Analysis of Subdivision Curves

Analysis is the process of proving properties of the limit curves in terms of properties of the mask and arity of a given scheme.

The relationships identified by analysis are key to designing new schemes with desired limit curve properties.

Limitations

- We deal only with **stationary** subdivision schemes, in which the coefficients, by which new vertices are computed as linear combinations of the old, are the same all the way along the control polygon and the same at every subdivision step.
- We deal with symmetric schemes, so that a given vertex has the same effect on both of its sides.
- We ignore end conditions, considering only the properties relevant to the interior of a long polygon.

Scenario

We express one step of the subdivision as a formal matrix product

$$P' = S.P$$

where P is the original vector of vertices (or polygon), S is the matrix implementing the refinement, and P' is the new polygon after one step.

The matrix S has columns which are identical to each other (by our first limitation), offset vertically each from its neighbour by a number of rows called the **arity**.

The column contains **coefficients** which are referred to collectively as the **mask**. From our second limitation,the mask is palindromic.

The height of the non-zero portion of the column is termed the **width** of the scheme.

The rows of the matrix are called the **stencils** or **templates**. There are *arity* different distinct templates, although these are strongly related by symmetry.

Repetition of the subdivision step is equivalent to raising S to a power, and taking the limit of repeated subdivision is equivalent to raising it to an infinite power.

The matrix is formal, because of our third limitation.

The matrix (and thus the scheme and its limit curve properties) are fully defined by the mask and the arity, and the process of designing a scheme is essentially that of choosing values for the arity and the width and then the coefficients within the mask of that width.

Basis functions

The first result we need to establish is that even though we do not start out from explicit basis functions, the equivalent still exists.

Each column of S defines how a given old vertex contributes to the set of new vertices. When we repeat the process, we effectively raise S to a high power, n. Each column of S^n defines how a given original vertex contributes to the (very dense) set of vertices at level n. In the limit each column of S^∞ defines how a given original vertex contributes to the limit curve.

Equivalently, we may view the column of S^∞ as the derivative of the limit curve with respect to an original vertex. This is exactly the kind of role which the basis functions fulfil in the cases where we start with equations.

In the regular situation, all the columns of S^n are identical, merely being shifted, and so we can talk about **the** basis function, all functions of the basis being shifted copies of the same shape.

Support

Given: a scheme of width w and arity a

To Construct: the support of the basis function.

Construction:

$$s = \frac{w-1}{a-1}$$

Proof: Start with cardinal data.

After the first step the effect of the non-zero vertex extends a distance $(w-1)/2a$ on each side, that being the distance to the last non-zero coefficient in the mask.

During the second step that last non-zero coefficient itself causes a further effect of $(w-1)/2a^2$, and in the third the region influenced grows by a further $(w-1)/2a^3$.

These contributions form a geometric series which can easily be summed to find the total effect of all the steps:

$$\begin{aligned}
\frac{s}{2} &= \frac{w-1}{2a}\left(1 + \frac{1}{a} + \frac{1}{a^2} + \dots\right) \\
s &= \frac{w-1}{a} \Big/ \left(1 - \frac{1}{a}\right) \\
&= \frac{w-1}{a} \Big/ \left(\frac{a-1}{a}\right) \\
&= \frac{w-1}{a-1}
\end{aligned}$$

■

From this we may determine corollaries:-

- that provided $a > 1$ the support of any finite scheme is itself finite.
- that the interpolating splines (which have unbounded support) cannot be reproduced by a finite subdivision scheme.
- the limit curve consists of pieces, each of which is influenced only by a particular subsequence of consecutive original vertices.

Positivity

If the mask contains only non-negative coefficients, the basis function will be non-negative.

The proof is by observing that if a matrix has only non-negative entries, no power can ever acquire a negative one.

An important corollary is that

- if a scheme is non-negative and is coordinate-system independent, the limit curve lies inside the convex hull of the vertices.

This is proved by taking a coordinate system aligned with one edge of that convex hull (in three dimensions take a skew coordinate system with two adjacent faces of the convex hull as two of the coordinate planes). If the basis functions are everywhere positive, the limit curve must lie on the positive side of the edge (of the two coordinate planes).

A sharper version of this uses the fact that each piece of curve is influenced only by a finite number of vertices to observe that the limit curve consists of pieces each within the convex hull of a finite subsequence of the vertices.

Containment

It is possible to provide some containment information even when the scheme is not positive semi-definite.

If all the coefficients in the mask lie between $-\alpha$ and $1+\alpha$, we can identify that all the vertices after one iteration lie inside a scaled version of the bounding box[1], scaled up by a factor of $1+2\alpha$.

For specific schemes it is possible to go further and prove bounds on the limit curve. The proofs depend on details of the scheme being considered. I have not yet found a way of doing this generically for all possible schemes subject to some conditions. It is obviously related to the convergence question addressed below.

Zero-crossings

The 4-pt binary scheme ($[-1, 9, 9, -1]/16$) has the property that the most extreme coefficients are negative. Thus given a cardinal initial polygon, the most extreme vertices at step 1 are also negative. The most extreme vertices after two steps are influenced only by the extreme ones at step 1, and so after two iterations from a cardinal polygon the most extreme vertices become positive again, and the sign of the extremes switches at every step.

This scheme is interpolating, and so all of these vertices lie on the basis function, which must (if continuous, and it is) have an infinite number of zero-crossings. No finite polynomial has an infinite number of roots, and so the basis function for the 4-pt binary scheme cannot be a polynomial (or a piecewise polynomial with a finite number of pieces).

[1] The corresponding result for the convex hull is a bit more complicated.

Precision Set

The precision set is technically the set of functions which are reproduced by the subdivision.

i.e. A function is in the precision set if, when all the vertices of the control polygon lie on that function, the limit curve is the same function.

This will clearly happen if, when all the vertices lie on the function, all the vertices after a single iteration lie on the same function. (Proof by induction.)

Only if seems likely, but I don't have a proof to hand.

The most important precision set results are those for low degree polynomials. These map directly into linear constraints on the coefficients, easily set up by choosing compatible values for P and P' and considering each stencil as defining a linear equation in the coefficients.

Constant Precision

This property holds if the constant function is in the precision set.

It maps to the constraint that the coefficients in each stencil sum to one.

This is also the constraint that the stencil should be a legitimate set of weights for computing a new point as a linear combination of old points, required in order to make the computation independent of the coordinate system being used. It is therefore a constraint which needs to be applied to all schemes. Schemes which do not satisfy it can be disregarded.

Linear Precision

This property holds if the general linear function is in the precision set.

Because the system is linear in the vertex values, and we are only interested in this property if constant precision also holds, we can set up constraints by choosing, for each stencil, a data set where the linear function (of unit slope) has a root at the corresponding new vertex.

We need to distinguish between **functional subdivision**, where the new vertices are deemed to be equally spaced in abscissa, and **geometric subdivision**, where all coordinates are equally determined by the subdivision process.

In the latter case, assuming constant precision, we have coordinate system independence, and so we may always choose a coordinate system in which the linear function has zero slope. All old vertices then have a zero ordinate, and so all new vertices given by linear combinations of zero will also have zero ordinate and lie on the same line.

Functional linear precision is thus the touchstone which allows us to use functional subdivision as a useful mental simplification of the geometric subdivision which we are really interested in. It is a constraint which should be applied, and schemes which do not satisfy it can be disregarded.

Quadratic Precision

This property holds if the general quadratic function is in the precision set.

Because the system is linear in the vertex values, and we are only interested in this property if constant and linear precisions also hold, we can set up constraints by choosing, for each stencil, a data set where the quadratic function (of unit second derivative) has a double root at the corresponding new vertex.

In general, only interpolating schemes can have this property, but there is a slightly relaxed version of great significance for non-interpolating schemes. This is the constraint that if all of the original vertices lie on a quadratic function, the new vertices after one step should lie on a quadratic with the same leading coefficients, but possibly displaced from it by an offset which is the same for all stencils.

This version of the constraint is significant for continuity of second derivative, because it is a necessary condition. If a scheme has continuity of second derivative, then after sufficiently many iterations it will be locally almost quadratic, and if quadratic precision in the relaxed sense did not hold, then the next iteration would not be locally quadratic.

Lower bound on width in terms of arity.

Each level of polynomial precision provides a number of constraints on the coefficients given by

$$\begin{array}{cc} \lfloor (a+1)/2 \rfloor & \text{for } w \text{ even} \\ \lfloor (a+2)/2 \rfloor & \text{for } w \text{ odd} \end{array}$$

The number of independent values within the mask is

$$\begin{array}{cc} w/2 & \text{for } w \text{ even} \\ (w+1)/2 & \text{for } w \text{ odd} \end{array}$$

which must be at least as large as the number of independent constraints.

The conditions arising from precision set arguments are useful in designing new schemes, because they are linear, and can be used to reduce the number of independent parameters of the scheme. Non-linear constraints are rather harder to handle.

Continuity and Convergence

The above has assumed that a limit curve exists. In fact it can fail to exist if the process of refinement diverges. It is also possible for the limit to be highly discontinuous, or to be undefined within some range.

Further, even if the curve itself is continuous, it is possible for the first derivative to diverge (infinite everywhere), to be discontinuous or to be undefined within some range.

In fact it seems typical for first derivatives to behave, and for higher derivatives to misbehave, but that is mainly because the really badly behaved curves tend to get dismissed early.

definitions

A function $f(x)$ is continuous if at every point x and for every value of $\epsilon > 0$ we can find a value of $\delta > 0$ such that

$$f(x+\delta) - f(x) < \epsilon$$

The derivative df/dx converges if the limit $(f(x+\delta)-f(x))/\delta$ remains below some upper bound as $\delta \to 0$.

Examples

The function $f(x) = x^{1/3}$ is continuous but its derivative does not converge at $x = 0$.

The function $f(x) = x^{4/3}$ has a continuous first derivative, but its second derivative does not converge at $x = 0$.

The function $f(x) = a + b\,sin(log(x))$ is undefined between the bounds $a - b$ and $a + b$ at $x = 0$.

The function $f(x) = a + bx\,sin(log(x))$ is continuous, but its first derivative is undefined between bounds at $x = 0$.

Necessary conditions

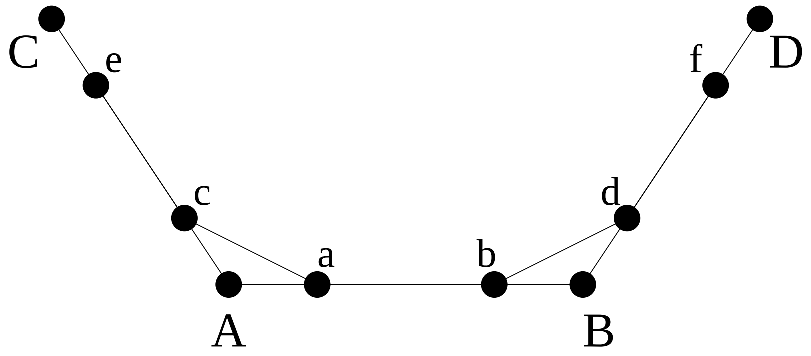
Our first tool for addressing these issues is the use of eigenanalysis on the matrix S. This is not surprising, as the limit curve is essentially given by the behaviour of S^n as $n \to \infty$.

We must be careful here, since on the face of it, S is infinite and is not square. However, we can create a matrix which is closely related to S for which we can safely carry out eigenanalysis. The key is that the neighbourhood of any one point of the curve is influenced only by a finite number of control points, and if we resequence the rows and columns of the matrix, so that the vertices near the point appear at the top and those further away further down and to the right, we find a matrix which is well behaved, with a finite set of eigenvalues.

A reasonable example is the matrix from the Chaikin (quadratic B-spline) scheme. The initial matrix S is

$$\frac{1}{4}\left|\begin{array}{cccccc} \ddots & & & & & \\ & 1 & 3 & & & \\ & & 3 & 1 & & \\ & & 1 & 3 & & \\ & & & 3 & 1 & \\ & & & 1 & 3 & \\ & & & & & \ddots \end{array}\right|$$

To turn this into something analyzable, we observe that this is a dual scheme, so that the midpoint of a span maps into the midpoint of a span. We call the points about which the configuration has mirror symmetry **mark points**. In this case the mark point of relevance is the midpoint of a span: in the case of a primal scheme the relevant mark point would have been a vertex. In the case of 'both' schemes, we would need to carry out the following analysis at both mark points, and in the case of a 'neither' scheme we would need to convert it to a both scheme by squaring.



Label the original vertices in sequence along the polygon as $[\ldots, E, C, A, B, D, F, \ldots]$ and the new vertices as $[\ldots, E', C', A', B', D', F', \ldots]$ and resequence the rows of S so that they occur in the sequence $A', B', C' \ldots$, and the columns of S so that they are in the sequence $A, B, C, \ldots$. Permutation of rows and columns is not an operation which preserves eigenproperties, but we can reasonably argue that the eigenproperties of this matrix are just as relevant as those of the original.

$$\begin{bmatrix} A' \\ B' \\ C' \\ D' \\ E' \\ F' \\ \vdots \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 3 & 1 & 0 & 0 & 0 & 0 & \\ 1 & 3 & 0 & 0 & 0 & 0 & \\ 3 & 0 & 1 & 0 & 0 & 0 & \\ 0 & 3 & 0 & 1 & 0 & 0 & \\ 1 & 0 & 3 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 3 & 0 & 0 & \\ & & & & & & \ddots \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \\ F \\ \vdots \end{bmatrix}$$

Note that after the first 4x4 top left component, there are no further diagonal or super-diagonal terms. Thus as the new S is raised to a high power, it reduces to a matrix 4 wide, with all other columns vanishing.

This 4x4 has eigenvalues 4,2,1,1, giving eigenvalues for the complete matrix of 1, 1/2, 1/4 and 1/4.

The first eigenvalue of unity was predictable because every row summed to 1. Its eigenvector on the right is [1/2,1/2,0,0] (scaled to make the sum unity because we are going to multiply it into a vector of points, and the result is supposed to be geometrically meaningful), and that on the left is $[1, 1, 1, 1, 1, 1, 1, 1, \ldots]^T$

This eigencomponent dominates, as all other eigenvalues are strictly smaller, and it predicts that as refinement proceeds, all points will converge to $[A + B]/2$

We can remove the effect of this component by taking $[A + B]/2$ as our origin. The next component has eigenvalue 1/2, with an eigenvector on the right of [1,-1,0,0] (this sums to zero, and therefore gives a displacement, and does not need scaling), and on the left of $[1/2, -1/2, 3/2, -3/2, 5/2, -5/2, \ldots]^T$, which indicates that the vertices converge to the straight line along $[A - B]$ and are evenly spaced thereon. The spacing halves at each iteration.

We are left with two equal remaining eigenvalues of $1/4$, and we can consider their effect by taking coordinates perpendicular to $[A-B]$.

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 1 & -1 \\ 3 & 3 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} -1/2 & -1/2 & 1/2 & 1/2 \\ 3/2 & -3/2 & 1/2 & -1/2 \end{bmatrix}$$

One of these components is even (symmetric about the vertical axis) and the other odd (antisymmetric). The even component has continuity of curvature, the odd one a discontinuity of curvature.

Because the two eigenvalues are equal we could have taken any pair of eigenvectors within the space that they span, but the odd/even distinction does actually have a lot of significance, as we can see if we apply a further transformation to our matrix S, so that it applies to an even/odd factorisation of the original configuration.

$$\begin{bmatrix} a+b \\ a-b \\ c+d \\ c-d \\ e+f \\ e-f \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \\ 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} A+B \\ A-B \\ C+D \\ C-D \\ E+F \\ E-F \end{bmatrix}$$

where because of the zeroes on the black squares of the chessboard, we can factorise the whole system into two independent parts.

$$\begin{bmatrix} a+b \\ c+d \\ e+f \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 4 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} A+B \\ C+D \\ E+F \end{bmatrix} \qquad \begin{bmatrix} a-b \\ c-d \\ e-f \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 2 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} A-B \\ C-D \\ E-F \end{bmatrix}$$

This partitioning always happens, because of the symmetry of the mask.

For schemes with constant precision, there is always a unit eigenvalue emerging from the even half.

For schemes with linear precision, there is always an eigenvalue of one half emerging from the odd half.

If a scheme is to have continuity of first derivative, the remaining even eigenvalues must all have magnitudes less than one half. If the next even eigenvalue is $1/4$ ($= (1/2)^2$) we have bounded second derivative. If it is less than $1/4$, the second derivative has a limit of zero.

If the next odd eigenvalue also is less than $1/4$, the second derivative is continuous, and it may be also be continuous if the curvature associated with the eigenvector on the left is zero at the mark point.

Sufficient conditions

The above methods allow us to determine the convergence and continuity of a scheme at its mark points. We can also observe that at every iteration new mark points arrive on the scene, and they become dense. In the case of binary subdivisions, this analysis tells us the continuity and convergence properties at all dyadic points ($a/2^b,\ a, b$ integers).

However, this is not enough.

If we take the variant of Chaikin with the mask $[1, 4, 4, 1]/5$ and apply this analysis, we find that the 2nd derivative is zero at every dyadic point. But if you draw the curve you find it can still curl round corners !

Sufficient conditions for convergence and continuity have been worked out for many schemes, by considering the polygon joining the control vertices as a fully continuous curve,h and looking at the sizes of the differences between one polygon and the next.

As far as I can see, these analyses have been specific to individual schemes, and I have not yet been able to articulate a proof method which can be applied to an arbitrary scheme to identify its properties.

However, one important technique common within this set of approaches is the use of **generating functions**. We have met these already as a way of computing the coefficients of a Box-spline subdivision. They are more than that.

The particular result which needs recording here is that if we are given a scheme as a mask, we can (sometimes) construct a mask which gives a subdivision scheme for the first differences, which eventually converges (if it converges) to the derivative.

We use **Laurent Polynomials** (polynomials which can have negative powers as well as positive ones) and associate with the sequence $[s_i, i = 1, , n]$ the polynomial[1] $s(z) = \sum_i s_i z^i$.

The transform of the convolution of two such sequences is given by the product of the generating functions.

Shifting by one step corresponds to multiplying by z and addition or subtraction to addition or subtraction. Shifting by two steps corresponds to multiplying by z twice, i.e. by z^2.

We can now see that a binary refinement scheme can be articulated as the equation

$$P'(z) = S(z)P(z^2)$$

The first differences of $P(z^2)$ are $(z^2 - 1)P(z^2)$, and those of $P'(z)$ are $(z - 1)P'(z)$

$$\begin{aligned}\delta P'(z) =& (z-1)P'(z) \\ =& (z-1)S(z)P(z^2) \\ =& \frac{(z-1)S(z)}{(z^2-1)}\delta P'(z)\end{aligned}$$

[1] This technology is also known as **z-transforms**, and it was developed as an extension of Laplace transforms as a way of analysing the stability of digital control systems where time delays are normal.

Thus $S'(z) = \frac{(z-1)}{(z^2-1)}S(z) = S(z)/(z+1)$ is a scheme for getting differences from differences. If we have a method of determining whether a scheme is convergent or not, we can use it on these derived schemes to determine if the differences are uniformly convergent (i.e. the limit function is continuous).

For a general arity, a, the derivation becomes

$$S'(z) = \frac{(z-1)}{(z^a-1)}S(z).$$

To get the property of the derivative we have to multiply $S'(z)$ by a. This derived scheme tells us whether the derivative is uniformly convergent.

Artifacts

Even if we have a smooth limit function, the limit curves may not be what we really want in a curve design system. For all the stationary schemes so far studied, if the control polygon has points evenly spaced around a circle, the limit curve does not lie on that circle, but has slight bumps or ripples which betray the original frequency and phase of the polygon.

In the case of more complicated polygons, it is less easy to see, but we do require that our curves should behave well for simple polygons.

We can determine how large these artifacts are going to be by doing a Fourier series analysis on the refined polygon and comparing its frequency components with those of the original.

Shannon's sampling theorem tells us that any frequencies greater than half the original frequency of the polygon are not justified by the original data, and so must be artifacts, and indeed, this issue is just another instance of aliasing, familiar in graphics as the 'jaggies' which appear on low resolution screen when sloping lines are drawn.

The artifacts vary in size with the number of polygon vertices round the circle. If there are many, the effect is small, and so most of the damage to the curve shape is done at the first refinement.

The rate at which the wiggles reduce with sampling density depends on the width of the mask. Wide masks take into account a large part of the circle, and can do a better job of blending.

However, wide masks are also more expensive, and complicate surface implementations quite disproportionately. The cure for artifacts is therefore to be sought in non-stationary schemes, which can apply a small amount of non-linearity to very good effect.