Fractal homeomorphism for bi-affine iterated function systems

Michael Barnsley*

Mathematical Sciences Institute,
The Australian National University,
Canberra, ACT 0200, Australia
E-mail: michael.barnsley@anu.edu.au
*Corresponding author

Andrew Vince

Department of Mathematics,
University of Florida,
438 Little Hall, P.O. Box 118105,
Gainesville, FL 32611-8105, USA
E-mail: avince@ufl.edu

Abstract: The paper concerns fractal homeomorphism between the attractors of two bi-affine iterated function systems. After a general discussion of bi-affine functions, conditions are provided under which a bi-affine iterated function system is contractive, thus guaranteeing an attractor. After a general discussion of fractal homeomorphism, fractal homeomorphisms are constructed for a specific type of bi-affine iterated function system.

Keywords: iterated function system; fractal transformation; bi-affine transformation.

Reference to this paper should be made as follows: Barnsley, M. and Vince, A. (2013) 'Fractal homeomorphism for bi-affine iterated function systems', *Int. J. Applied Nonlinear Science*, Vol. 1, No. 1, pp.3–19.

Biographical notes: Michael Barnsley is a Professor at the Mathematical Sciences Institute, Australian National University where he teaches and researches fractal geometry. Previously, he was an itinerant post-doc in England, France, and Italy (1973–1979); a Professor at Georgia Tech (1979–1991); and an entrepreneur and Chief Scientist. The company he co-founded, Iterated Systems (1987–1998), developed fractal image compression technology, used for example by Microsoft.

Andrew Vince is a Professor at the University of Florida. His mathematical interests are in discrete geometry, fractal geometry in particular, combinatorics and graph theory. He has held visiting positions in Australia, Malawi, New Zealand, Turkey and Uganda.

1 Introduction

The purpose of this paper is to investigate bi-affine iterated function systems and fractal homeomorphisms between their attractors. The class of bi-affine functions from $\mathbb{R}^2$ to $\mathbb{R}^2$ is more general than affine transformations but less general than quadratic transformations. These are the functions $f: \mathbb{R}^2 \to \mathbb{R}^2$ that are, for a fixed x or a fixed y, affine in the other variable:

$$\begin{aligned} &f\left((1-\alpha)x_1+\alpha x_2, y\right) = (1-\alpha) f\left(x_1, y\right) + \alpha f\left(x_2, y\right) \\ &\text{and } f\left(x,(1-\alpha)y_1+\alpha y_2\right) = (1-\alpha) f\left(x, y_1\right) + \alpha f\left(x, y_2\right) \end{aligned} \tag{1}$$

for all $x_1, x_2, y_1, y_2, \alpha \in \mathbb{R}$. Interpreted geometrically, these equations mean that

1 horizontal and vertical lines are taken to lines

2 proportions along horizontal and vertical lines are preserved.

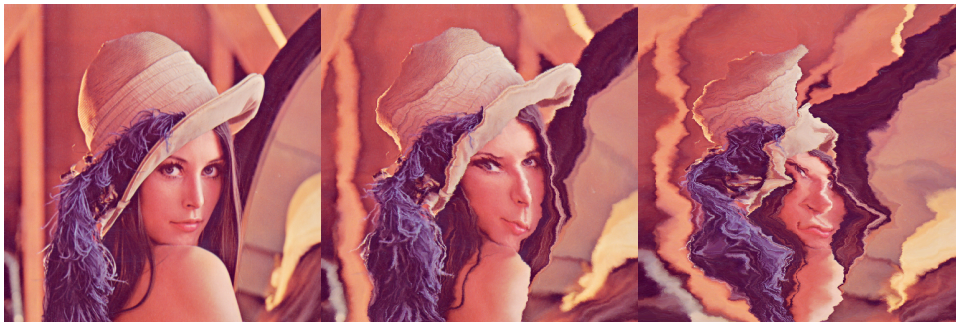
This elementary class of functions, with connections to classic geometric results of Brianchon and Lampert dating back to the 18th century, proves extremely versatile for the applications described in this paper.

Our main motivation for investigating bi-affine functions comes from the representation and transformation of certain fractal images. A standard method for constructing a deterministic self-referential fractal is by an iterated function system (IFS). The attractor of the IFS maybe a fractal. Barnsley (2009a, 2009b) has introduced a method for transforming the attractor of one IFS to the attractor of another IFS, a method that has applications to digital imaging such as image encryption, filtering, compression, watermarking, and various special effects. Figure 1 explained in more detail in Section 5, is obtained by applying such transformations – called fractal homeomorphisms. In constructing a fractal homeomorphism it is convenient to use an IFS whose maps have nice geometric properties but that are not too complicated. Linear transformations are easy to work with and have the property that lines are taken to lines. Affine transformations are not much more complicated than linear transformations and do not have the restriction that the origin be taken to the origin. There is a tradeoff; the more properties required, the more complicated the mapping. In the fractal geometry literature, in particular for fractals constructed using an iterated function system, affine transformations are frequently used. For the applications described in Section 5, requiring slightly more general functions, bi-affine functions are ideally suited.

The paper is organised as follows. The geometry of bi-affine functions is the subject of Section 2. Theorem 1 gives basic properties of a bi-affine function, in particular properties of the folding line and folding parabola. Theorem 2 gives a geometric construction for finding the image of a given point under a bi-affine function and precisely describes the 2-to-1 nature of a bi-affine function. Section 3 provides background on iterated function systems and their attractors. It is a classic result that, if an IFS is contractive, then it has an attractor. Theorem 3 gives fairly general conditions under which a bi-affine IFS is contractive. Fractal homeomorphism between the attractors of two IFSs is the subject of Section 4. The construction of a fractal homeomorphism depends on finding shift invariant sections of the coding maps of the two IFSs. Theorem 4 states that any such shift invariant section comes from a mask (the terms

coding map, section, shift invariant, and mask are defined in Section 4). Theorem 5 concerns how to obtain a fractal homeomorphism between two attractors from the respective sections. The visual representation of a fractal homeomorphism for bi-affine IFSs is the subject of Section 5. Theorem 6 states that a particular type of bi-affine IFS can be used to construct fractal homeomorphisms. The main issue is proving the continuity of the map and its inverse. The pictures in Figure 1 are obtained by this method.

Figure 1 Two fractal homeomorphisms applied to the original picture



2 Geometry of bi-affine functions

Boldface letters represent vectors in $\mathbb{R}^2$. A function $f: \mathbb{R}^2 \to \mathbb{R}^2$ is called bi-affine if it has the form

$$f(x, y) = \mathbf{a} + \mathbf{b}x + \mathbf{c}y + \mathbf{d}xy. \tag{2}$$

It is easy to verify that the class of bi-affine functions is exactly the class characterised by the two properties listed in the introduction. The class of bi-affine functions is not closed under composition, but the composition of a bi-affine and an affine function is bi-affine.

Call a bi-affine function *non-degenerate* if $\mathbf{d} \neq 0$ and neither **b** nor **c** is a scalar multiple of **d**. In particular, neither **b** nor **c** is the zero vector. If $\mathbf{d} = 0$, then f is 'degenerate' in the sense that it is affine and well understood. If $\mathbf{d} \neq 0$ and both **b** and **c** are scalar multiples of **d**, then f is 'degenerate' in the sense that the range of f degenerates to a line. If $\mathbf{d} \neq 0$ and just one of **b** and **c** is a scalar multiple of **d**, then f is 'degenerate' in the sense that the image of the folding line, as defined in the next section, is just a point, a fact that can be verified by equation (4) in the proof of Theorem 1.

Basic properties of bi-affine functions are described in this section. According to statement 2 of Theorem 1 below, the image of a line L under a bi-affine function is a parabola. Such a parabola can be *degenerate* in the sense that it is either a line (focal distance 0) or a line that doubles back on itself (focal distance ∞). The first case occurs if and only if L is parallel to either the x or y-axis. An example of the second case is $(X, Y) = f(x, y) = (-x + xy, -1 - y + xy)$, in which case the image of the line $y = x + 1$ is

given by the parametric equation $X = t^2$, $Y = t^2 - 2$, a degenerate parabola with vertex at $(0, -2)$ that doubles back along the line $y = x - 2$.

For vectors **a** and **b**, let $|\mathbf{a}\ \mathbf{b}|$ denote the determinant of the matrix whose columns are **a** and **b**. For a non-degenerate bi-affine function f, call the line L_f with equation

$$|\mathbf{b}\ \mathbf{d}|\,x + |\mathbf{d}\ \mathbf{c}|\,y = |\mathbf{c}\ \mathbf{b}| \tag{3}$$

the *folding line* of the function f. Note that, by the non-degeneracy of f, neither $|\mathbf{b}\ \mathbf{d}|$ nor $|\mathbf{d}\ \mathbf{c}|$ is zero, and hence the folding line is not parallel to either the x or y-axis. The terminology 'folding line' is justified by the next theorem, in particular statement 4. The image $P_f = f(L_f)$ of the folding line is called the *folding parabola*, which, according to statement 3 is non-degenerate. Let L_+ and L_- denote the closed half spaces above and below L_f, respectively.

Theorem 1: If f is a non-degenerate bi-affine map, then

1 any line parallel to either coordinate axis is mapped to a line

2 any line is mapped to a, possibly degenerate, parabola

3 the folding parabola $P_f := f(L_f)$ is non-degenerate

4 the map f is injective when restricted to either L_+ or L_-.

Proof: Statement (1) follows from the fact that, for a fixed x or a fixed y, a bi-affine function is an affine function in the other variable.

The image of a line under a bi-affine transformation has a parametric equation of the form $f(t) = \mathbf{u} + \mathbf{v}t + \mathbf{w}t^2$, which is a, possibly degenerate, parabola.

To show that P_f is non-degenerate, first note that the image of the line L_f under f is given by the parametric equation $(X(x), Y(x))$ with parameter x by

$$\begin{aligned}(X,Y) &= f(x,x)\\ &= \frac{1}{|\mathbf{d}\ \mathbf{c}|}\Big[\big(|\mathbf{d}\ \mathbf{c}|\,\mathbf{a} + |\mathbf{c}\ \mathbf{b}|\,\mathbf{c}\big) + \big(|\mathbf{d}\ \mathbf{c}|\,\mathbf{b} - |\mathbf{b}\ \mathbf{d}|\,\mathbf{c} + |\mathbf{c}\ \mathbf{b}|\,\mathbf{d}\big)x - |\mathbf{b}\ \mathbf{d}|\,\mathbf{d}x^2\Big] \qquad (4)\\ &= \frac{1}{|\mathbf{d}\ \mathbf{c}|}\Big[\big(|\mathbf{d}\ \mathbf{c}|\,\mathbf{a} + |\mathbf{c}\ \mathbf{b}|\,\mathbf{c}\big) + 2|\mathbf{d}\ \mathbf{b}|\,\mathbf{c}x - |\mathbf{b}\ \mathbf{d}|\,\mathbf{d}\,x^2\Big],\end{aligned}$$

the first equality obtained simply by substituting from equation (3) into equation (2) and the second equality by a direct calculation. The tangent vector to this parabola is

$$T(x) = \frac{2|\mathbf{d}\ \mathbf{b}|}{|\mathbf{d}\ \mathbf{c}|}(\mathbf{c} - \mathbf{d}x).$$

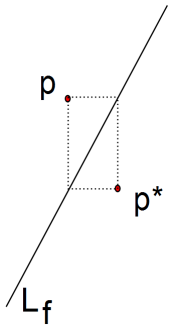
By the non-degeneracy of f, the vectors **c** and **d** are linearly independent, implying that the direction of the tangent vector is not constant. Hence the parabola is not degenerate.

The Jacobian determinant of a bi-affine function is $|\mathbf{b}\ \mathbf{d}|x + |\mathbf{d}\ \mathbf{c}|y - |\mathbf{c}\ \mathbf{b}|$, which is nonzero except on the folding line. It follows from the inverse function theorem that f is injective when restricted to either L_+ or L_-. □

A parabola P divides the plane into two regions; let $\widehat{P}$ denote the closed region 'outside' P (and including P). The set $\widehat{P_f}$ will be called simply the *parabolic region of f*. For a point $\mathbf{p} = (x, y) \in \mathbb{R}^2$, let

$$\mathbf{p}^* = \left(\frac{|\mathbf{c}\ \mathbf{b}| - |\mathbf{d}\ \mathbf{c}|\, y}{|\mathbf{b}\ \mathbf{d}|}, \frac{|\mathbf{c}\ \mathbf{b}| - |\mathbf{b}\ \mathbf{d}|\, x}{|\mathbf{d}\ \mathbf{c}|} \right). \tag{5}$$

Figure 2 The location of the point $\mathbf{p}^*$ with respect to the folding line L_f (see online version for colours)



This somewhat complicated formula is merely an analytic expression of the simple geometry shown in Figure 2. Statement 3 in Theorem 2 below makes precise the 2-to-1 nature of a bi- affine function given in statement 4 of Theorem 1. Statement 1 in Theorem 2 is illustrated in Figure 3. Statement 2 gives a geometric construction of the image of a given point under a bi-affine function and is illustrated in Figure 4.

Figure 3 Folding parabola P_f, parabolic region $\widehat{P_f}$, tangent line T_A (see online version for colours)

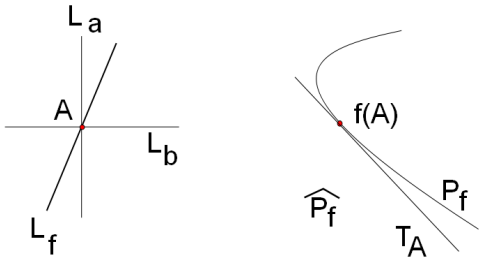
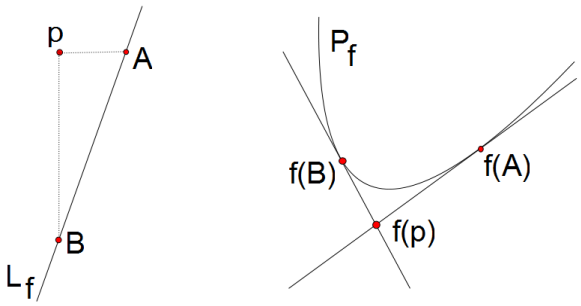


Figure 4 The image of a point **p** under a bi-affine function (see online version for colours)



Note: The point $f(\mathbf{p})$ has 'coordinates' (A, B).

Theorem 2 Assume that f is a non-degenerate bi-affine map. We use the notation V_a for the line $x = a$ and H_b for the line $y = b$.

1 if $A = (a, b)$ is any point on the folding line L_f and T_A is the tangent line to P_f at $f(A)$, then $T_A = f(V_a) = f(H_b)$

2 if $\mathbf{p} = (a, b) \in \mathbb{R}^2$, then $f(\mathbf{p}) = T_A \cap T_B$, where $A = L_f \cap V_a$ and $B = L_f \cap H_b$

3 $f(L_+) = f(L_-) = \widehat{P_f}$ and, in particular, $f(\mathbf{p}) = f(\mathbf{p}^*)$ for all $\mathbf{p} = (x, y) \in \mathbb{R}^2$.

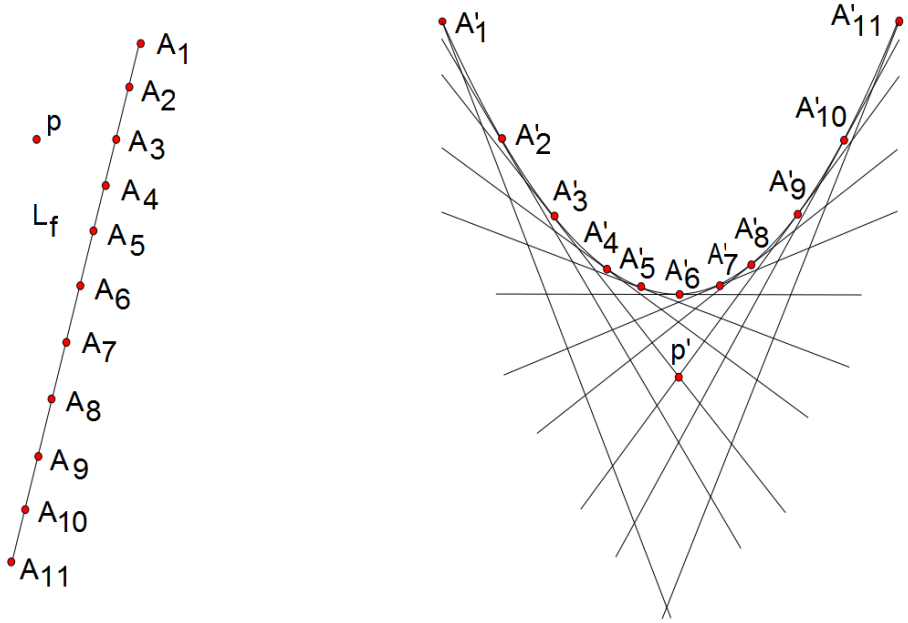
Proof: Concerning statement 1, any horizontal or vertical line L intersects L_f in a single point. Therefore $f(L)$ intersects $f(L_f) = P_f$ in a single point, which implies that $f(L)$ is tangent to P_f. Since the intersections $P_f \cap f(V_a)$ and $P_f \cap f(H_b)$both consist of the same single point, $f(V_a)$ and $f(H_b)$ both equal the tangent line T_A to P_f at $f(A)$.

By statement 1, the point $f(\mathbf{p})$ lies both on the tangent to P_f at $f(A)$ and on the tangent to P_f at $f(B)$. This proves statement 2.

Consider $\mathbf{p}$ and $\mathbf{p}^*$ as in Figure 2. Statement 2 implies that $f(\mathbf{p}) = T_A \cap T_B = f(\mathbf{p}^*)$. In particular $f(L_+) = f(L_-)$. Since the union of all tangents to P_f is the parabolic region $\widehat{P_f}$ of f, we have $f(L_+) = f(L_-) = f(\mathbb{R}^2) = \widehat{P_f}$. □

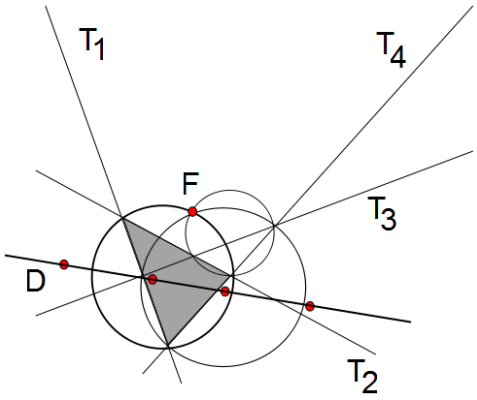
It is a consequence of Theorem 2 that the parabolic region can be coordinatised as follows. Each point $\mathbf{p} \in \widehat{P_f}$ has a unique set $\{A, B\}$ of (unordered) coordinates where A and B are points on the folding line. Specifically $\mathbf{p} = T_A \cap T_B$. This is illustrated in Figures 4 and 5.

Figure 5 Coordinatisation of $\widehat{P_f}$ (see online version for colours)



Notes: $A_i' = f(A_i)$, $1 \le i \le 11$, and $\mathbf{p}' = f(\mathbf{p})$. The point $\mathbf{p}'$ has coordinates $\{A_3, A_9\}$.

Figure 6 Construction of the folding parabola: focus F and directrix D (see online version for colours)



The folding parabola P_f can be constructed geometrically as follows (see Figure 6). Choose four pairwise distinct points A_1, A_2, A_3, A_4 on the folding line and construct the four tangent lines T_{A_1}, T_{A_2}, T_{A_3}, T_{A_4}. It is a direct consequence of the classic Brianchon Theorem (actually the converse) that there is a unique parabola with these lines as tangents. According to what has been shown above, this must be the folding parabola. The parabola can be explicitly constructed using a theorem of Lambert. According to Lambert, the circumcircle of a tangent triangle of a parabola (see Figure 6) goes through

the focus of the parabola. So the focus F is determined as the intersection of three such circumcircles. Then reflect F about two of the tangents to get two points on the directrix.

3 Iterated function systems

This section reviews the standard notation and definitions related to IFS. These concepts are then applied to the bi-affine case.

Let $\mathbb{X}$ be a complete metric space. If $f_m : \mathbb{X} \to \mathbb{X}$, $m = 1, 2, \ldots, M$, are continuous mappings, then $\mathcal{F} = (\mathbb{X}; f_1, f_2, \ldots, f_M)$ is called an *IFS*. An IFS that consists of bi-affine functions will be called a *bi-affine* IFS. To define the attractor of an IFS, first define

$$\mathcal{F}(B) = \bigcup_{f \in \mathcal{F}} f(B)$$

for any $B \subset \mathbb{X}$. By slight abuse of terminology we use the same symbol $\mathcal{F}$ for the IFS, the set of functions in the IFS, and for the above mapping. For $B \subset \mathbb{X}$, let $\mathcal{F}^k(B)$ denote the k-fold composition of $\mathcal{F}$, the union of $f_{i_1} \circ f_{i_2} \circ \cdots \circ f_{i_k}(B)$ over all finite words $i_1\, i_2 \cdots i_k$ of length k. Define $\mathcal{F}^0(B) = B$. A non-empty compact set $A \subset \mathbb{X}$ is said to be an *attractor* of the IFS $\mathcal{F}$ if

1 $\mathcal{F}(A) = A$

2 $\lim_{k\to\infty} \mathcal{F}^k(B) = A$, for all compact sets $B \subset \mathbb{X}$, where the limit is with respect to the Hausdorff metric.

Attractors for bi-affine IFSs consisting, respectively of 2, 3 and 4, functions are shown in Figure 7.

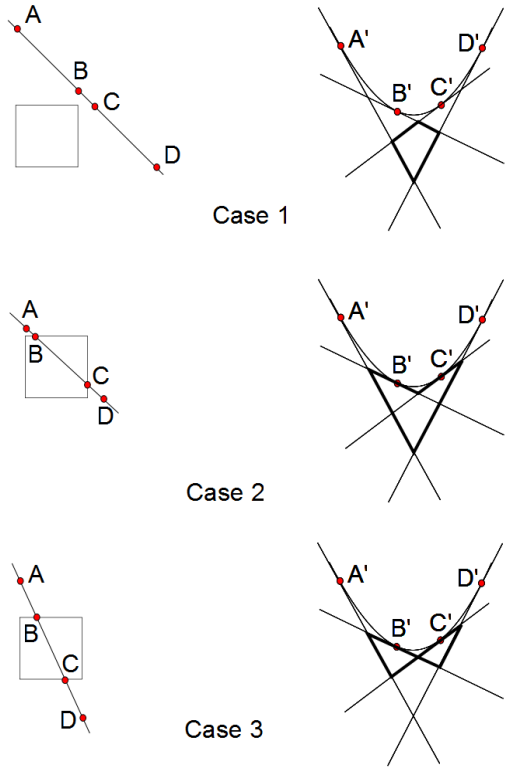
Figure 7 Attractors of IFSs consisting of two, three, and four bi-affine functions, respectively

A function $f : \mathbb{X} \to \mathbb{X}$ is called a *contraction* with respect to a metric d if there is an s, $0 \le s < 1$, such that $d(f(x), f(y)) \le sd(x, y)$ for all $x, y \in \mathbb{R}^n$. An IFS with the property that

each function is a contraction will be called a contractive IFS. In his seminal paper Hutchinson (1981) proved that a *contractive* IFS on a complete metric space has a unique attractor. Theorem 3 below gives fairly general conditions under which a bi-affine IFS is contractive.

Let □ denote the unit square with vertices (0, 0), (1, 0), (1, 1), (0, 1). The shape of the image of □ under a non-degenerate bi-affine map f depends on the location of the folding line L_f relative to □. It follows from Theorem 2 that there are three possible cases as shown in Figure 8. It is only in Case 1 (L_f disjoint from the interior of □) that the image of the sides of □ form a convex quadrilateral. If this is the case, call f *proper*.

Figure 8 The image of the sides of the unit square (on the left) is shown by thick lines (on the right): $A' = f(A)$, $B' = f(B)$, $C' = f(C)$, $D' = f(D)$ (see online version for colours)



Note that the unique bi-affine function taking (0, 0), (1, 0), (1, 1), (0, 1) to the points $\mathbf{p}_0$, $\mathbf{p}_1$, $\mathbf{p}_2$, $\mathbf{p}_3$, respectively, is

$$f(x,y) = \mathbf{p}_0 + (\mathbf{p}_1 - \mathbf{p}_0)x + (\mathbf{p}_3 - \mathbf{p}_0)y + (\mathbf{p}_2 + \mathbf{p}_0 - \mathbf{p}_1 - \mathbf{p}_3)xy. \tag{6}$$

Theorem 3: Let $f(x, y) = \mathbf{p}_0 + (\mathbf{p}_1 - \mathbf{p}_0)x + (\mathbf{p}_3 - \mathbf{p}_0)y + (\mathbf{p}_2 + \mathbf{p}_0 - \mathbf{p}_1 - \mathbf{p}_3)xy$ be a proper, non-degenerate bi-affine function. If there is an s, $0 \le s < 1$, such that

$$\left|\mathbf{p}_{i+1}-\mathbf{p}_i\right| \le s, \quad \left|\mathbf{p}_{i+2}-\mathbf{p}_i\right| \le \sqrt{2}s, \quad \left|\mathbf{p}_{i+1}+\mathbf{p}_{i-1}-2\mathbf{p}_i\right| \le \sqrt{2}s$$

for $i = 0, 1, 2, 3 \pmod 4$, then f is a contraction on $\square$.

In terms of the quadrilateral $\mathbf{p}_0\ \mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3$, the first of the three inequalities states that each side has length less than or equal to s, the second that each diagonal has length less than or equal to $\sqrt{2}s$, and the third that the vector sum of any two incident sides has length less or equal to $\sqrt{2}s$. For a bi-affine function taking $\square$ into itself, for example, these conditions are not too restrictive. Two lemmas help in proving Theorem 3.

Lemma 1: If f is a proper, non-degenerate bi-affine function, then f is injective when restricted to $\square$.

Proof: By the comments above, the folding line L_f lies outside the interior of the square $\square$. The lemma then follows by statement 4 of Theorem 1. □

Lemma 2: Let f be a non-degenerate bi-affine function that is injective on $\square$, and let $W \subseteq \square$ be any rectangle with sides parallel to the x and y-axes and with diagonal of length $\rho(W)$. Further let $\rho_1(W)$ and $\rho_2(W)$ be the lengths of the two diagonals of $f(W)$ and let

$$M\left(\rho_0\right) = \max_{W:\rho(W)=\rho_0} \max\left(\rho_1(W), \rho_2(W)\right).$$

For any $\rho_0 \le \sqrt{2}$, if W_0 is a rectangle that maximises $M(\rho_0)$, then W_0 and $\square$ have a common vertex.

Proof: If one side of W_0 lies on the line $x = 0$ or $x = 1$ and another side lies on $y = 0$ or $y = 1$, then the proof is complete. So, without loss of generality, assume that W_0 has no side that lies on $x = 0$ or $x = 1$. Let V_0 be the rectangle bounded by the lines $x = 0$, $x = 1$ and the lines determined by the upper and lower sides of W_0. Let A, B, C, D be the vertices of $f(V_0)$. Then by the conditions 1 in the introduction, there is an α such that the four vertices of $f(W_0)$ are $(1 - \alpha)A + \alpha B$, $(1 - \alpha - \Delta)A + (\alpha + \Delta)B$, $(1 - \alpha - \Delta)C + (\alpha + \Delta)D$, $(1 - \alpha)D + \alpha C$, where Δ is the horizontal length of W_0. As α varies between 0 and α_0, the rectangle W_0 shifts left or right, from the extreme left side of $\square$ to the extreme right side of $\square$. The lengths of the two diagonals of $f(W_0)$ are $|U_1 + (B + C - A - D)\ \alpha|$ and $|U_2 + (B + C - A - D)\ \alpha|$, where vectors U_1 and U_2 depend on A, B, C, D and Δ. As α varies in the range $0 \le \alpha \le \alpha_0$, the quantities $U_1 + (\mathrm{B} + \mathrm{C} - \mathrm{A} - \mathrm{D})\ \alpha$ and $U_2 + (\mathrm{B} + \mathrm{C} - \mathrm{A} - \mathrm{D})\ \alpha$ describe (parallel) line segments. Hence the maximum of $|U_1 + (B + C - A - D)\ \alpha|$ and $|U_2 + (B + C - A - D)\ \alpha|$ occur at an end, i.e. $\alpha = 0$ or $\alpha = \alpha_0$, contradicting the assumption that W_0 has no side that lies on $x = 0$ or $x = 1$. □

Proof (of Theorem 3): It is sufficient to show that there is an s, $0 \le s < 1$ such that $|f(x, y) - f(x', y')| \le s\,|(x, y) - (x', y')|$. Let W_0 be the rectangle whose diagonal is the line segment joining (x, y) and (x', y'). By Lemma 1, the function f is injective on □, and by Lemma 2, the quantity

$$R : \frac{\left|f(x,y) - f\left(x',y'\right)\right|}{\left|(x,y) - \left(x',y'\right)\right|}, (x,y) \neq \left(x',y'\right)$$

is maximised when W_0 and □ have a common vertex, i.e. when W_0 lies on the corner of □. Without loss of generality, it may be assumed to be the lower left corner. Otherwise, replace f with the composition of f with the rotation that moves the lower left corner to the relevant corner. Now let $\Delta x = |x' - x|$ and $\Delta y = |y' - y|$. With $\mathbf{b} = \mathbf{p}_1 - \mathbf{p}_0$, $\mathbf{c} = \mathbf{p}_3 - \mathbf{p}_0$, $\mathbf{d} = \mathbf{p}_2 + \mathbf{p}_0 - \mathbf{p}_1 - \mathbf{p}_3$ and setting $r = \Delta y/\Delta x$, we have two possible formulas for R:

$$R^2 = \frac{\left|f(\Delta x, \Delta y) - f(x', y')\right|^2}{|(\Delta x, \Delta y)|^2} = \frac{1}{1+r^2}\,|\,\mathbf{b} + r\mathbf{c} + r\Delta x\mathbf{d}\,|^2$$

or

$$R^2 = \frac{\left|f(\Delta x, 0) - f(0, \Delta y)\right|^2}{|(\Delta x, \Delta y)|^2} = \frac{1}{1+r^2}\,|\,\mathbf{b} - r\mathbf{c}\,|^2 .$$

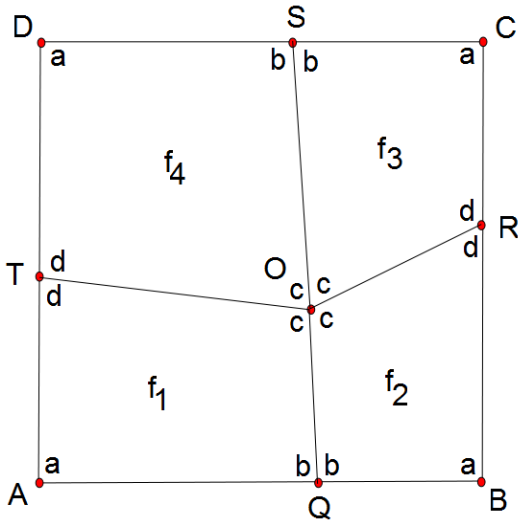
Since $0 < \Delta x \le 1$, the quantity R^2 in the first formula is maximised when either $\Delta x = 0$ or $\Delta x = 1$. So it is now sufficient to prove that

$$\max_{r>0} = \left(\frac{1}{1+r^2}\,|\,\mathbf{b} + r\mathbf{c}\,|^2, \frac{1}{1+r^2}\,|\,\mathbf{b} + r\mathbf{c} + r\mathbf{d}\,|^2, \frac{1}{1+r^2}\,|\,\mathbf{b} - r\mathbf{c}\,|^2 \right) < 1.$$

The inequalities in the statement of the theorem, in terms of **b**, **c**, **d** are that $|\mathbf{b}|$, $|\mathbf{c}|$, $|\mathbf{b} + \mathbf{d}|$, $|\mathbf{c} + \mathbf{d}|$ are all less than or equal to s and $|\mathbf{b} - \mathbf{c}|$, $|\mathbf{b} + \mathbf{c} + \mathbf{d}|$, $|\mathbf{b} + \mathbf{c}|$, $|\mathbf{b} - \mathbf{c} + \mathbf{d}|$, $|\mathbf{c} - \mathbf{b} + \mathbf{d}|$, $|\mathbf{b} + \mathbf{c} + 2\mathbf{d}|$ are all less than or equal to $\sqrt{2}s$. Maximising the three quantities over r, in order to verify the above inequality, is a not quite trivial calculus problem whose details are omitted. □

Example 1: Consider a bi-affine IFS $\mathcal{F} = \{□; f_1, f_2, f_3, f_4\}$, where the four functions are determined by the images of the four vertices of □ as shown in Figure 9. Each of the four images $f_i(□)$, $i = 1, 2, 3, 4$, of the square contains exactly one vertex of □. The attractor of $\mathcal{F}$ is □ itself. For 'most' choices of the centre and side points, $\mathcal{F}$ satisfies the conditions of Theorem 3 and hence $\mathcal{F}$ is a contractive IFS. Why this simple example should be of interest is the subject of the next two sections.

Figure 9 This IFS consists of four bi-affine functions (see online version for colours)



Notes: Points labelled with lower case letters are the images of points labelled with upper case letters. The attractor is the unit square.

4 Fractal homeomorphism

In the case of a contractive IFS, it is possible to assign to each point of the attractor an 'address'. Given two IFSs $\mathcal{F}$ and $\mathcal{G}$ with respective attractors A_F and A_G, a fractal homeomorphism is basically a homeomorphism $h : A_F \to A_G$ that sends a point in A_F to the point in A_G with the same address. To make this notion precise, let $\mathcal{F} = (\mathbb{X}; f_1, f_2, ..., f_N)$ be a contractive IFS on a complete metric space $\mathbb{X}$ with attractor A. Let $\Omega = \{1, 2, \ldots, N\}^\infty$ denote the set of infinite strings using symbols 1, 2, . . ., N, and for $\sigma \in \Omega$, let $\sigma|_k$ denote the string consisting of the first k symbols in σ. Moreover, if $\sigma = i_0\, i_1, i_2 \cdots$, then we use the notation

$$f_{\sigma|_n} := f_{i_0} \circ f_{i_1} \circ \cdots \circ f_{i_n}.$$

Now define a map $\pi : \Omega \to A$, called the *coding map*, by

$$\pi(\sigma) := \lim_{k\to\infty} f_{\sigma|_k}(x).$$

It is well known (Atkins et al., 2010; Hutchinson, 1981) that the limit above exists and is independent of $x \in \mathbb{X}$. Moreover π is continuous, onto, and satisfies the following commuting diagram for each $n = 1, 2, \ldots, N$.

$$\begin{array}{ccc} \Omega & \xrightarrow{s_n} & \Omega \\ \pi\downarrow & & \downarrow\pi \\ \mathbb{X} & \xrightarrow[f_n]{} & \mathbb{X} \end{array} \tag{7}$$

The symbol $s_n : \Omega \to \Omega$ denotes the inverse shift map defined by $s_n(\sigma) = n\sigma$. A *section* of the coding map π is a function $\tau : \Omega \to A$ such that $\pi \circ \tau$ is the identity. A section selects, for each $x \in A$, a single *address* in Ω from the ones that come from the coding map. Call the set $\Omega_\tau := \tau(A)$ the *address space* of the section τ. Let S denote the shift operator on Ω, i.e., $S(n\sigma) = \sigma$ for any $n \in \{1, 2, \ldots, N\}$ and any $\sigma \in \Omega$. A subset $W \subset \Omega$ will be called *shift invariant* if $\sigma \in W$ implies that $S(\sigma) \in W$. If Ω_τ is shift invariant, then τ is called a *shift invariant section*. The following example demonstrates the naturalness of shift invariant sections.

Example 2: Consider the IFS $\mathcal{F} = (\mathbb{R}; f_0, f_1)$ where $f_0(x) = \frac{1}{2}x$ and $f_1(x) = \frac{1}{2}x + \frac{1}{2}$. The attractor is the interval [0, 1]. An address of a point x is a binary representation of x. In choosing a section τ one must decide, for example, whether to take $\tau\left(\frac{1}{4}\right) = .01$ or $\tau\left(\frac{1}{4}\right) = .00111\cdots$. If the section τ is shift invariant, this would imply, for example, that if $\tau\left(\frac{1}{4}\right) = .00111\cdots$. then $\tau\left(\frac{1}{2}\right) = .0111\cdots$, not $\tau\left(\frac{1}{2}\right) = .100\cdots$. Call an IFS injective if each function of the IFS is injective.

Lemma 3: With notation as above, a section τ of an injective IFS is shift invariant if and only if, for any $x \in A$, if $\tau(x)|_1 = n$, then $(S \circ \tau)(x) = (\tau \circ f_n^{-1})(x)$.

Proof: Given the right hand statement above, we will prove that τ is shift invariant. Assume that $n\sigma \in \Omega_\tau$. Then there is an $x \in A$ such that $\tau(x) = n\sigma$, and hence $\sigma = (S \circ \tau)(x) = (\tau \circ f_n^{-1})(x) = \tau(f_n^{-1})(x))$. Thus $\sigma \in \Omega_\tau$.

Conversely, assume that τ is shift invariant. Assume that $x \in A$ and $\tau(x) = n\sigma$ for some $\sigma \in \Omega$. By shift invariance, there is a $y \in A$ such that $\tau(y) = \sigma$. Now

$$\begin{aligned} x &= (\pi \circ \tau)(x) = \pi(n\sigma) = f_n\left(\lim_{k\to\infty} f_{\sigma|_k}(A)\right) \\ &= f_n(\pi(\tau(y))) = f_n(y). \end{aligned}$$

Therefore $y = f_n^{-1}(x)$ and

$$(S \circ \tau)(x) = S(n\sigma) = \sigma = \tau(y) = \tau\left(r_n^{-1}(x)\right) = \left(\tau \circ f_n^{-1}\right)(x).$$

□

Theorem 4 below states that every shift invariant section of an injective IFS can be obtained from a mask. For an IFS $\mathcal{F}$ with attractor A, a *mask* is a partition $M = \{M_i, 1 \le i \le N\}$ of A such that $M_i \subseteq f_i(A)$ for all $f_i \in \mathcal{F}$. Given an injective IFS $\mathcal{F}$ and a mask M, consider the function $T : A \to A$ defined by $T(x) := f_i^{-1}(x)$ when $x \in M_i$. The *itinerary* $\tau_M(x)$ of a point $x \in A$ is the string $i_0\, i_1\, i_2 \cdots \in \Omega$, where i_k is the unique integer $1 \le i_k \le N$ such that

$$T^k(x) \in M_{i_k}.$$

Theorem 4: Let $\mathcal{F}$ be a contractive and injective IFS.

1 if M is a mask, then τ_M is a shift invariant section of π

2 if τ is a shift invariant section of π, then $\tau = \tau_M$ for some mask M.

Proof: To show that τ_M is a section, let $x \in A$. Then $(\pi \circ \tau_M)(x) = \lim_{k\to\infty} f_{\tau_M(x)|_k}(A)$. It follows immediately from the definition of τ_M that $x \in f_{\tau_M(x)|_k}(M_{i_{k+1}}) \subseteq f_{\tau_M(x)|_k}(A)$ for all k. Hence $(\pi \circ \tau_M)(x) = x$. Concerning the shift invariance, it follows from the definition of T that the following diagram commutes.

$$\begin{array}{ccc} A & \overset{T}{\to} & A \\ \tau_M \downarrow & & \downarrow \tau_M \\ \Omega & \underset{S}{\to} & \Omega \end{array} \tag{8}$$

If $\sigma \in \Omega_{\tau_M}$, then there is an $x \in A$ such that $\sigma = \tau_M(x)$ and, from the diagram, $(S \circ \tau_M)(x) = (\tau_M \circ T)(x) = \tau_M(f_n^{-1}(x)) \in \Omega_{\tau_M}$ for some $n \in \{1, 2, \ldots, N\}$.

Concerning the second statement, define a mask $M = \{M_i, 1 \le i \le N\}$ as follows:

$$M_i = \{x : \tau(x) = i\,\sigma \text{ for some } \sigma \in \Omega\}.$$

It is sufficient to show that $M_i \subseteq f_i(A)$, and that $\tau_M(x) = \tau(x)$ for all $x \in A$. If $x \in M_i$, then $\tau(x) = i\sigma$ for some $\sigma \in \Omega$ and $x = (\pi \circ \tau)(x) = f_i(\lim_{k\to\infty} f_{\sigma|_k}(A)) \in f_i(A)$.

To show that $\tau_M(x) = \tau(x)$, let $\tau(x) = j_0\, j_1\, j_2 \cdots$ and $\tau_M(x) = k_0\, k_1\, k_2 \cdots$. That $j_0 = k_0$ follows from the definitions. By induction, assume that $j_i = k_i$, for $i = 0, 1, \ldots, m-1$. Applying Lemma 3 for m times yields

$$j_m = \left(S^m \circ \tau\right)(x)\Big|_1 = \tau\left(f_{j_{m-1}}^{-1} \circ \cdots \circ f_{j_1}^{-1} \circ f_{j_0}^{-1}(x)\right)\Big|_1,$$

where the j_i's are determined by the recursive formula $\tau\left(f_{j_{r-1}}^{-1} \circ \cdots \circ f_{j_1}^{-1} \circ f_{j_0}^{-1}(x)\right)|_1 = j_r$. By the definition of the mask, $\tau(f_{j_{m-1}}^{-1} \circ \cdots \circ f_{j_1}^{-1} \circ f_{k_0}^{-1}(x)) \in M_{j_m}$. But by the definition of the itinerary, the k_i's are determined by the recursive formula

$$\tau\left(f_{k_{r-1}}^{-1} \circ \cdots \circ f_{k_1}^{-1} \circ f_{k_0}^{-1}(x)\right) \in M_{k_r}$$

for $r = 0, 1, 2, \ldots, m$. But, since $k_i = j_i$ for $i = 0, 1, \ldots, m-1$, we have $\tau(f_{j_{m-1}}^{-1} \circ \cdots \circ f_{j_1}^{-1} \circ f_{j_0}^{-1}(x)) \in M_{k_m}$. Therefore $k_m = j_m$. □

To define fractal homeomorphism, consider two contractive IFSs $\mathcal{F} = (\mathbb{X}, f_1, f_2, ..., f_N)$ and $\mathcal{G} = (\mathbb{X}, g_1, g_2, ..., g_N)$ with the same number N of functions on a complete metric space $\mathbb{X}$. Let A_F and A_G be the attractors and π_F and π_G the coding maps of $\mathcal{F}$ and $\mathcal{G}$, respectively. A homeomorphism $h : A_F \to A_G$ is called a *fractal homeomorphism* if there exist shift invariant sections τ_F and τ_G such that the following diagram commutes:

$$\begin{array}{ccc} A_F & \underset{h}{\to} & A_G \\ \tau_F \searrow & & \swarrow \tau_G \\ & \Omega & \end{array} \tag{9}$$

i.e., the homeomorphism h takes each point $x \in A_F$ with address $\sigma = \tau_F(x)$ to the point $y \in A_G$ with the same address $\sigma = \tau_G(y)$. Theorem 5 below states that the fractal homeomorphisms between attractors A_F and A_G are exactly mappings of the form $\pi_G \circ \tau_F$ or $\pi_F \circ \tau_G$ for some shift invariant sections τ_F, τ_G.

Theorem 5: Let $\mathcal{F}$ and $\mathcal{G}$ be contractive IFSs. With notation as above:

1 If $h : A_F \to A_G$ is a fractal homeomorphism with corresponding sections τ_F and τ_G, then $\Omega_{\tau_F} = \Omega_{\tau_G}$. Moreover $h = \pi_G \circ \tau_F$ and $h^{-1} = \pi_F \circ \tau_G$.

2 If τ_F is a shift invariant section for $\mathcal{F}$ and $h := \pi_G \circ \tau_F$ is a homeomorphism, then h is a fractal homeomorphism.

Proof: Concerning statement 1, since h is a bijection, the commuting diagram 9 implies that the images of τ_F and τ_G are equal, i.e., $\Omega_{\tau_F} = \Omega_{\tau_G}$. Now $\tau_F = \tau_G \circ h$ from the diagram implies $\pi_G \circ \tau_F = (\pi_G \circ \tau_G) \circ h = h$. The formula involving h^{-1} is likewise proved.

Concerning statement 2, the section τ_F is a bijection from A_F onto Ω_{τ_F}. Since h is also a bijection, the equality $h = \pi_G \circ \tau_F$ implies that $\pi_G|_{\Omega_F}$, the restriction of π_G to Ω_F, is a bijection onto A_G. If τ_G is the inverse of $\pi_G|_{\Omega_F}$, then τ_F and τ_G satisfy the commuting diagram 9. That τ_F is shift invariant means that $\Omega_{\tau_G} = \Omega_{\tau_F}$, i.e., τ_G is shift invariant. □

5 Image from a fractal homeomorphism

This section concerns images on the unit square □. Define an image as a function $c : \square \to \mathcal{C}$, where $\mathcal{C}$ denotes the colour palate, for example $\mathcal{C} = \{0, 1, 2, \ldots, 255\}^3$. If h is any homeomorphism from □ onto □, define the *transformed image* $h(c) : \square \to \mathcal{C}$ by

$$h(c) := c \circ h.$$

We are interested in the case where h is a fractal homeomorphism. The remainder of this section concerns fractal homeomorphism based on bi-affine IFSs with four functions as described in Example 1.

Consider Example 1 depicted in Figure 9. For the bi-affine IFS $\mathcal{F} = \{\square; f_1, f_2, f_3, f_4\}$, we will construct a section τ_F that is referred to in Barnsley (2009a) as the *top section*. Consider the mask $M_F = \{M_1, M_2, M_3, M_4\}$ defined recursively by

$$M_i = f_i(\square) \setminus \bigcup_{j=1}^{i-1} f_j(\square)$$

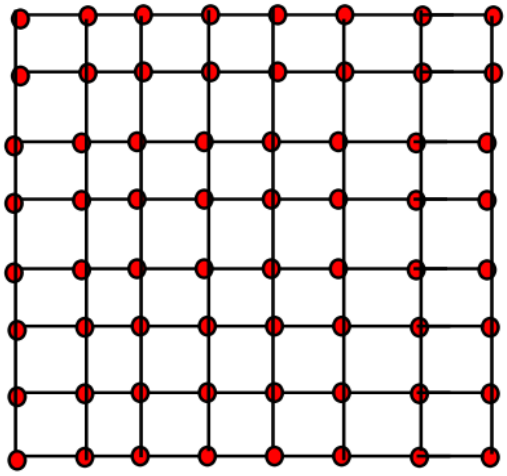
for i = 1, 2, 3, 4. Explicitly, M_1 is the closed quadrilateral $ATOQ$, M_2 is the open quadrilateral $OQBR$ together with the segments $(Q, B]$, $[B, R]$, $[R, O)$, M_3 is the open quadrilateral $ORCS$ together with the segments $(R, C]$, $[C, S]$, $[S, O)$, and M4 is the open quadrilateral $OSDT$ together with the segments $(S, D]$, $[D, T)$. The section τ_F corresponding to the mask M_F is given by $\tau(x) = \max \pi^{-1}(x)$, where the maximum is with respect to the lexicographic order on Ω.

Now consider a second bi-affine IFS $\mathcal{G} = \{\square; g_1, g_2, g_3, g_4\}$ of the same type with points O', Q', R', S', T' replacing O, Q, R, S, T, and with mask M_G defined exactly as it was for M_F. The masks M_F and M_G induce shift invariant sections τ_F and τ_G, respectively, as verified by Theorem 4. Theorem 6 below states that $\pi_G \circ \tau_F$ and $\pi_F \circ \tau_G$ are continuous and hence, by Theorem 5, fractal homeomorphisms.

To prove Theorem 6, the following lemma will be used. The proof is routine and will be omitted. All partitions $\mathcal{P}$ will be of the unit square $\square$ into regions whose closures are topological polygons. The *dual graph* of such a partition is the graph $\Gamma_{\mathcal{P}}$ whose points are the regions and where two vertices are joined if and only if the corresponding regions share a side. A partition $\mathcal{P}$ is *nested* in partition $\mathcal{Q}$ if each region in $\mathcal{P}$ is contained in some region of $\mathcal{Q}$. Assume that partition $\mathcal{P}_1$ is nested in partition $\mathcal{P}_2$ and $\mathcal{Q}_1$ is nested in $\mathcal{Q}_2$, and that there are graph isomorphisms $\Phi_1 : \Gamma_{\mathcal{P}_1} \to \Gamma_{\mathcal{Q}_1}$ and $\Phi_2 : \Gamma_{\mathcal{P}_2} \to \Gamma_{\mathcal{Q}_2}$. Call Φ_1 and Φ_2 *compatible* if whenever $P_1 \in \mathcal{P}_1$ and $P_2 \in \mathcal{P}_3$ with $P_1 \subseteq P_2$ we have $\Phi_1(P_1) \subset \Phi_2(P_2)$. The *mesh* $|\mathcal{P}|$ of a partition $\mathcal{P}$ is the maximum diameter of the regions. If $\lim_{n\to\infty} |\mathcal{P}_n| = 0$, then, for any $x \in \square$, there is a unique nested sequence $\{P_n\}$ of regions $P_n \in \mathcal{P}_n$ such that $x = \bigcap_{n\in\mathbb{N}} P_n$.

Lemma 4: Let $\mathcal{P}_n$ and $\mathcal{Q}_n$, n = 0, 1, 2, . . ., be two nested sequences of partitions of the unit square $\square$ with $\lim_{n\to\infty} |\mathcal{P}_n| = \lim_{n\to\infty} |\mathcal{Q}_n| = 0$. Assume that there are corresponding sequences of compatible graph isomorphisms $\Phi_n : G_{\mathcal{P}_n} \to G_{\mathcal{Q}_n}$. The map $h : \square \to \square$ defined as follows is a homeomorphism. For $x = \bigcap_{n\in\mathbb{N}} P_n \in \square$ with $P_n \in \mathcal{P}_n$, and define $h(x) = \bigcap_{n\in\mathbb{N}} \Phi_n(P_n)$.

Figure 10 The dual graph of the partition $\mathcal{P}_F^2$ (see online version for colours)



Theorem 6: For the two bi-affine IFSs $\mathcal{F} = \{\square, f_1, f_2, f_3, f_4\}$ and $\mathcal{G} = \{\square, g_1, g_2, g_3, g_4\}$ defined above, the map $h = \pi_G \circ \tau_F$ is a homeomorphism.

Proof: For each $n \geq 0$, let Ω_n denote the set of strings of length n using symbols $\{1, 2, 3, 4\}$. For the IFS $\mathcal{F}$, define a partition $\mathcal{P}_F^n = \{P_\sigma : \sigma \in \Omega_n\}$ of $\square$ recursively by taking $\mathcal{P}_F^0 = \mathcal{P}_F$ and

$$\mathcal{P}_F^{n+1} = \left\{P_{\sigma j} = P_\sigma \bigcap f_\sigma\left(P_j\right) : \sigma \in \Omega_n,\ \ 1 \leq j \leq 4\right\}.$$

A straightforward induction shows that $\{M_F^n\}$ is a nested sequence of partitions of $\square$. The dual graph Γ_F^n of $\mathcal{P}_F^n$ is the grid graph shown in Figure 10 for $n = 2$. This construction of a nested partition can be repeated for the IFS $\mathcal{G}$. Since the obvious graph isomorphisms between Γ_F^n and Γ_G^n are compatible with the nested partitions, Lemma 4 implies that h is a homeomorphism. $\square$

References

Atkins, R., Barnsley, M., Wilson, D.C. and Vince, A. (2010) 'A characterization of point-fibred affine iterated function systems', *Topology Proceedings*, Vol. 38, pp.189–211.

Barnsley, M. (2009a) 'Transformations between self-referential sets', *American Mathematical Monthly*, Vol. 116, No. 4, pp.291–304.

Barnsley, M. (2009b) 'Transformations between fractals', *Fractal Geometry and Stochastics IV*, pp.227–250.

Hutchinson, J. (1981) 'Fractals and self-similarity', *Indiana University Mathematics Journal*, Vol. 30, No. 5, pp.713–747.