* n *from* 1 *to* $A+B-1$.

Proof. First of all, the given values may be simplified by the use of the identity $Ab_k - Ba_k = (-1)^k$ from Continued Fraction Theorem 2 to $(-a_kn(-1)^k-1)_{\text{ mod } A}$ and $(b_kn(-1)^k-1)_{\text{ mod } B}$:

[14]We are assuming the result that any rational number has a continued fraction representation which is unique and finite. However, the specific case we are interested in is when A/B is a convergent or semi-convergent of $\log_2 3$, in which case the penultimate convergent to A/B and the full convergent of $\log_2 3$ immediately preceding A/B are the same thing.

$a_kB = Ab_k - (-1)^k$, so

$$\left(a_k(B-n)(-1)^k\right)_{\bmod A} \equiv \left(Ab_k - (-1)^k - a_kn\right)(-1)^k{}_{\bmod A}$$
$$\equiv \left(-a_kn(-1)k - ((-1)^k)^2\right)_{\bmod A} \equiv \left(-a_kn(-1)^k - 1\right)_{\bmod A}$$

and $b_kA = (-1)^k + Ba_k$, so

$$\left(b_k(n-A)(-1)^k\right)_{\bmod B} \equiv \left(b_kn - ((-1)^k + Ba_k)\right)(-1)^k{}_{\bmod B}$$
$$\equiv \left(b_kn(-1)k - ((-1)^k)^2\right)_{\bmod B} \equiv \left(b_kn(-1)^k - 1\right)_{\bmod B}$$

Note that $A\left(b_kn(-1)^k\right) + B\left(-a_kn(-1)^k\right) = n$ for any value of n whatsoever, since $A\left(b_kn(-1)^k\right) + B\left(-a_kn(-1)^k\right) = (Ab_k - Ba_k)(-1)^kn = \left((-1)^k\right)^2 n = n$. Then $A\left(b_kn(-1)^k - 1\right) + B\left(-a_kn(-1)^k - 1\right) = n - (A+B)$ for any value of n.

Next recall Lemma 1 which states that $\left(A(y)_{\bmod A} + B(x)_{\bmod A}\right)_{\bmod AB} \equiv (Ay+Bx)_{\bmod AB}$

Applying this lemma to the results derived above we have the following:

$$\left(A\left(\left(b_k(n-A)(-1)^k\right)\right)_{\bmod B} + B\left(\left(a_k(B-n)(-1)^k\right)\right)_{\bmod A}\right)_{\bmod AB} \equiv$$
$$\left(A\left(\left(b_kn(-1)^k\right) - 1\right)_{\bmod B} + B\left(\left(-a_kn(-1)^k - 1\right)\right)_{\bmod A}\right)_{\bmod AB} \equiv$$
$$\left(A\left(\left(b_kn(-1)^k\right) - 1\right) + B\left(\left(-a_kn(-1)^k - 1\right)\right)\right)_{\bmod AB} \equiv$$
$$(n - (A+B))_{\bmod AB} \equiv (AB - (A+B) + n)_{\bmod AB}$$

We know that Z_{AB} may be partitioned into three blocks: Out of the first $(A-1)(B-1)$ elements, $\frac{(A-1,B-1)}{2}$ elements z such that $z = Ay + Bx$ has a non-negative solution; their $\frac{(A-1,B-1)}{2}$ complements $z\prime$ which have no such solution; and the remaining elements of the abstract region

$$(A-1)(B-1) = AB - (A+B) + 1, \ldots AB - (A+B) + n, \ldots AB - 1$$

for n from 1 to $A+B-1$ which all have non-negative solutions, which are precisely $x = \left(a_k(B-n)(-1)^k\right)_{\bmod A}$, and $y = \left(b_k(n-A)(-1)^k\right)_{\bmod B}$.

We wish to show that when the elements $AB - (A+B) + n$ are mapped into $G_{(A,B)}$ via the mapping ν, their order is preserved. To do this we will show that if the mapping $(x,y) \to Ay + Bx$ is increasing, so is $\lambda : (x,y) \to 2^x3^y$. It will be useful to show this for $(x,y) \in Z_A \times Z_B$ plus the two elements $(A,0)$ and $(0,B)$. Then we will be able to say that as the solvable elements of Z_{AB} get larger, their images under ν in $G_{(A,B)}$ also increase. From now on we assume A/B is a full or semi-convergent of $\log_2 3$.

Theorem 15. *If $z_1, z_2 \in Z_{AB} \cup \{AB\}$, where $z_2 > z_1$ and $z_2 = Ay_2 + Bx_2$ and $z_1 = Ay_1 + Bx_1$ for $(x_1,y_1),(x_2,y_2) \in Z_A \times Z_B \cup \{(A,0),(0,B)\}$, then $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$.*

Proof. $Ay_2 + Bx_2 > Ay_1 + Bx_1$ implies $\frac{A}{B} > \frac{(x_1-x_2)}{(y_2-y_1)}$ as long as $(y_2 - y_1) > 0$. If $(y_2 - y_1) < 0$, $\frac{A}{B} < \frac{(x_1-x_2)}{(y_2-y_1)}$ so in this case $(x_1 - x_2) < 0$, and $\frac{A}{B} < \frac{(x_2-x_1)}{(y_1-y_2)}$ in positive integers. If $y_1 = y_2$, $x_2 > x_1$, so $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$ and order is preserved. So we have two cases, $(y_2 - y_1) > 0$ and $(y_2 - y_1) < 0$.

Suppose $\frac{A}{B} > \log_2 3$. If $(y_2 - y_1) > 0$, $\frac{A}{B} > \frac{(x_1-x_2)}{(y_2-y_1)}$, but $(y_2 - y_1) < B$, and $\frac{A}{B}$ is a convergent or semi-convergent of $\log_2 3$, so it must be that $\frac{A}{B} > \log_2 3 > \frac{(x_1-x_2)}{(y_2-y_1)}$. Therefore $3^{y_2-y_1} > 2^{x_1-x_2}$ and $\frac{3^{(y_2-y_1)}}{2^{(x_1-x_2)}} > 1$, $3^{(y_2-y_1)}2^{(x_2-x_1)} > 1$, so $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$. Therefore, in this case also, λ is increasing if $Ax + By$ is increasing.

Now suppose $\frac{A}{B} > \log_2 3$ but $(y_2 - y_1) < 0$. Then $\frac{A}{B} < \frac{(x_1-x_2)}{(y_2-y_1)}$, therefore $\frac{(x_2-x_1)}{(y_1-y_2)} > \log_2 3$, that is $2^{(x_2-x_1)} > 3^{(y_1-y_2)}$, so $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$.

If $\frac{A}{B} < \log_2 3$ and $(y_2 - y_1) > 0$, then $\frac{A}{B} > \frac{(x_1-x_2)}{(y_2-y_1)}$, therefore $\frac{(x_1-x_2)}{(y_2-y_1)} < \log_2 3$ and again $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$. Finally, if $\frac{A}{B} < \log_2 3$ and $(y_2 - y_1) < 0$, $\frac{A}{B} < \frac{(x_2-x_1)}{(y_1-y_2)}$, and since $(y_1 - y_2) < B$, it must be the case that $\frac{A}{B} < \log_2 3 < \frac{(x_2-x_1)}{(y_1-y_2)}$. Therefore as above $2^{(x_2-x_1)} > 3^{(y_1-y_2)}$, so $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$, and the theorem holds in every case. □

We have just seen that if $\frac{A}{B}$ is a convergent or semi-convergent of $\log_2 3$, for the elements of $Z_A \times Z_B$ and the ordered pair $(A, 0), (0, B)$, if $Ay_2 + Bx_2 > Ay_1 + Bx_1$, then $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1}$. This information, together with the partitioning of Z_{AB} according to solvability gives the order of the corresponding $G_{(A,B)}$ elements in $\hat{P}$:

Theorem 16. *The mapping* $\nu = \lambda\mu^{-1} \to G_{(A,B)} : z \xrightarrow{\mu^{-1}} (x, y) \xrightarrow{\lambda} 2^x 3^y$ *sends the solvable elements of* Z_{AB} *onto* $S_{(A,B)}$*, the elements which are not solvable onto the rest of* $G_{(A,B)}$*. In particular,* ν *maps the elements* $AB - (A + B) + n$ *for* $n = 1, \ldots A + B - 1$ *onto the region* $R_{(A,B)}$ *with their order preserved.*

Proof. By the previous theorem [Theorem 15], the $\frac{(A-1)(B-1)}{2}$ elements of Z_{AB} which are less than $(A-1)(B-1)$ and are solvable, ordered as ordinary integers, are mapped into $G_{(A,B)}$ and if $z_2 > z_1$, $\lambda(z_2) > \lambda(z_1)$, i.e., with their order preserved.

The next block of solvable integers $AB - (A+B) + n$ for $n = 1, \ldots A+B-1$ are likewise mapped into $G_{(A,B)}$ in order, all of them greater than the elements from the previous block of solvable elements.

Since $AB = A \cdot B + B \cdot 0 = A \cdot 0 + B \cdot A$, we don't have enough information to know whether $2^A > 3^B$ or vice versa. However, it is clear from the theorem that both 2^A and 3^B are greater than the previous $\frac{(A-1)(B-1)}{2} + (A + B - 1) = \frac{AB+A+B-1}{2}$ elements.

The remaining $\frac{(A-1)(B-1)}{2}$ elements of Z_{AB} are not solvable. If $z_1 < z_2$ are two such elements we have $z_1 + AB = Ay_1 + Bx_1$ and $z_2 + AB = Ay_2 + Bx_2$ for some $(x_1, y_1), (x_2, y_2) \in Z_A \times Z_B$, and therefore $2^{x_2}3^{y_2} > 2^{x_1}3^{y_1} > 2^A$ and 3^B. The element $AB - (A + B)$ corresponds to $(A - 1, B - 1)$, the maximum values for x and y in $Z_A \times Z_B$, and it is mapped by ν onto the greatest element of $G_{(A,B)}$, $2^{A-1}3^{B-1}$.

From this information we know that an element of $G_{(A,B)}$, s, is less than 2^A and 3^B, i.e., is a member of $S_{(A,B)}$ if, and only if $\nu^{-1}(s)$ is solvable. Therefore, the solvable elements in Z_{AB}, mapped into $G_{(A,B)}$, constitute the subset $S_{(A,B)}$.

Now recalling the result from page 30 that if and ony if an element of Z_{AB} and its complement are both solvable, they both belong to the subset from $AB-(A+B)+n$ to $AB - 1$, we can see that $R_{(A,B)}$ is the image of this subset, the abstract region, under ν: $R_{(A,B)}$ is connected, so $R_{(A,B)} \subset S(A, B)$, so only solvable elements may be mapped into $R_{(A,B)}$. Furthermore, $R_{(A,B)}$ was defined to be the largest non-empty set of [connected] complements, and the largest subset of Z_{AB} which is a set of complements and in which all elements are solvable is the abstract region. Finally, complementarity is preserved by ν, that is, if $z \in Z_{AB}$, $\nu(z\prime) = (\nu(z))\prime$, so ν maps the elements $AB - (A + B) + n$ for $n = 1, \dots A + B - 1$ onto the region $R_{(A,B)}$ with their order preserved.

□

Theorem 10 (Characterization Theorem) restated:

$$R_{(A,B)} = \left\{ 2^{(a_k(-1)^k(B-n)_{\text{ mod } A}} \cdot 3^{(b_k(-1)^k(n-A)_{\text{ mod } B}} \;\middle|\; 0 < n < A + B \right\}$$

and as n increases, $R_{(A,B)}(n)$ increases.

This follows immediately from the theorem just proved [Theorem 16] and from Theorem 14 that $(a_k(B-n)(-1)^k)_{\text{ mod } A}$ and $(b_k(n-A)(-1)^k)_{\text{ mod } B}$ are solutions x and y respectively in positive integers to $Ay + Bx = AB - (A + B) + n$ for $1 \leq n \leq A + B - 1$.

□

---

4. Regions and Well-Formed Scales

With this result the proof of the Characterization Theorem [Theorem 10] is completed, providing a great deal of information. Now the necessary and sufficient conditions for a region to exist are available, beyond those provided in the original definition, and a means of computing the ordered pairs which give rise to regions is at hand. The theorem provides the formula for computing, in order, the elements of any given region, incidentally revealing that every region contains $A + B - 1$ elements. However, the central result is the way the cyclic ordering of the group coincides with the natural ordering of arithmetic in the region. The mapping between Z_{AB} and $G_{(A,B)}$ serves as a kind of x-ray to lay bare the cyclic ordering of the region. We will return to this topic after reconsidering well-formed scales and the intervals of which they are composed.

The following is a corollary which isolates the octave frame in a region:

Corollary 17. *If $A > n \geq 1$, $R_{(A,B)}(n + B) = 2 \cdot R_{(A,B)}(n)$.*

That is, the last element in any connected sequence of $B + 1$ region elements is twice the value of the first.

Proof. $R_{(A,B)}(n) = 2^{a_k(B-n)\left((-1)^k\right)_{\text{ mod } A}} \cdot 3^{b_k(n-A)\left((-1)^k\right)_{\text{ mod } B}}$ and

$$\begin{aligned} R_{(A,B)}(n + B) &= 2^{a_k(B-(B+n))\left((-1)^k\right)_{\text{ mod } A}} \cdot 3^{b_k((B+n)-A)\left((-1)^k\right)_{\text{ mod } B}} \\ &= 2^{a_k(-n)\left((-1)^k\right)_{\text{ mod } A}} \cdot 3^{b_k(n-A)\left((-1)^k\right)_{\text{ mod } B}} \end{aligned}$$

The exponents on 3 are the same for $R_{(A,B)}(n)$ and $R_{(A,B)}(n + B)$, so compare the exponents on 2. On page 30 [in Theorem 14] we showed, using the identity $Ab_k - Ba_k = (-1)^k$ that $a_k(B-n)\left((-1)^k\right)_{\text{mod } A} \equiv -a_k n\left((-1)^k\right) - 1_{\text{ mod } A}$, so if $A > 1$, and $0 < n < A$, the exponent on 2 is increased by 1.

Then $R_{(A,B)}(n + B) = 2 \cdot R_{(A,B)}(n)$.

□

In the light of the above and information from the Characterization Theorem, every sequence of elements $R_{(A,B)}(n)$ to $R_{(A,B)}(n+B)$ forms a well-formed scale spanning an octave. Henceforth, we can think of a well-formed scale as any sequence of $B+1$ consecutive region elements, or any Pythagorean scale that can be reduced to such a sequence. The modal system of well-formed Pythagorean scales associated with$R_{(A,B)}$ consists of the different scales, each of length $B+1$ and numbering B in all, whose initial note corresponds to $R_{(A,B)}(1)$, $R_{(A,B)}(2)$, $\ldots R_{(A,B)}(B)$.

Since $R_{(A,B)}$ is a connected set, the ratios between successive elements of the region must correspond to (semi-)convergents[15] of $\log_2 3$, that is, the intervals between adjacent tones of any well-formed scale correspond to the prime ratios $\rho(A.B) = 2^A 3^{-B}$ or $2^{-A}3^B$. Suppose r_1 and r_2 are successive region elements, $r_2 > r_1$. We wish to compute the possible ratios r_2/r_1, where the division is the division of ordinary arithmetic. In ordinary arithmetic, r_2/r_1 yields a P number $2^x 3^y$ where $1 < 2^x 3^y \leq 2$ (see page 28). So x and y are of opposite signs. On the other hand, division in the group $G_{(A,B)}$ yields an element of $G_{(A,B)}$, i.e., an element of $\hat{P}$, $2^a 3^b$, where a and b are non-negative integers. However, the results of these two different operations must be consistent, that is, they must satisfy the relations $x \equiv a_{\text{ mod } A}$ and $y \equiv b_{\text{ mod } B}$. Let

$$r_1 = 2^{(a_k(-1)^k(B-n) \text{ mod } A} \cdot 3^{(b_k(-1)^k(n-A) \text{ mod } B}$$
$$r_2 = 2^{(a_k(-1)^k(B-(n+1)) \text{ mod } A} \cdot 3^{(b_k(-1)^k((n+1)-A) \text{ mod } B}$$

Then in the group $r_2/r_1 = 2^{(-a_k((-1)^k) \text{ mod } A} \cdot 3^{(b_k((-1)^k) \text{ mod } B}$.

If k is even, $(-a_k((-1)^k)_{\text{ mod } A} = A - a_k$, $(b_k((-1)^k)_{\text{ mod } B} = b_k$.
If k is odd, $(-a_k((-1)^k)_{\text{ mod } A} = a_k$, $(b_k((-1)^k)_{\text{ mod } B} = B - b_k$

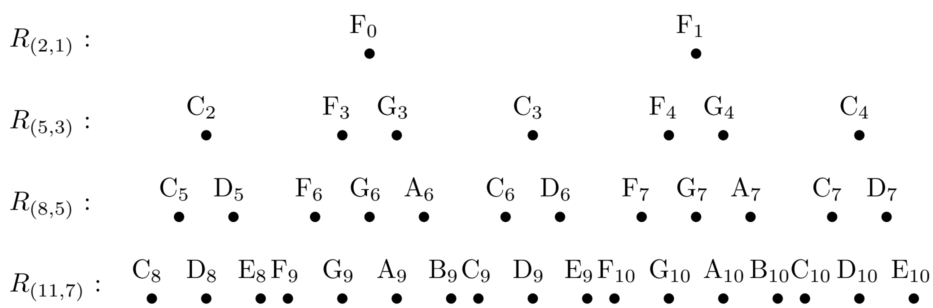
Then if k is even, $3^{b_k} > 2^{a_k}$ and $2^A > 3^B$, so the ordinary ratio r_2/r_1 must be $3^{b_k}/2^{a_k}$ or $2^{(A-a_k)}/3^{(B-b_k)}$. If k is odd, $2^{a_k} > 3^{b_k}$ and $3^B > 2^A$, so r_2/r_1 must be either $2^{a_k}/3^{b_k}$ or $3^{(B-b_k)}/2^{(A-a_k)}$. For example, for k even, $3^{b_k}/2^{a_k} = 3^{b_k}2^{-a_k}$, and $-a_k \equiv (A-a_k)_{\text{ mod } A}$ and $b_k \equiv (b_k)_{\text{ mod } B}$,
so $-a_k \equiv -a_k((-1)^k)_{\text{ mod } A}$, $b_k \equiv b_k((-1)^k)_{\text{ mod } A}$ and the necessary relations are satisfied.

At most two intervals, then, exist between adjacent elements of a given region. If $B > 1$, at *least* two step intervals must appear in the region: Since $B > 1$, it is true that $A - 1 > 0$, so the region has more than B elements, i.e., the region spans at least an octave, and the ratio 2/1 is the ratio of the $(B+1)^{st}$ element to the first element of the region. If there were only one ratio between adjacent elements in the region, we would have $\left(\frac{2^x}{3^y}\right)^B = 2$ for some integers x and y, which cannot be since $\sqrt[B]{2}$ is irrational. Thus if $B > 1$, the intervals associated with $\frac{a_k}{b_k}$ and $\frac{A-a_k}{B-b_k}$ must be distinct and must both appear in any well-formed scale. Of course, in the case $B = 1$ only one interval appears: The two cases are $R_{(1,1)} = \{1\}$, unison, and the Octave Region, $R_{(2,1)} = \{1, 2\}$, spanning the open octave.

[15] From this point on, we use the term *(semi-)convergent* to stand for *either* a convergent or a semi-convergent. -NC

4.1. **Families of Regions.** Since a_k/b_k is a full convergent, at least one of the intervals in any given region is associated with a convergent. We say that the regions which have this interval in common are *related*. If $A_1/B_1, \ldots A_t/B_t$ are semi-convergents with the preceding convergent being a_k/b_k, then the regions associated with them have the interval determined by a_k/b_k in common. An extension of this collection is the *family* of regions associated with a_k/b_k: $\left\{R_{(na_k+a_{k-1}, nb_k+b_{k-1})} \mid 0 \le n \le t_{k+1}+1\right\}$ All of the regions except the initial $R_{(a_{k-1}b_{k-1})}$ have an interval in common, $2^{a_k}/3^{b_k}$ if k is odd, $3^{b_k}/2^{a_k}$ if k is even. The initial region of a family does not contain the relating interval, and the final region of the family is the last region which does contain it. The size of the various families is determined by the partial quotients of the continued fraction of $\log_2 3$, the terms t_i. with the number of regions in a family being $t_i + 2$.

From a musical point of view the most interesting family is the family whose final region is the Diatonic Region $R_{(11,7)}$. The interval which relates these regions is the whole step, $3^2/2^3$. The Structural, Pentatonic, and Diatonic Regions all contain whole steps, which fill in the initial octave of the Octave Region. This family is depicted below:



That the diatonic region $R_{(11,7)}$, is the final member of this family is related to the coincidence up until 11/7 of the convergents and semi-convergents of $\log_2 3$ with the convergents to the golden number, ϕ. Since the continued fraction for $\log_2 3$ begins $[1; 1, 1, 2, \ldots]$, the first five elements in the sequence of (semi-)convergents are $[1] = \frac{1}{1}$, $[1; 1] = \frac{2}{1}$, $[1; 1, 1] = \frac{3}{2}$,$[1; 1, 1, 1] = \frac{5}{3}$, and $[1; 1, 1, 2] = [1; 1, 1, 1, 1] = \frac{8}{5}$, coinciding with the first five convergents to $\phi = [1; 1, 1, 1, 1, \ldots]$. The Diatonic Region is then the last region where both of the intervals which make up the region and the associated well-formed Pythagorean scale are determined by convergents to the golden number.

The relationship of the golden section to musical scales and intervals has been invoked by many writers, usually on the basis of isolated, perhaps accidental correspondences. For example, the pleasing quality of the major sixth and the minor sixth in just intonation is sometimes attributed to the fact that the frequency-ratios for these intervals, 5/3 and 8/5, are the ratios of successive Fibonacci numbers. The relationship we have exhibited is based on coincidence, in both senses of the word, but is not accidental: The way these scales are formed is directly related to the Fibonacci sequence.

4.2. **Two Alternative Models.** Before concluding the mathematical development we mention two topics which offer alternative views of the material, without giving any background or development.

One view which offers geometric insights is to consider the lattice $Z \times Z$, isomorphic to P. Each convergent or semi-convergent A/B corresponds to a unimodular transformation of the lattice onto itself, whose matrix has determinant 1 or -1. If $M_{(A,B)}$ is the matrix of this linear transformation,

$M_{(A,B)} = (-1)^k \begin{bmatrix} a_k & A - a_k \\ b_k & B - b_k \end{bmatrix}$ and $\det M_{(A,B)} = (-1)^{k+1}$. The eigenvalues of the first four transformations are $\left(\frac{1}{\phi}\right)^n$ for $n = 0, 1, 2, 3$. It is evident that any of the interval pairs associated with a region form a basis for the lattice $Z \times Z$, that is, for P, where the canonical basis corresponds to the intervals octave and twelfth.

We mention another approach for readers familiar with Farey series. On the basis of our Continued Fraction Theorems 7 and 8 and their converses one can see that in any Farey series the adjacent elements which have between them $\frac{1}{\log_2 3} = \log_3 2$ are always elements B/A, convergents or semi-convergents to $\log_3 2$. One of the properties of adjacent Farey series elements $a/b, c/d$ is that $ad - bc = -1$ when $a/b < c/d$, and many of the continued fraction theorems have equivalents in the theory of Farey series. Less elegant than continued fractions as a point of departure for our purposes, nevertheless it gives an interesting perspective on the sequence of rational approximations to $\log_2 3$.

Although the chain of syllogisms leading to the Chacterization Theorem is long, the idea behind the final result is a simple though powerful one which can be grasped most easily by working with concrete examples. The reader is urged to work through a few examples, translating into tones, even playing the resulting scales, before going on to the conclusion. The interpretation which follows considers some of the musical implications of the mathematical theory.

---

5. Musical Implications

The central notion of our theory is that every well-formed Pythagorean scale is a smooth scale. In our terms a scale is *smooth* if, in the context of the group within which it is embedded, all of its scale-step intervals are the same. Since the interval between any two adjacent steps in a well-formed scale is always $2^{(a_k((-1)^k) \bmod A} 3^{(b_k((-1)^k) \bmod B}$, in the group $G_{(A,B)}$ the interval is a constant and the scale is smooth. Of course, this is purely a reflection of the way the scale coincides with a part of the cyclic ordering of the group: The Characterization Theorem tells us that when a well-formed scale is mapped into the cyclic group Z_{AB}, the elements of the scale go onto a connected part of Z_{AB}. We can think of a cyclic group as a clock, and the region, hence the well-formed scale, is represented by a portion of the clock.

Smoothness appears to be a highly abstract quality. After all, the primary elements of music are tones and intervals, and another consequence of the theory is that all of theses scales are composed of two distinct intervals. A less abstract smoothness would be involved in scales, not well-formed, such as the whole tone scale in equal temperament, or the equal-tempered chromatic scale, in which the octave is divided into 6 or 12 equal intervals. Then the scale itself is equivalent to a cyclic group.

To the mind, smoothness of the latter type is immediately apparent, while detailed analysis is required to see the smoothness of well-ordered scales. To the ear, however, the smoothness of the equal-interval scales is abstract, that of the diatonic scale immediately and sensorially apprehended. Indeed, the word scale derives from the Latin "*scalae*" meaning steps or ladder, which is in accord with the intuitive sense of the scale as a succession of steps spanning an octave. In the most familiar scales the progression from step to step is felt to be even, natural, smooth. Contrasted with the familiar diatonic and pentatonic scales, for example, the whole-tone and equal-tempered chromatic scales are considered awkward or in some sense less satisfactory. Musicians tend to take this contrast for granted, but it is occasionally considered to by paradoxical. Douglas Hofstadter, for example, writes:

> The seven intervals between successive notes of the diatonic scale are not all equal. Some are twice as large as others, yet to the ear there is a perfect intuitive logic to it. Rather paradoxically, in fact, most people can sing a major scale without any trouble, uneven intervals notwithstanding, but few can sing a chromatic scale accurately, even though it 'ought' to be much more straightforward—or so it would seem, since all its intervals are exactly the same size.[16]

Trained listeners hearing a mode of the diatonic scale readily distinguish between whole steps and half steps, and others, perhaps untrained but sensitive to music, hear a distinct difference in quality between major and minor second. In fact, "quality" is the word musicians use to characterize the size of a given diatonic interval: The major third differs from the minor third in *quality.* Finally, in experiments with musically untrained listeners we find that a large majority identifies the mixolydian or phrygian mode of the diatonic scale as a scale of *equal* steps, as against the whole-tone scale or some non-well-ordered arrangements of 5 whole steps and 2 half steps, such as one containing two half steps in a row. Such scales are perceived as being composed of unequal steps, wrongly in the former case and rightly in the latter case.

We would argue that most listeners in fact perceive the different scale-steps not so much in terms of a difference in size, but more as a difference in quality, though the musician is so thoroughly trained to think in terms of interval size that this becomes second nature.

To illustrate the dual nature of a well-ordered scale and also perhaps the way the scale is perceived, consider the following spatial analogy: Imagine a being, X, who lives entirely in a two-dimensional plane, and another, Y, who inhabits three-dimensional space. Y moves two and a half units in the plan inhabited by X and turns the corner moving a half unit further on a line perpendicular to X's plane. From the limited point of view of X, Y traveled two and a half units and disappeared, while Y experience traveling three equal units, with the third step distinguished by the sensation of turning the corner.

The experience of listening to the tetrachords of the major scale might be described similarly: From a one-dimensional point of view, we measure two whole

[16] Douglas Hofstadter *Metamagical Themas* p. 174 Bantam Books 1986

steps and a half step, but the ear experiences the tetrachord of the scale as step, step, step, with a difference in quality on the third step, as if turning a corner. In like manner the description of the diatonic scale as a well-formed scale affords two viewpoints: From one point of view it is made up of two distinctly different intervals, from another it is composed of one interval which admits two equivalent representations.

The concept of dimension also carries over: While the twelve tones of the equal-tempered chromatic scale can be represented as the powers of $\sqrt[12]{2}$ and the six tones of the whole-tone scale by power of $\sqrt[6]{2}$, thus as elements of a one-dimensional space, the tones of Pythagorean space and of any well-formed scale are represented by means of powers of two and three, that is, as points in the two-dimensional lattice $Z \times Z$.

There is a strong suggestion here of a relationship between the mathematical properties of well-formed scales and the audible properties of certain well-formed scales. The correlation remains a hypothesis, but suggests a point of departure for cognitive and perceptual research as well as for studies in music theory.

In fact, there is no generally accepted explanation for the selection of certain tones as constituents of a scale. A priori, there is no compelling reason why the notes of the diatonic scale, for example, should be preferred to some other selection of seven pitches. Attempts have been made to find the basis for the scale in the harmonic (overtone) series, and, as we will see, the axioms for well-ordered scales have their origins in the fact of the harmonic series, but the harmonic series by itself has not provided a satisfactory explanation for the particular notes of the scale. In fact, by itself it predicts a rather different scale. Taking C as unity, the tones associated with the beginning of the harmonic series are as follows:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C_0	C_1	G_1	C_2	E_2^*	G_2	-$B\flat_2$	C_3	D_3	E_3^*	-$F\sharp_3$	G_3	-A_3	-$B\flat_3$	B_3^*	C_4

The E in this system is slightly lower than the E of the Pythagorean system, the Pythagorean major third being $(9/8) \cdot (9/8) = 81/64$ and the just major third being $5/4 = 80/64$, so the variants differ by 81/80 the syntonic comma. There is no objection to deriving this tone from the harmonic series, since *in a harmonic context* the ear prefers the major third of just intonation. However, the other prime numbers greater than 5 and their multiples yield tones distinctly flatter than the tones of the scale which they approximate, and the tone F, a perfect fourth above C, will never appear in this system.

Much more could be said about the harmonic series and there is an extensive literature on this topic, but for our purposes it suffices to say that while the laws of harmony are governed by this acoustical fact of nature, we must look elsewhere for the rules which govern, not the construction of melodies, but the materials of which melody is constructed. Of course melody is not utterly in conflict with harmony, and therefore it is not surprising that the strongest harmonic facts, the octave and the twelfth, should also be the forces which shape the elements of melody, the tones and intervals of the scale.

We are left with the implication that the ear, that is, the mind, is drawn in some way to the structures of well-formed scales, perhaps responding to the deep smoothness of these scales.The only other explanation for the multiplicity of well-formed scales in world music would be that they were consciously constructed according to these mathematical principles. In some cases this is undoubtedly the way they arose, although the mathematics of the age may not have permitted the formulation of these principles. In the case of the Chinese 53-tone scale it is reasonable to assume that the scale was derived theoretically, since the intervals it is made up of are so small. The seventh convergent of $\log_2 3$ is $84/53 = [1; 1, 1, 2, 2, 3, 1]$, and a version of the Chinese 53-tone scale is given by Daniélou spans an entire region, though no hint is given as to how it was derived. The Chinese had very precisely tuned bells dating back to the 5^{th} century BCE (cf *Scientific American*, Dec. 1984) and were extraordinarily adept calculators. Evidentally some Chinese theorists extrapolated from the structure of the pentatonic, diatonic, and chromatic scales to form this precisely calibrated scale.

The same is likely to be true in the case of the Arabic 17-tone scale. We may recall here the distinction between scales which are melodic structures in themselves and scales which constitute a repertoire of available tones. Curt Sachs, for example, reserves the word scale for the former, referring to latter as a "set of elements." The Arabic 17-tone division of the octave is not a scale, in his terms, but a set of elements, from which may be constructed scales and melodies based on this set; similarly the 53-tone division of the octave.

We posited 100 modal varieties of well-formed scales from world music. Since there are B different modes of the well-formed scale associated with $R_{(A,B)}$, we obtain the total number of modes by summing the values of B associated with musically significant regions: $1 + 2 + 3 + 5 + 7 + 12 + 17 + 53 = 100$. However, most of these modes are of an artificial nature: Unless the scale has the status of a melodic structure by itself, its modes do not have much musical interest. The scales which are decidedly of melodic character are the pentatonic and diatonic scales, and here it seems unlikely that some sort of theory of well-formed scales preceded the use of these scales.

If well-formed scales are preferred by the ear, what accounts for the pre-eminent status of the diatonic scale in particular out of all possible well-formed scales? The answer is undoubtedly complicated. The limits of the human ear's ability to discriminate between different pitches is probably the most important. Intervals which are extremely small do not register as discrete entities. The relative sizes of the two intervals in a well-formed scale may play a role as well: In the Arabic Region, for example, the two intervals are the chromatic half step and the Pythagorean comma, whose ratio is approximately $3 : 1$, and in the Structurals the intervals are the perfect fourth and the whole step, whose ratio is about $5 : 2$. The ratio in the Pentatonic region is about $3 : 2$, and in the Diatonics, between the whole step and the diatonic half step the ratio is close to $2 : 1$. In the Chromatics, however, the ratio of the chromatic to diatonic half-step is close to $1 : 1$, that is, the ear finds it difficult to distinguish between the two intervals. As k increases without bound, $\frac{2^{a_k}}{3^{b_k}} / \frac{3^{B-b_k}}{2^{A-a_k}}$ approaches unity, so when A/B is a higher order convergent, the two intervals which make up the region are approximately the same size. So it may be that not only the absolute size but also the relative size of the intervals plays a role in determining which well-formed scales are musically useful: If the sizes of the two intervals are too different or too much the same they do not yield a melodic scale.

We refer to the culmination with the diatonic scale of the progression of melodic well-formed scales as *closure*. There are several mathematical correlations with the phenomenon of closure. We have mentioned the correspondence of the first five regions with the convergents or semi-convergents to the golden number,

$\phi = (1 + \sqrt{5})/2$. This correspondence breaks down with the Diatonic Region $R_{(11,7)}$, so the Diatonic Region is the last region where both intervals correspond to convergents of ϕ.

A related feature is the status of the Diatonic Region as the last region in the family of regions generated by the whole step. Recall that each convergent of $\log_2 3$ generates a family
$F_{(a_k,b_k)} = \{R_{(A,B)} \mid A = na_k + a_{k-1}, B = nb_k + b_{k-1};\ 0 \le n \le t_{k+1} + 1\}$ where t_{k+1} is the $(k+1)^{st}$ partial quotient of $\log_2 3$. Then the whole step $\rho_{(3,2)}$ generates $F_{(A,B)} = \{R_{(3n+2,2n+1)} \mid 0 \le n \le 3\}$, that is the set consisting of $R_{(2,1)}$, $R_{(5,3)}$, $R_{(8,5)}$, $R_{(11,7)}$. [See diagram on page 35.] Similarly, the Arabic Region is the last member of the family generated by the diatonic half step, $\rho_{(8,5)} = 256/243$, and the Chinese system of 53 tones is the final element in $F_{(19,12)}$, the family generated by the Pythagorean comma.

Evidentially an equivalent formulation of the closure problem is to ask why the whole step should be considered the fundamental unit, the basic scale-step. That this was the way the ancient Greek theorists conceived of the whole step is evident in their terminology, in which the whole step is $\delta\iota\alpha\tau o\nu o\nu$ (*diatonon*), "through a tone," and the diatonic half step is $\lambda\iota\mu\mu\alpha$ (*limma*), "the remainder." Despite the extrapolation to systems of a 17-tone division or a 53-tone division of the octave by Arabian and Chinese theorists, in the music of their cultures diatonic and pentatonic scales predominate, at least in the sense that the whole step is the basic scale-step.

The way the whole step family imposes a type of mathematical closure on the infinite process which gives rise to the set of regions is an example of a principle of intersection, the coincidence of mathematical substructures within a larger system. The region itself, as we have stated, is the intersection of two types of ordering within the group $G_{(A,B)}$: the coincidence of the linear ordering of P with what might be called the canonical cyclic ordering of $G_{(A,B)}$. Goethe's image of the arrow and the circle, *Kreis und Pfeil*, comes to mind.

The way the octave frame superimposed on the region gives rise to the modal system of well-formed scales is a further intersection. While the mathematical notion of region owes nothing to the number 2 in particular, as was mentioned in footnote 13 on page 26, in purely mathematical terms knowledge of a region over the interval of an octave is in a sense the minimal information necessary to know the entire region. We may mention here that the number of each type of interval within an octave of any well-formed scale is invariant: in one octave of a well-formed scale associated with $R_{(A,B)}$ there are $B - b_k$ intervals of the form $\dfrac{2^{a_k}}{3^{b_k}}$ or $\dfrac{3^{b_k}}{2^{a_k}}$ and b_k of the form $\dfrac{3^{B-b_k}}{2^{A-a_k}}$ or $\dfrac{2^{A-a_k}}{3^{B-b_k}}$. Assume k is even. Then $3^{b_k} > 2^{a_k}$, and $\left(\dfrac{3^{b_k}}{2^{a_k}}\right)^{B-b_k} \cdot \left(\dfrac{2^{A-a_k}}{3^{B-b_k}}\right)^{b_k} = \dfrac{3^{b_k(B-b_k)}}{3^{(B-b_k)b_k}} \cdot \dfrac{2^{(A-a_k)b_k}}{2^{a_k(B-b_k)}} = \dfrac{2^{(A-a_k)b_k}}{2^{a_k(B-b_k)}} = 2^{Ab_k - Ba_k} = 2^{(-1)^k} = 2$. Similarly if k is odd. The acoustic fact that the octave is determined by the ratio $2:1$ and that the octave is the equivalence relation between tones motivates the concept of a modal system obtained by taking the various intersections of the octave and the region.

Similarly, the Structural Region coincides with the harmonic series in the sense that the intervals possible within the structural system, the octave, fifth, fourth, and whole step, correspond to the ratios $2:1$, $3:2$, $4:3$, and $9:8$, which are all in the early part of the harmonic series, that is satisfying the so-called rule of small numbers.

The diatonic system is at the culmination of the pattern of growth following the lines of the whole step family and the point at which the Fibonacci pattern of growth leading to the golden section diverges from the pattern leading to $\log_2 3$. Up to this point these patterns are superimposed on the $\log_2 3$ continued fraction pattern.

In the *Timaeus*, the formation of the diatonic scale is used as a metaphor for the myth of creation. It is interesting to note that all of the intervals mentioned by Plato correspond to convergents of the golden number, in our sense of considering the ratios of the exponents. "And he went on to fill all the intervals of 4/3 with the interval 9/8, leaving over in each a fraction. This remaining interval of the fraction had its terms in the numerical proportion of 256 to 243. By this time the mixture from which he was cutting these portions was all used up."

Is this order of the diatonic scale an example of *kosmos* or *taxis*, a spontaneously occuring order or a consciously constructed order? The perennial question of mathematics likewise presents itself: Is this order discovered or invented?

We hasten to acknowledge the claims of culture in the formation of the scales and intervals which "provide music with an initial level of articulation," in the words of Lévi-Strauss.[17] We can see that the Pythagorean diatonic system has deep roots in our culture, and in a broad sense any scale and any mathematical formulation is an aspect of culture. Whether discovery or invention, however, that the Pythagorean system should contain within it such subtle and fascinating structures as well-formed scales confers further authority upon it and upon the Pythagorean conception of music as the power of number made manifest.

E-mail address: Norman Carey ncarey@gc.cuny.edu, David Clampitt Clampitt.4@osu.edu.

[17] *The Raw and the Cooked* Claude Lévi-Strauss, tr. John and Doreen Weightsman Harper and Row 1969