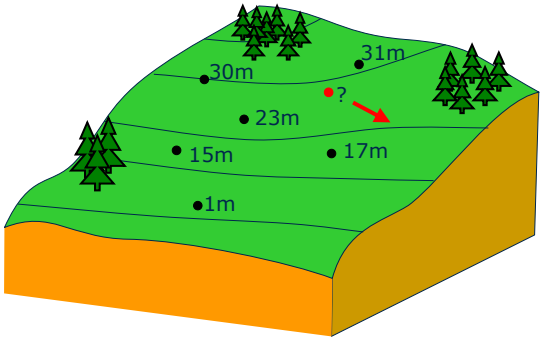
INTERPOLACIÓN

La interpolación es un proceso por el cual se define un valor en un punto cualquiera a partir de los valores conocidos en algunos puntos dados. Por ejemplo: tenemos estaciones meteorológicas que nos dan la temperatura, la presión y el viento en distintas localidades y queremos estimar el clima en otro pueblo cualquiera. Suponiendo a las estaciones razonablemente cercanas, surgen dos preguntas: ¿Cuáles estaciones considero? ¿Cómo calculo los valores estimativos para el pueblo, a partir de los datos conocidos?



En la figura se esquematiza un terreno, del cual se conocen las alturas en distintos puntos. Se requiere interpolación para estimar la altura en un punto cualquiera o para definir las líneas de altura constante (curvas de nivel o isocurvas).

En general, tenemos un conjunto finito de datos (punto, valor), que suelen denominarse **nodos** y **valores nodales**: $\{P_i,v_i\}$ o $\{(\mathbf{x}_i, v_i)\}$, donde $P_i \equiv \mathbf{x}_i = \{x_i, y_i[, z_i]\}$ es la posición del i-ésimo punto y v_i el valor conocido en ese lugar. Se pretende estimar el valor v, en un punto variable de coordenadas **x** conocidas.

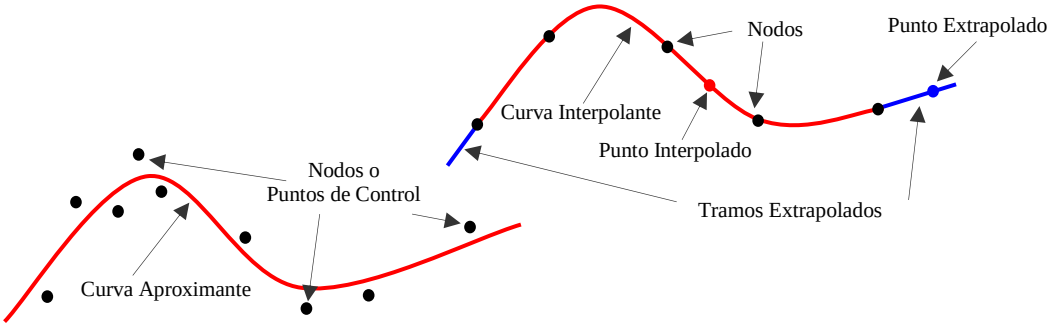
La interpolación define un valor aproximado para un punto cualquiera, en función de los valores conocidos en algunos puntos. Podría asignarse el valor del punto más cercano o una combinación de algunos valores cercanos o de todos los conocidos. En una dimensión, es muy sencillo encontrar el punto más cercano, o los dos nodos que "encierran" el punto, pero en más dimensiones resulta más complicado.

La forma estándar, para variables continuas, consiste en hacer un promedio ponderado:

$$v(\mathbf{x}) = (\sum \beta^i(\mathbf{x})\, v_i) / (\sum \beta^i(\mathbf{x})) = \sum [\beta^i(\mathbf{x}) / \sum \beta^i(\mathbf{x})]\, v_i = \sum \alpha^i(\mathbf{x})\, v_i \quad (\Sigma\, \alpha^i = 1)$$

El valor v, en la posición **x** del punto variable, se estima asignando a cada nodo un **peso** α^i, que depende de la posición del punto variable. Normalmente, el peso de cada nodo estará en relación inversa con la distancia al punto variable, de modo que el valor resulte más influenciado por los puntos más cercanos. Por ejemplo, puede asignarse peso unitario al nodo más cercano y cero al resto, con lo que el punto toma el valor del nodo más cercano; también se puede usar, como peso, la inversa de la distancia (con la suma de inversas como cociente normalizador), teniendo en cuenta todos los puntos, aún los muy lejanos, que prácticamente no tienen influencia. Parece natural buscar unos pocos nodos que influyan en el punto y medir, de alguna forma, esa influencia; los puntos que no influyan no se consideran, o equivalentemente: se les asigna peso nulo. Hay muchas formas aceptadas para definir los pesos; es justamente la selección de pesos la que define el método de interpolación.

No hay una única clasificación de métodos y denominaciones, pero debemos mencionar algunas:



a) Métodos Locales versus Globales: En un método global influyen todos los nodos. Un ejemplo es el método de Kriging, utilizado en las ciencias de la tierra, donde se construyen los pesos en relación inversa a las distancias, con algunos supuestos de variabilidad. Para los métodos locales, en cambio, sólo se consideran los nodos cercanos; a los más lejanos se les asigna peso nulo. En general se utilizan estos últimos métodos pues son algorítmicamente más veloces, sobre todo cuando se dispone de un método rápido para encontrar un conjunto de nodos cercanos.

b) Interpolación versus Extrapolación: En la interpolación se asume que el punto, cuyo valor se busca, esta "entremedio de los nodos", el caso contrario es extrapolación. El concepto de "entremedio" es trivial en 1D, pero en más dimensiones se suele hablar de ámbito, órbita, esfera o zona de influencia y definirlo es un poco más complejo (envoltorio convexo); en breve lo desarrollaremos.

c) Interpolación versus Aproximación: En la interpolación se pretende que el resultado de la fórmula sea valor del nodo cuando el punto variable coincide con el nodo; en una aproximación eso no es necesario. Parece extraño admitir una aproximación; pero, en general, es lo más razonable: imaginemos un proceso de medición de puntos de un objeto 3D; si hay muchísimos puntos medidos, naturalmente con cierto error, pretendemos que la superficie "aproximada" pase por entre los puntos; del mismo modo que hacíamos "regresión lineal" (que es la forma más conocida de aproximación) o un proceso de aproximación por mínimos cuadrados. Si la superficie fuera obligada a pasar por todos los nodos aparecería muy zigzagueante o arrugada. En otros casos, como el CAD, los pocos puntos aproximados se denominan puntos de control y sirven para definir una curva o superficie compleja.

   En síntesis: se denomina interpolación cuando el valor calculado en un nodo coincide con el valor nodal: $\forall j\in[1,n]$, $v(\mathbf{x}_j) = \sum_1^n \alpha^i(\mathbf{x}_j)\, v_i = v_j$ (o: $\alpha^i(\mathbf{x}_j) = \delta^i_j$); en caso contrario es una aproximación.

Interpolación afín

Haciendo una combinación afín de los puntos fijos, se obtiene un punto variable, que depende de los coeficientes utilizados. La interpolación afín utiliza los mismos coeficientes como "pesos" para interpolar valores en el punto calculado; es decir que los valores nodales se combinan del mismo modo que las coordenadas espaciales, con los mismos coeficientes.

Recordemos que la combinación afín de n+1 puntos, da un nuevo punto; que se obtiene a través de una combinación lineal de n vectores, formados por la diferencia de n puntos E_i, con un punto especial O:

$$(P - O) = \sum^n \alpha^i (E_i - O)$$

No se pueden sumar puntos ni multiplicar escalares por puntos, pero hagamos un trato: cuando digamos "punto" P estamos hablando de sus coordenadas, de su vector posición **x**, respecto de un sistema de referencia fijo, cuyo origen y base arbitrarios no tienen nada que ver con O ni los E_i. Así podemos hacer:

$$P = \sum^n \alpha^i E_i - \sum^n \alpha^i O + O = \sum^n \alpha^i E_i + (1 - \sum^n \alpha^i) O.$$

Siendo O el (n+1)-ésimo punto dato, consideremos a $(1-\sum^n \alpha^i)$ como el (n+1)–ésimo coeficiente:

$$E_{n+1} = O \qquad \alpha^{n+1} = 1 - \sum^n \alpha^i$$

Y así llegamos a la forma estándar (simétrica y muy cómoda) de expresar una combinación afín:

$$P = \sum^{n+1} \alpha^i E_i, \text{ con la condición: } \sum^{n+1} \alpha^i = 1$$

Esta es la forma que más utilizaremos en computación gráfica. Diremos que una combinación afín de puntos es un **promedio ponderado** o combinación lineal con coeficientes que suman uno.

Conocemos los nodos E_i, en posición $\mathbf{x}_i = \{x_i^j\}$ y con valor nodal v_i, las coordenadas $\mathbf{x} = \{x^j\}$ y el valor v, asociado a esa posición, se obtienen con el mismo promedio ponderado:

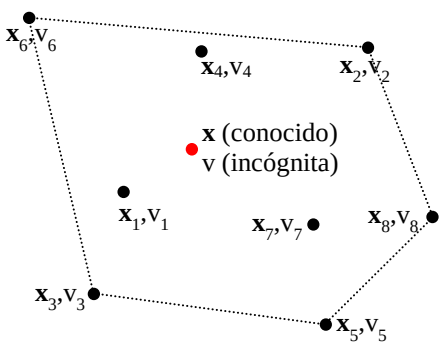
$$\mathbf{x} = \sum \alpha^i \mathbf{x}_i \qquad \Sigma \alpha^i = 1$$

y luego, con esos mismos pesos, se calcula:

$$v = \sum \alpha^i v_i.$$

En síntesis: **la interpolación afín consiste en calcular las variables del mismo modo que las coordenadas de una combinación afín**: $\{x,y,z,v\} = \sum \alpha^i \{x_i,y_i,z_i,v_i\}$

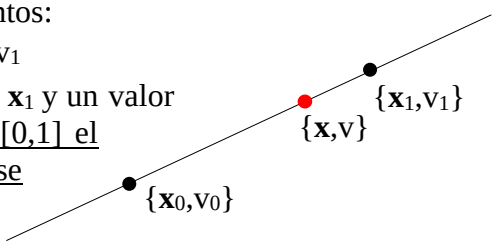
Aún no queda claro cómo se definen los pesos $\alpha^i(\mathbf{x})$, si viene dada la posición **x** del punto variable. Antes de responder eso, vamos a desarrollar detalladamente la combinación afín de varios puntos.



En la figura se representa un conjunto $\{\mathbf{x}_i,v_i\}$ de puntos fijos con valores nodales asignados y un punto variable en una posición conocida **x**, para el cual se pretende interpolar el valor v, asignando pesos a los puntos fijos. El sistema de referencia del plano no importa, es arbitrario. Se muestra también el **envoltorio convexo**, cápsula convexa o ***convex-hull*** (CH) del conjunto de puntos fijos. El envoltorio convexo se define como el menor convexo que contiene al conjunto de puntos; es la forma que tomaría una banda elástica envolviendo al conjunto. Enseguida veremos que ese es el límite de las posiciones que pueden obtenerse haciendo combinaciones afines de todos los puntos fijos, con pesos $\alpha^i \geq 0$ (no negativos).

El problema de interpolar un valor entre muchos puntos lo iremos resolviendo en forma progresiva. El caso más simple lo constituye la combinación afín 1D de dos puntos:

$\mathbf{x} = \alpha^0 \mathbf{x}_0 + \alpha^1 \mathbf{x}_1$, con $\alpha^0 + \alpha^1 = 1 \quad \Rightarrow \quad v = \alpha^0 v_0 + \alpha^1 v_1$



Definidos los α^i, se obtiene un punto en la línea que une $\mathbf{x}_0$ con $\mathbf{x}_1$ y un valor asociado al punto. Cuando los pesos están limitados al rango [0,1] el punto variable estará en el segmento que une los dos puntos y se garantiza que cualquier variable asociada asumirá también un valor intermedio. Si alguno de los parámetros sale de [0,1], el punto sale del segmento, hacemos extrapolación; el valor asociado también sale del intervalo $[v_0, v_1]$.

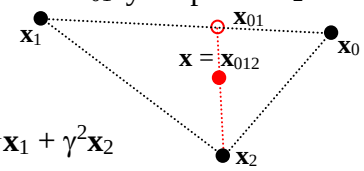
Para aclarar: reescribimos la ecuación anterior, pero asumiendo que ξ representa cualquier valore en un punto, ya sea una coordenada (x, y o z) o un valor asociado (temperatura, presión, color R, G o B, etc.):

$\xi = \alpha^0 \xi_0 + \alpha^1 \xi_1 = (1-\alpha^1) \xi_0 + \alpha^1 \xi_1 = \xi_0 + (\xi_1 - \xi_0) \alpha^1$.

La ecuación muestra que cuando $\alpha^1 \in [0,1]$, ξ adopta valores intermedios entre ξ_0 y ξ_1, para cualquier variable: coordenada o valor anexado; cada una puede ser vista como una componente o se puede ver a ξ como un vector de componentes diversas. El punto se mantiene en el envoltorio convexo de los dos puntos fijos, que es simplemente el segmento $[\mathbf{x}_0, \mathbf{x}_1]$ y cualquier valor se mantiene en $[v_0, v_1]$. La combinación, en este caso ($\alpha^i \in [0,1]$), se denomina interpolación en sentido estricto o bien **combinación convexa**, un término que quedará claro en más dimensiones. Si los valores de α^i salen del intervalo [0,1] las coordenadas y los valores asociados pasan más allá de los límites y estamos extrapolando.

Mediante una combinación afín de dos puntos fijos (interpolación o extrapolación) no pueden obtenerse puntos fuera de la recta que los une y por lo tanto la interpolación afín no está definida fuera de la recta. Por definición utiliza los mismos pesos que la combinación afín, entonces está restringida a la expansión afín de los dos nodos. No se puede estimar así el valor en un punto fuera de la recta que definen.

En la figura derecha está marcado $\mathbf{x}_{01}$ como una combinación afín entre los puntos $\mathbf{x}_0$ y $\mathbf{x}_1$. Ahora agregamos $\mathbf{x}_2$ fuera de la recta que forman y hacemos $\mathbf{x} = \mathbf{x}_{012}$ combinando $\mathbf{x}_{01}$ y el punto $\mathbf{x}_2$. La multiplicidad de índices (01, 012, …) es sólo para indicar la secuencia.



$\mathbf{x}_{01} = \alpha^0 \mathbf{x}_0 + \alpha^1 \mathbf{x}_1$; con $\alpha^0 + \alpha^1 = 1$;

$\mathbf{x}_{012} = \beta^{01} \mathbf{x}_{01} + \beta^2 \mathbf{x}_2$; con $\beta^{01} + \beta^2 = 1$;

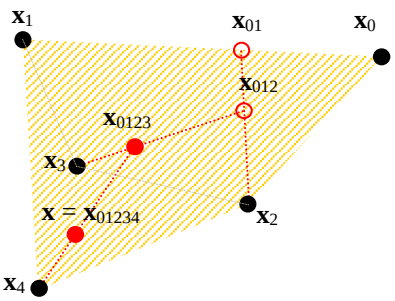
$\mathbf{x}_{012} = \beta^{01}(\alpha^0 \mathbf{x}_0 + \alpha^1 \mathbf{x}_1) + \beta^2 \mathbf{x}_2 = \beta^{01}\alpha^0 \mathbf{x}_0 + \beta^{01}\alpha^1 \mathbf{x}_1 + \beta^2 \mathbf{x}_2 = \gamma^0 \mathbf{x}_0 + \gamma^1 \mathbf{x}_1 + \gamma^2 \mathbf{x}_2$

También es una combinación afín, porque los parámetros suman uno:

$\gamma^0 + \gamma^1 + \gamma^2 = \beta^{01} \alpha^0 + \beta^{01} \alpha^1 + \beta^2 = \beta^{01} (\alpha^0 + \alpha^1) + \beta^2 = \beta^{01} + \beta^2 = 1$

Sin ser muy rigurosos, podemos dar por demostrado que la combinación afín de combinaciones afines es una combinación afín (suma uno) de los puntos fijos. Además, si los coeficientes están entre cero y uno en cada paso, los coeficientes finales (γ) también (son productos de números entre cero y uno), por lo tanto: la combinación convexa de combinaciones convexas es una combinación convexa. Un sencillo análisis de los límites de cada combinación (hacerlo) nos revela que en el caso convexo (pesos positivos) el resultado estará dentro del triángulo marcado que es el envoltorio convexo de los tres puntos fijos.

Añadiendo al esquema un cuarto punto, quinto, etc. todos en el mismo plano. llegaremos fácilmente a la conclusión de que la combinación afín de un conjunto de puntos, cuando los parámetros están limitados al rango [0,1], dará por resultado un punto entre los límites del envoltorio convexo del conjunto. Es por ello que este tipo de combinación recibe el nombre de combinación convexa. (afín: Σ 1, convexa: afín y [0,1])

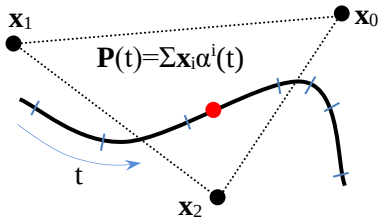


Obviamente, el resultado anterior es válido en cualquier cantidad de dimensiones y no solo en el plano. La extensión es trivial puesto que con $\mathbf{x}_i$ representamos cualquier coordenada o variable asociada. De hecho, podemos considerar x, y, z, temperatura y presión, como un punto en cinco dimensiones: $\xi = \{x,y,z,t,p\}$.

Se puede pensar en el dibujo anterior como la proyección plana de un resultado multidimensional; con lo cual queda claro que la proyección ortogonal de una combinación afín es combinación afín de los puntos proyectados y por lo tanto está dentro del envoltorio convexo de los puntos proyectados (la proyección ortogonal es una transformación afín y por lo tanto preserva las combinaciones afines).

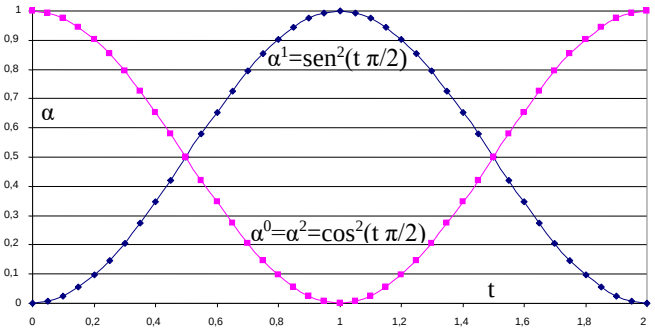
Un plano se define por combinación afín de tres puntos, una recta por dos. Pero como vimos, con más de tres puntos coplanares también se puede definir cualquier otro punto del plano, por interpolación o extrapolación afín. En general, con d+1 puntos independientes (puede haber más, pero dependientes) se obtiene cualquier punto del espacio de soporte d dimensional. La dependencia afín sólo implica que el mismo punto se puede obtener de distintas maneras (distintos α^i).

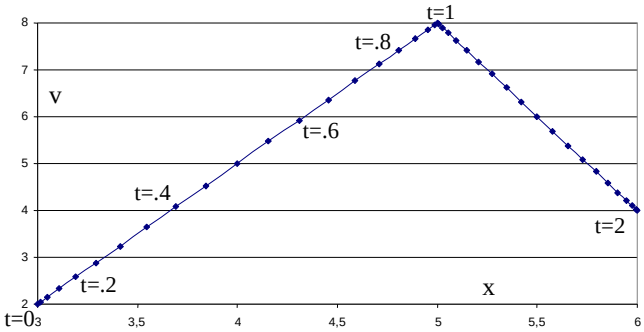
Si al conjunto de parámetros se le impone una restricción funcional se obtiene un subespacio. Por ejemplo, en dos o más dimensiones, si cada parámetro α^i es función continua de otro parámetro t: $\alpha^i = \alpha^i(t)$, el resultado es una curva uniparamétrica como en la figura. Si en tres o más dimensiones, $\alpha^i = \alpha^i(s,t)$ se obtiene una superficie biparamétrica. Este es el mecanismo que se utiliza para construir curvas y superficies paramétricas (Bézier, splines, NURBS, etc.) por combinación convexa.



Los pesos pueden definirse mediante cualquier función, lineal o no, continua o discontinua, mientras sumen uno. Por ejemplo, para dos puntos fijos $\mathbf{x}_0$ y $\mathbf{x}_1$ hacemos $\alpha^0=\text{sen}^2(t)$ y $\alpha^1=\cos^2(t)$, que suman siempre uno; si asumimos que t es el tiempo, tendremos un punto variable rebotando entre los nodos a velocidad variable. **La interpolación es afín si los pesos suman uno y convexa si no son negativos, pero se denomina "lineal" sólo cuando los pesos son funciones lineales de las coordenadas**.

Para ayudar a distinguir el efecto afín del lineal veamos un ejemplo no-lineal: Vamos a hacer la interpolación afín de valores en dos tramos, entre tres puntos en una dimensión: ($x_0 = 3$, $v_0 = 2$); ($x_1 = 5$, $v_1 = 8$); ($x_2 = 6$, $v_2 = 4$). Los pesos no-lineales dependerán de un parámetro t que varía entre 0 y 1 para el primer tramo y entre 1 y 2 para el segundo tramo: $\alpha^1=\text{sen}^2(t\,\pi/2)$ y $\alpha^0=\alpha^2=\cos^2(t\,\pi/2)$ cuyos valores se muestran en la gráfica izquierda. Se calculan **x** y v, que se grafican a la derecha.





La "velocidad" (dv/dt) es variable, pero el valor v asignado a cada punto cumple $(v-v_0)/(v_1-v_0) = (x-x_0)/(x_1-x_0) = \cos^2(t\,\pi/2)$, es decir que v(x) es una recta en el primer tramo y otra en el segundo. Esto es así, obviamente, pues ambos (**x**, v) reciben el mismo tratamiento. Lo que se pretende dejar claro es que no siempre **x** y v varían linealmente con un dado parámetro t.

Generalizando a más dimensiones: la interpolación afín será lineal cuando el valor dependa sólo de los datos en los nodos del "**símplice**" (símplice = segmento, triángulo, tetraedro y, en general, el convex-hull de *n*+1 puntos con independencia afín en *n* dimensiones) que rodea al punto y las funciones de peso de cada nodo valen uno en su nodo y cero en el resto y entremedio varían en forma lineal. Veremos el método estándar para construir funciones de peso que varían linealmente en el espacio, a diferencia del ejemplo anterior. La interpolación afín y lineal se suele llamar interpolación lineal (a secas).

Coordenadas Baricéntricas – Inversión de la interpolación lineal

Hasta aquí actuábamos como si conociéramos o asignáramos los pesos: En una interpolación afín, se conocen los nodos, sus valores y las funciones de peso para cada punto; entonces, dado un punto, se puede calcular el valor en esa ubicación. La inversión permite encontrar las funciones que definen los pesos en un cada punto. En este caso buscamos pesos que varíen en forma lineal.

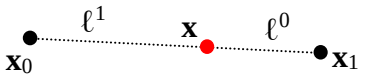
Para definir funciones de peso lineales, que recorren el espacio a velocidad constante en cualquier coordenada (derivada parcial constante), se utilizan las coordenadas baricéntricas o coordenadas de área o funciones de forma lineales, que son formas de denominar a los pesos lineales de un conjunto de nodos con independencia afín (símplice), por ejemplo: dos puntos distintos, tres puntos no colineales (no alineados) o cuatro no coplanares (que no están en el mismo plano).

Los pesos de una combinación afín se denominan "coordenadas" porque la ecuación $\mathbf{x} = \sum^d \alpha^i \mathbf{x}_i$ define el punto tomando los d nodos fijos y con independencia afín, como sistema de referencia de un espacio afín (d-1)-dimensional; los pesos α^i son las coordenadas de **x**. En particular, cuando todos los pesos valen 1/d, el punto obtenido es el baricentro (bari ~ peso) o centro de gravedad o promedio de los puntos.

Partamos del caso más sencillo, que es el de dos puntos fijos que definen un segmento [$\mathbf{x}_0$, $\mathbf{x}_1$]; hay una variable en particular que sabemos que varía en forma lineal entre los extremos: la longitud $\ell^1(\mathbf{x})$ del segmento [$\mathbf{x}_0$, $\mathbf{x}$]: si $\mathbf{x} = \mathbf{x}_0 \Rightarrow \ell^1(\mathbf{x}_0) = 0$ y, si $\mathbf{x} = \mathbf{x}_1 \Rightarrow \ell^1(\mathbf{x}_1) = \ell$, entonces se la puede interpolar:

$$\ell^1(\mathbf{x}) = \alpha^0(\mathbf{x})\ \ell^1(\mathbf{x}_0) + \alpha^1(\mathbf{x})\ \ell^1(\mathbf{x}_1), \text{ con } \alpha^0 + \alpha^1 = 1$$

$$\ell^1(\mathbf{x}) = \alpha^0(\mathbf{x})\ 0 + \alpha^1(\mathbf{x})\ \ell; \quad \Rightarrow \quad \alpha^1 = \ell^1/\ell$$



El peso de $\mathbf{x}_1$ es proporcional a la longitud ℓ^1 del segmento **opuesto** al punto. Del mismo modo, o utilizando la suma unitaria, se obtiene que $\alpha^0 = \ell^0/\ell$. Es decir que la relación (cociente) entre las longitudes de los segmentos opuestos es idéntica a la relación entre los pesos.

La relación se cumple también para la extrapolación, pero necesitaremos "longitudes con signo". Para definir el signo se debe establecer un **orden** de los puntos o una dirección preferencial, por ejemplo: de cero a uno o el vector $\boldsymbol{\ell} = (\mathbf{x}_1 - \mathbf{x}_0)$, con $\ell = |\boldsymbol{\ell}|$; y entonces, dado que no se puede dividir por un vector, recurrimos a proyectar numerador y denominador sobre el versor $\boldsymbol{\ell}/\ell$ que define la dirección positiva:

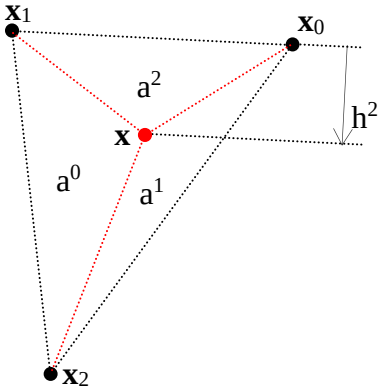
$$\ell^1 = (\mathbf{x}-\mathbf{x}_0)\cdot\boldsymbol{\ell}/\ell \qquad \ell^0 = (\mathbf{x}_1-\mathbf{x})\cdot\boldsymbol{\ell}/\ell$$

$$\alpha^1 = (\mathbf{x}-\mathbf{x}_0)\cdot\boldsymbol{\ell}/\ell^2 \qquad \alpha^0 = (\mathbf{x}_1-\mathbf{x})\cdot\boldsymbol{\ell}/\ell^2$$

En una extrapolación hay una longitud ℓ^i mayor que el total ℓ y una negativa, entonces habrá un peso α^i mayor que uno y el otro será negativo, pero los ℓ^i (con signo) suman el total ℓ y los α^i suman uno.

En un algoritmo para calcular los pesos se implementa la última ecuación, no se necesita invocar ningún condicional para averiguar si el punto móvil está entremedio de los nodos. Dado que $\ell^2 = \boldsymbol{\ell}\cdot\boldsymbol{\ell}$, no es necesario calcular la longitud del segmento que es la raíz cuadrada del denominador.

Pasamos ahora a tres puntos no alineados. A la derecha se pueden ver los triángulos definidos por el punto interpolado $\mathbf{x}$ con cada uno de los lados del triángulo que lo contiene, el área total es $a = a^0 + a^1 + a^2$, la denominación de cada uno corresponde al punto fijo **opuesto**.



El área a^2 (léase: "a dos" y no: "a al cuadrado") del triángulo opuesto a $\mathbf{x}_2$: ($\mathbf{x}$, $\mathbf{x}_0$, $\mathbf{x}_1$) varía linealmente; vale 0 cuando el punto variable está en la línea ($\mathbf{x}_0$, $\mathbf{x}_1$) (en particular $a^2(\mathbf{x}_0) = a^2(\mathbf{x}_1) = 0$) y es igual al área total a, cuando el punto variable está en una línea paralela a la anterior y que pasa por $\mathbf{x}_2$ (en particular $a^2(\mathbf{x}_2) = a$). Es lineal porque es un triángulo de base fija $|\mathbf{x}_1-\mathbf{x}_0|$ y altura h^2 (h dos) que varía en forma lineal (la gráfica de $h^2(x,y)$ es un plano). Interpolando para calcular el área a^2:

$$a^2 = \alpha^0\ a^2(\mathbf{x}_0) + \alpha^1\ a^2(\mathbf{x}_1) + \alpha^2\ a^2(\mathbf{x}_2) = \alpha^0\ 0 + \alpha^1\ 0 + \alpha^2\ a \quad \Rightarrow \quad \alpha^2 = a^2/a.$$

Del mismo modo, $\alpha^i = a^i/a$ para cualquier vértice en cualquier punto dentro del triángulo. Esta es la razón por la que también se denominan coordenadas de área. Para calcularlas utilizaremos el producto cruz, que es el área del paralelogramo que definen dos vectores; el área de un triángulo es la mitad, pero los cocientes entre áreas hacen innecesarias las divisiones por dos (se "tachan").

Para poder extrapolar por fuera del triángulo, se utilizan los productos vectoriales proyectados y con signo: Todos los vectores área se proyectan en una dirección elegida, y lo lógico es utilizar la dirección del vector área total. Siendo anti-conmutativos, hay que definir un orden que logre que, para un punto interior, todas las áreas apunten en una misma dirección. Definimos primero el área total:

$$\mathbf{a} = (\mathbf{x}_1-\mathbf{x}_0) \times (\mathbf{x}_2-\mathbf{x}_0)$$

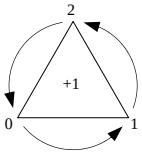
Ese orden elegido para $\mathbf{a}$, nos fuerza a definir las áreas parciales de modo que, cuando el punto está en el interior del triángulo, apunten al mismo lado del plano que el vector área total:

$$\mathbf{a}^0(\mathbf{x}) = (\mathbf{x}_1-\mathbf{x}) \times (\mathbf{x}_2-\mathbf{x}) \qquad \mathbf{a}^1(\mathbf{x}) = (\mathbf{x}_2-\mathbf{x}) \times (\mathbf{x}_0-\mathbf{x}) \qquad \mathbf{a}^2(\mathbf{x}) = (\mathbf{x}_0-\mathbf{x}) \times (\mathbf{x}_1-\mathbf{x})$$

El orden utilizado puede (suele) generalizarse como:

$$\mathbf{a}^i(\mathbf{x}) = (\mathbf{x}_{i+1}-\mathbf{x}) \times (\mathbf{x}_{i+2}-\mathbf{x})$$

donde las sumas de índices se hacen módulo tres ((i+1)%3; ej:(2+1)%3 = 0)



Entonces, proyectando numerador y denominador en la dirección del vector $\mathbf{a}$:

$$\alpha^i = \mathbf{a}^i\cdot\mathbf{a} \,/\, \mathbf{a}\cdot\mathbf{a} = \mathbf{a}^i\cdot\mathbf{a} \,/\, a^2$$

Lo mismo sucede con volúmenes de tetraedros. Los volúmenes orientados de los tetraedros son productos triples (volumen del paralelepípedo dividido por seis, pero el seis se cancela en las divisiones):

$$v = ((\mathbf{x}_1-\mathbf{x}_0) \times (\mathbf{x}_2-\mathbf{x}_0)) \cdot (\mathbf{x}_3-\mathbf{x}_0) \qquad \alpha^i = ((\mathbf{x}_{i+1}-\mathbf{x}) \times (\mathbf{x}_{i+2}-\mathbf{x})) \cdot (\mathbf{x}_{i+3}-\mathbf{x}) / v$$

con las sumas de índices en módulo cuatro.

Es muy sencillo hacer los algoritmos de modo que realicen estas operaciones en forma ordenada.

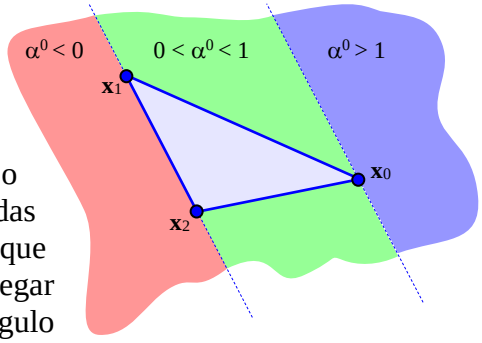
Se puede recordar así: El peso de un nodo en un punto variable es igual al cociente entre el área opuesta y la total. (La longitud en 1D o el volumen en 3D). También hay que recordar que las áreas, longitudes y volúmenes se definen mediante operaciones ordenadas de modo que resulten positivos en el interior.

El procedimiento que acabamos de explicar se suele denominar "inversión de la interpolación lineal" pues utilizamos una variable que se interpola linealmente (longitud, área o volumen), para encontrar los pesos que logran ese resultado. El proceso que consiste en encontrar los pesos que logren un determinado resultado, es la "inversa" de usar esos pesos para lograr una interpolación. Invertir otras formas de interpolación que veremos en breve, puede resultar muy difícil y a veces imposible.

Toda combinación afín, de n+1 puntos $\{\mathbf{x}_i\}$ con independencia afín, es la representación de un punto en un sistema de referencia n-dimendsional, cuyo origen es $\mathbf{x}_0$ y sus bases los $\{\mathbf{x}_i-\mathbf{x}_0\}$. Para estimar las coordenadas baricéntricas de un punto, basta con construir mentalmente ese sistema y estimar todas las coordenadas; excepto la 0, que es uno menos la suma de las otras.

Para los vértices de un triángulo, cada línea paralela al segmento $[\mathbf{x}_1,\mathbf{x}_2]$, tendrá una coordenada de área α^0 constante; pues, cualquier punto de la misma define un triángulo de altura constante con base en el segmento. Dichas líneas se denominan isolíneas o líneas isoparamétricas. Para un tetraedro, las isosuperficies o superficies isoparamétricas serán los planos paralelos a la cara opuesta de un nodo.

En la figura, la recta $(\mathbf{x}_1, \mathbf{x}_2)$ divide al plano en dos, podemos ver que $\alpha^0(\mathbf{x})$ es positivo para cualquier punto $\mathbf{x}$ en el semiplano que contiene a $\mathbf{x}_0$, es nula en la recta y negativa en el semiplano opuesto y es constante en las rectas paralelas. La simpleza del cálculo de las áreas permite averiguar fácilmente si un punto está o no dentro de un triángulo: solo basta ver que las tres coordenadas baricéntricas son positivas (sin hacer las divisiones) y, en caso de que no lo sean, nos da una pista inmediata de cómo movernos para llegar al triángulo; en una triangulación, nos permite averiguar cuál triángulo contiene al punto variable (lo veremos más adelante).



Las aplicaciones más interesantes no son nunca las más simples, normalmente se cuenta con varios puntos definidos (sin independencia afín); por ejemplo, se suelen obtener datos del clima (más de un valor: presión, temperatura, dirección del viento, velocidad, etc.) en distintos nodos o puntos de muestreo, que en este caso son estaciones meteorológicas. Para obtener en forma aproximada un valor en una localidad cualquiera se puede realizar una interpolación afín, pero no queda claro cuales estaciones meteorológicas deben tenerse en cuenta, cuales tienen peso no-nulo y qué peso tiene cada una. Un método estándar consiste en realizar una división del envoltorio convexo en triángulos no solapados y cuyos vértices son los nodos, esto es: una triangulación. Para calcular el valor en un punto cualquiera se pueden utilizan sólo las estaciones del triángulo que encierra al punto. Ésta técnica se conoce como interpolación poli-lineal o lineal por tramos (*piecewise linear interpolation*) o lineal a secas; las pocas variantes están dadas por los distintos métodos para realizar la triangulación.

Hay muchas otras técnicas (lineales o no) pero casi todas se basan en triangulaciones u otras divisiones estratégicamente simples del dominio (mallas). Para calcular la solución aproximada de una ecuación diferencial (Cálculo Numérico y Método de Elementos Finitos) también se utilizan estas mismas técnicas para la definición de las variables en el espacio.

En las aplicaciones gráficas, el color y las coordenadas de textura se interpolan de este modo, en los fragmentos de triángulos. Estos mismos cálculos de áreas están implementados en la placa gráfica. En el espacio del modelo se realizan sobre los triángulos 3D; pero en el espacio de la imagen, con los triángulos distorsionados por la proyección central, se realizan en 4D y ahora veremos cómo.

Interpolación afín en coordenadas homogéneas e interpolación hiperbólica

En coordenadas homogéneas, la coordenada w también resulta interpolada, como cualquier variable más. Lo extraño aparece al hacer la división por w, al volver al espacio "real": En la foto de una fila de árboles equiespaciados, los árboles no están equiespaciados. Pueden suceder cosas tan raras como que, en la imagen, el quinto árbol este justo en el medio entre el primer árbol y un árbol en el infinito.

Sean $\{w\mathbf{x},w\}$ (con $w\neq0$) las coordenadas homogéneas de nuestros puntos $\mathbf{x}$. Interpolemos en 4D:

$$\{w\mathbf{x},w\} = \alpha^0\{w_0\mathbf{x}_0,w_0\}+\alpha^1\{w_1\mathbf{x}_1,w_1\} = \{\alpha^0w_0\mathbf{x}_0,\alpha^0w_0\}+\{\alpha^1w_1\mathbf{x}_1,\alpha^1w_1\} \qquad (\alpha^0 + \alpha^1 = 1)$$

$$w = \alpha^0w_0+\alpha^1w_1 \qquad \Rightarrow \qquad \mathbf{x} = (\alpha^0w_0\mathbf{x}_0+\alpha^1w_1\mathbf{x}_1) / (\alpha^0w_0+\alpha^1w_1) = \beta^0\mathbf{x}_0+\beta^1\mathbf{x}_1$$

Los "pesos proyectados" (β) son distintos que los pesos en el espacio homogéneo (α):

$$\beta^i = \alpha^iw_i / \sum_j\alpha^jw_j = \alpha^i\, w_i / w \qquad \text{(sin suma en i, la suma en j está explícita)}$$

Los β_i también suman uno, pues cada peso del numerador está dividido por w que es la suma de todos. Sigue siendo una combinación afín, pero otra. Si los α^i son lineales, los β^i son hiperbólicos. La división por w no preserva (los coeficientes de) la combinación afín.

La combinación de dos puntos se extiende, en forma trivial y con idéntico resultado, a más puntos.

Hagamos un par de ejemplos. Tenemos dos puntos en 3D: $\mathbf{x}_0= \{0,0,0\}$ y $\mathbf{x}_1 = \{1,0,0\}$; los representamos en coordenadas homogéneas con el mismo w=1: $\mathbf{x}_0= \{0,0,0,1\}$ y $\mathbf{x}_1 = \{1,0,0,1\}$. Con pesos $\alpha^0 = \alpha^1 = 1/2$ el punto interpolado es: $\mathbf{x} = \{0,0,0,1\}/2+\{1,0,0,1\}/2 = \{1/2,0,0,1\}$ que equivale en 3D a $\{1/2,0,0\}$; en este caso, ambos w valen uno y el resultado no es sorprendente. Ahora, si aumenta el w del segundo punto: $\mathbf{x}_1=\{2,0,0,2\}$ (el mismo punto 3D) el nuevo resultado es $\mathbf{x} = \{0,0,0,1\}/2+\{2,0,0,2\}/2 = \{1,0,0,3/2\}$ que al pasar a 3D queda $\{2/3,0,0\}$, más cerca del segundo punto, que "pesa" el doble que el primero.

Puede pensarse que los w_i son pesos adicionales en el cálculo del promedio ponderado. De hecho, en algunos textos (y en particular para curvas y superficies NURBS) se introducen como pesos en lugar de coordenadas homogéneas. ($w \sim$ *weight* = peso)

La relación entre los pesos proyectados β^i y los originales α^i está alterada con la relación de los w_i:

$$\beta^i / \beta^j = \alpha^i\, w_i / \alpha^j\, w_j$$

Aquí se ve que **la transformación proyectiva no preserva la combinación afín** (el punto medio no se transforma en el punto medio). De hecho, puede haber un punto medio entre un punto "aquí" $\{\mathbf{x}_1,1\}$ y otro en el infinito $\{\mathbf{x}_2,0\}$; lo cual parece ridículo, a menos que estemos hablando de una imagen.

La interpolación hiperbólica[*] es la interpolación en el espacio proyectivo, no es más que una interpolación lineal en coordenadas homogéneas con una dimensión más.

Un problema algebraicamente idéntico se presenta para interpolar en la imagen o en la pantalla (posición, color, normal o textura) cuando, en realidad, lo que se pretende es interpolar linealmente en el espacio, antes de la proyección. En la proyección perspectiva, la coordenada visual z cumple un papel similar al w del espacio proyectivo, pero el plano de la imagen se identifica con el plano *near*, de donde z_{near} es como la "unidad" w=1 del plano proyectivo; entonces, en las fórmulas anteriores z/z_{near} cumple exactamente el mismo papel que w/1 pero, dado que son todos cocientes, z_{near} se cancela y se usa solo z. En la web este tema se encuentra como "*perspective correct*" (*interpolation* o *texture mapping*).

Ejemplo: se quiere averiguar donde se proyectará el punto medio entre $\mathbf{x}_0$ y $\mathbf{x}_1$ ($\alpha^0=\alpha^1=½$), se conocen las coordenadas en la imagen $\mathbf{y}_0$ e $\mathbf{y}_1$ y las profundidades visuales z_0 y z_1; se calculan:

$$\beta^1 = (0.5\, z_1)/[0.5\, (z_0+z_1)] = z_1/(z_0+z_1) \qquad \mathbf{y} = (1-\beta^1)\mathbf{y}_0+\beta^1\mathbf{y}_1 = (z_0\mathbf{y}_0+z_1\mathbf{y}_1)/(z_0+z_1).$$

Si el problema es el inverso: se está operando píxel a píxel en la pantalla, para ello se barre entre un punto y el otro con β^1 variable entre cero y uno y se necesita un valor correspondiente, interpolado con α^1 en el espacio real. Bastará con aplicar las relaciones anteriores, considerando suma uno:

$$\alpha^0/\alpha^1 = (\beta^0/z_0) / (\beta^1/z_1) \qquad \Rightarrow \qquad \alpha^0 = (\beta^0/z_0) / (\beta^0/z_0 + \beta^1/z_1); \quad \alpha^1 = (\beta^1/z_1) / (\beta^0/z_0 + \beta^1/z_1)$$

Esas ecuaciones se utilizan al rasterizar, para asignar colores o pegar las texturas correctamente.

[*] En la geometría plana de Euclides las paralelas no se cortan, en la hiperbólica (proyectiva) todo par de rectas distintas se corta en un punto y en la esférica, todo par de rectas distintas se corta en dos puntos

Interpolación bilineal

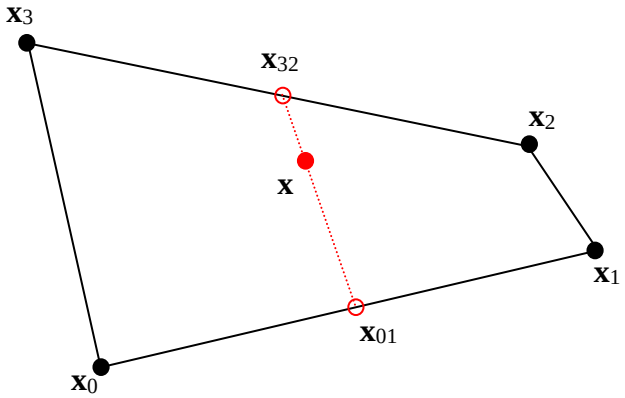
La combinación bilineal de los vértices de un cuadrilátero (no necesariamente en un mismo plano), consiste en usar una misma combinación lineal en dos segmentos opuestos, con un parámetro y luego hacer una combinación lineal de los dos puntos resultantes, con otro parámetro:

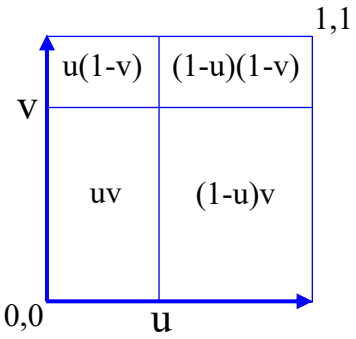
$$\mathbf{x}_{01} = (1-u)\,\mathbf{x}_0 + u\,\mathbf{x}_1$$

$$\mathbf{x}_{32} = (1-u)\,\mathbf{x}_3 + u\,\mathbf{x}_2$$

$$\mathbf{x} = (1-v)\,\mathbf{x}_{01} + v\,\mathbf{x}_{32} = \overbrace{(1-v)\,(1-u)}^{\alpha_0}\,\mathbf{x}_0 + \overbrace{(1-v)\,u}^{\alpha_1}\,\mathbf{x}_1 + \overbrace{v\,u}^{\alpha_2}\,\mathbf{x}_2 + \overbrace{v\,(1-u)}^{\alpha_3}\,\mathbf{x}_3$$

El resultado es idéntico si primero se hace la combinación con v en los segmentos ($\mathbf{x}_0$, $\mathbf{x}_3$) y ($\mathbf{x}_1$, $\mathbf{x}_2$), para luego combinar con u los puntos resultantes (probar).

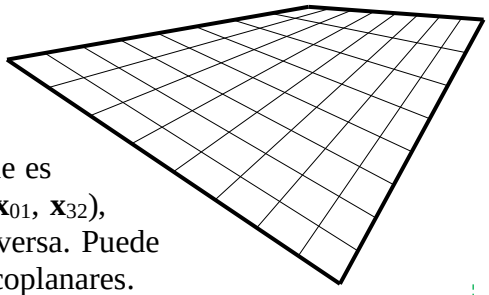




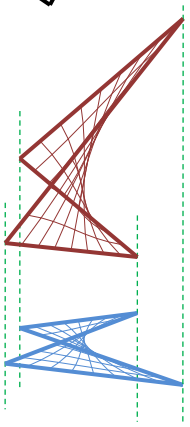
La denominación de bilineal proviene del uso de los parámetros, normalmente identificados con las coordenadas u y v de un cuadrado unitario en el espacio de parámetros. Un polinomio formado así, con productos de monomios lineales, sin ningún término cuadrático en u o en v, se denomina bilineal.

Los cuatro parámetros de la combinación afín resultante: {(1-v) (1-u), (1-v) u, v u, v (1-u)} suman uno y, si $(u,v) \in [0,1]^2$ también estarán en el rango [0,1]. El resultado es una interpolación afín convexa, pero no−lineal, sino bilineal, de los cuatro puntos y valores; se trata solo de una receta para la asignación de los pesos. Es decir: la interpolación bilineal es una forma particular de expresar los pesos de la combinación convexa de cuatro puntos que forman un cuadrángulo (en el plano o en el espacio), pero usando sólo dos parámetros. La construcción de los parámetros como áreas opuestas, en la figura de arriba a la derecha, es solamente válida en el espacio de parámetros, pues la no-linealidad hace que las áreas no se correspondan con las parcelas del cuadrilátero real.

En computación gráfica se suele utilizar este tipo de interpolación para mapear texturas o interpolar colores en cuadriláteros, el resultado es más suave que el obtenido dividiendo en dos triángulos (que es lo que hace OpenGL).



Los cuatro pesos de la combinación afín $\alpha^i(u,v)$ dependen de dos parámetros, tienen una restricción paramétrica del tipo que mencionamos antes. Al combinar así cuatro puntos no coplanares en 3D, se obtiene una superficie bilineal. La superficie es "reglada" porque está formada por segmentos rectos, como el ($\mathbf{x}_{01}$, $\mathbf{x}_{32}$), que se obtienen barriendo u entre cero y uno, con v fijo; o viceversa. Puede verse como una pompa de jabón entre cuatro alambres rectos no coplanares.



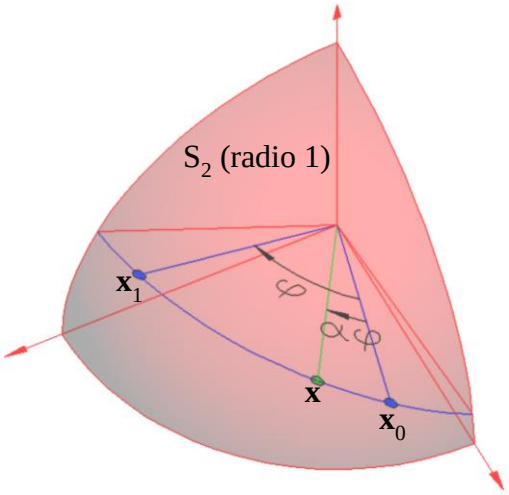
La función P(u,v) es biunívoca en el interior (se puede invertir: dado el punto se pueden calcular u y v) excepto cuando los cuatro puntos son coplanares y el cuadrilátero es cóncavo. Interpolando en un cuadrilátero cóncavo se obtienen, como siempre, puntos dentro del *convex-hull*; pero, a veces, fuera del cuadrilátero y los puntos exteriores al cuadrilátero cóncavo se pueden obtener con dos conjuntos de parámetros distintos. (Pensarlo como a la derecha: una superficie bilineal 3D proyectada)

El mismo concepto se extiende a un "cubo" y se utiliza la interpolación trilineal para mapear el interior de un cubo unitario en un hexaedro o cuboide de seis caras bilineales en el espacio. En CG eso se utiliza para mapear texturas (2D con *mipmaps* o 3D), también se utiliza en Cálculo Numérico (Elementos Finitos) en 3D.

Interpolación esférica lineal o *Slerp* (*Spherical Linear Interpolation*)

Se trata de un método de interpolación entre dos puntos de una esfera unitaria y se utiliza para interpolar direcciones, orientaciones o normales, que se describen mediante vectores unitarios; o bien rotaciones que se representan mediante cuaterniones, como veremos en breve.

A diferencia de todas las anteriores esta no es una interpolación afín y además se interpolan coordenadas, pero no variables asociadas (no las hay, al menos no en CG).



Se denomina n−esfera S_n a la generalización dimensional del concepto de esfera unitaria S_2. En n+1 dimensiones, es una superficie n−dimensional, cuyos puntos distan una unidad del origen. La 2−esfera S_2 es la superficie de una esfera común de radio 1 y centrada en el origen en 3D.

En cualquier superficie, las curvas que unen puntos, con mínima longitud del recorrido, se llaman **geodésicas**. Son el equivalente de las rectas en el plano (en el espacio euclídeo).

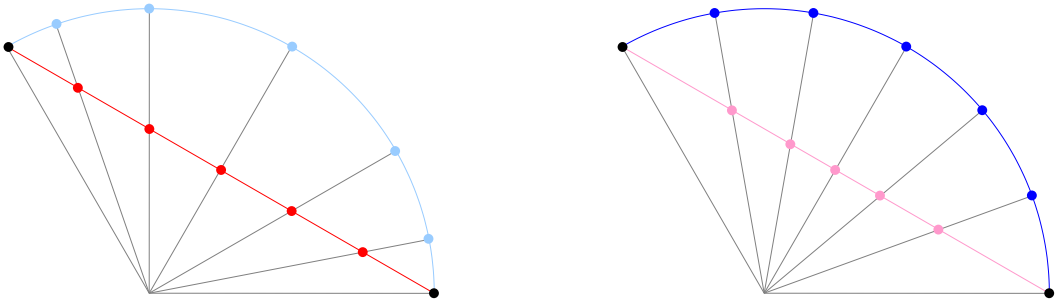
El camino más corto entre dos puntos de una esfera, pero recorriendo puntos de la superficie esférica, es un arco de circunferencia máxima. Las circunferencias máximas son las que tienen el mismo centro y radio que la esfera. Cualquier otra circunferencia en la superficie, tendrá un radio menor. Por ejemplo: de los paralelos de la tierra, solo el ecuador es un círculo máximo; mientras que todos los meridianos lo son. En S_n las geodésicas son circunferencias centradas en el origen y tienen radio unitario.

Si dos puntos de la esfera, $\mathbf{x}_0$ y $\mathbf{x}_1$, no son antípodas (no están alineados con el origen), habrá una sola circunferencia máxima que los une. Si son antípodas, habrá infinitas. De Buenos Aires a Sidney hay un único camino de mínima longitud, pero del Polo Norte al Polo Sur hay infinitos meridianos con la misma longitud. Rara coincidencia: también son antípodas Taipei en la isla Formosa o Taiwan y El Espinillo en la provincia argentina de Formosa.

La interpolación esférica, entre dos puntos de una esfera, es un punto intermedio, pero en una geodésica y no en la recta que los une. La interpolación esférica lineal o *slerp* es una interpolación lineal de los ángulos centrales: Si $\varphi \in [0,\pi)$ es el ángulo entre dos versores o el ángulo central de dos puntos de la esfera unitaria, y $u \in [0,1]$ es el parámetro de la interpolación; queremos un nuevo versor, en el mismo plano, pero angularmente distanciado $u\varphi$ del primero y en dirección al segundo:

$$\mathbf{x} = \text{slerp}(\mathbf{x}_0, \mathbf{x}_1, u)$$

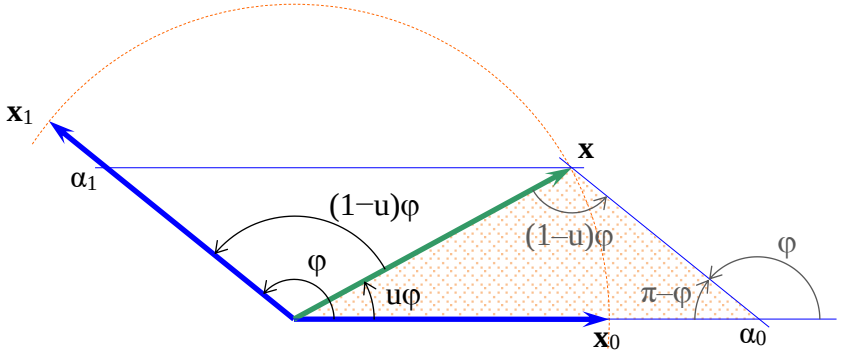
La interpolación esférica es lineal (celeridad o rapidez constante) en la superficie de la esfera, pero no en el espacio, el camino recorrido por el punto móvil es un arco de longitud $u\varphi$ (porque el radio es uno). El camino recorrido por unidad de "tiempo" (d/du) es constante, pero no el vector velocidad, que no cambia su módulo (celeridad) pero sí su dirección (sólo tiene aceleración centrípeta).



En las figuras anteriores se hace evidente la diferencia entre una interpolación lineal (*lerp*) de direcciones y una esférica (*slerp*). La diferencia (muy sutil) se puede apreciar visualmente al interpolar linealmente (mal, pero así se hace) las normales para iluminación; por ejemplo, para el sombreado de Phong.

Cualquiera sea la cantidad de dimensiones, todo el proceso de interpolación descansa en el plano de los dos versores; por lo tanto, haremos los cálculos en ese plano.

Partimos de los versores dados, $\mathbf{x}_0$ y $\mathbf{x}_1$ y del parámetro u variable. Debemos calcular el ángulo φ entre los versores y luego el versor interpolado **x**(u). Por lo antedicho, es necesario que los puntos a interpolar no sean antípodas; es decir: que los versores no sean opuestos.



En primer lugar, dado que los módulos son unitarios, notemos que podemos calcular:

$\varphi = \cos^{-1}(\mathbf{x}_0 \cdot \mathbf{x}_1)$

(Si el producto escalar fuese −1, los puntos son antípodas y hay que marcar un error).

Podemos representar **x** (o cualquier vector del plano) usando a $\mathbf{x}_0$ y $\mathbf{x}_1$ como bases:

$\mathbf{x} = \alpha^0 \mathbf{x}_0 + \alpha^1 \mathbf{x}_1$. (Notar que: $\alpha^0 + \alpha^1 \neq 1$ porque **x** no es combinación afín)

En el triángulo indicado, se muestran los ángulos: uφ, (1-u)φ y π–φ (cuyo seno es igual al seno de φ). Las longitudes de los lados opuestos son: $|\alpha^1\mathbf{x}_1|=\alpha^1$, $|\alpha^0\mathbf{x}_0|=\alpha^0$ y $|\mathbf{x}|=1$. Usando el teorema del seno:

$$\frac{1}{\text{sen}(\varphi)} = \frac{\alpha_0}{\text{sen}[(1-u)\varphi]} = \frac{\alpha_1}{\text{sen}(u\varphi)}$$

De allí se deducen α_0 y α_1, quedando:

$$\mathbf{x} = \frac{\text{sen}[(1-u)\varphi]}{\text{sen}(\varphi)}\mathbf{x}_0 + \frac{\text{sen}(u\varphi)}{\text{sen}(\varphi)}\mathbf{x}_1$$

Para evitar la posible división por cero o la inestabilidad, cuando φ sea muy pequeño, hay que hacer una aproximación lineal. Cuando φ tiende a cero, la ecuación anterior tiende a la simple interpolación lineal (el seno tiende al ángulo y el arco tiende a la cuerda):

$\mathbf{x} \approx (1 - u)\,\mathbf{x}_0 + u\,\mathbf{x}_1$

A nivel de implementación, para evitar la inestabilidad bastará saber si se pueden equiparar el ángulo y su seno. Dado que el mayor ángulo involucrado es φ, basta saber si $|\varphi - \text{sen}(\varphi)| < \varepsilon$, donde ε es un pequeño valor que depende de la precisión numérica utilizada. Haciendo la expansión en serie de Taylor, se ve fácilmente que $|\varphi - \text{sen}(\varphi)| \sim \varphi^3/6$ (φ siempre es positivo), podemos comparar φ con 10^{-3} o 10^{-6} en simple o doble precisión respectivamente y si es menor usar una interpolación lineal **normalizada** (dividiendo por el módulo del resultado) para que **x** vuelva a la esfera unitaria. Comparar φ con 10^{-3} es lo mismo que comparar el sen(φ) con 10^{-3} o su coseno (ya calculado) con .9999995.:

$c = \mathbf{x}_0 \cdot \mathbf{x}_1$ $c < -.9999995 \Rightarrow$ opuestos, error. $c > .9999995 \Rightarrow$ cercanos, lineal

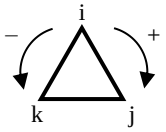
Un gran problema de la interpolación esférica es que, a diferencia de la afín: *slerp* de *slerp* no es *slerp*, la *slerp* de un versor con otro, obtenido este último por *slerp* de otros dos versores, depende del orden de selección. Es muy poco elegante que la interpolación de las normales de los vértices de un triángulo dependa del orden. Una solución consistiría en usar las áreas de los triángulos esféricos para hacer un equivalente de las coordenadas baricéntricas y renormalizar (los cocientes de áreas de triángulos esféricos son idénticos a los cocientes entre los respectivos ángulos sólidos). Esto aún no está estandarizado en matemáticas ni en la práctica y tampoco está implementado en las placas gráficas.

Cuaterniones

Si bien tienen múltiples aplicaciones, la única utilización práctica de los cuaterniones es en CG para representar e interpolar rotaciones y efectivamente son muy utilizados para eso en los videojuegos. La interpolación lineal entre dos vectores (sus componentes) produce un nuevo vector intermedio, un punto intermedio si se trata de vectores posición. La interpolación lineal entre dos matrices también produce una nueva matriz; vista como transformación, sus columnas dan vectores base interpolados. Pero en particular, la interpolación de dos matrices de rotación, no define una rotación intermedia, ni siquiera define una rotación, habría que "reortogonalizar" (si se hace slerp de las columnas ¿las bases quedan perpendiculares entre sí?). Este y otros problemas (*gimbal lock*) llevaron al uso de cuaterniones.

Los cuaterniones fueron inventados por Hamilton, buscando una división de vectores en 3D como la que permite el algebra de complejos en 2D. Un cuaternión se forma como un vector de cuatro dimensiones cuya base son los números {1,i,j,k}. El 1 es la unidad real estándar; i, j y k son extensiones del imaginario puro i. El cuadrado de cualquiera de ellos es −1. Los productos mutuos (más allá del trivial 1) son:

$i^2 = j^2 = k^2 = ijk = -1$; $ij = -ji = k$; $jk = -kj = i$; $ki = -ik = j$



Notar la antisimetría. La regla mnemotécnica es: ijk en orden circular, cada par sucesivo da el siguiente y en orden inverso da negativo.

Un cuaternión es entonces una combinación lineal q = ai + bj + ck + d, donde a, b, c y d son números reales. Tiene una componente real: d y tres imaginarias: a, b y c. También podemos representarlo como q = <d;**u**> donde **u** es el vector tridimensional {a,b,c} que engloba las tres componentes imaginarias ({i,j,k} que la historia confundió con la base euclídea), en este caso se habla de una componente escalar y una vectorial o cuaternión pura. (Notar la relación con la definición de planos.)

Con esa notación podemos representar escalares y vectores como cuaterniones haciendo la distinción:

$< d, \mathbf{0} >$ representa el escalar d;

$< 0, \mathbf{u} >$ representa el vector o cuaternión puro **u**.

La operatoria de los cuaterniones es idéntica a la de los complejos. Para el producto (todos contra todos) hay que tener en cuenta los productos mutuos de los números que conforman la base. Para poder usar la notación estándar con los vectores (**u**), se reservan para ellos los términos "producto escalar" y "producto vectorial" con sus símbolos "·" y "×"; mientras que se define el producto (a secas) entre cuaterniones, distribuyendo los productos. El producto de cuaterniones, ya reagrupados los términos, queda:

$$q_a q_b = < d_a, \mathbf{u}_a > < d_b, \mathbf{u}_b > = < (d_a d_b - \mathbf{u}_a \cdot \mathbf{u}_b), (d_a \mathbf{u}_b + d_b \mathbf{u}_a + \mathbf{u}_a \times \mathbf{u}_b) >$$

Se puede ver que el producto no es conmutativo; la no−conmutatividad está en el producto vectorial, pues si se invierten los factores, resulta opuesto. La conmutatividad sólo se da, entonces, cuando el producto vectorial es nulo, las partes vectoriales son paralelas u opuestas o alguna nula:

$$< a, \mathbf{u} > < b, \gamma\, \mathbf{u} > = < (ab - \gamma\, \mathbf{u}^2), (a\gamma + b)\, \mathbf{u} > = < b, \gamma\, \mathbf{u} > < a, \mathbf{u} >$$

El conjugado de $q = < d, \mathbf{u} >$ es otro cuaternión con la parte vectorial opuesta:

$$q^* = < d, -\mathbf{u} >$$

Se puede ver muy fácilmente que:

$$qq^* = q^*q = a^2+b^2+c^2+d^2 = d^2+\mathbf{u}^2$$ (es un escalar)

La norma o magnitud de un cuaternión se define como en el caso complejo:

$$||q|| = \sqrt{qq^*} = \sqrt{a^2 + b^2 + c^2 + d^2} = \sqrt{d^2 + \mathbf{u}^2}$$ (Pitágoras 4D)

Mirando la fórmula anterior es fácil deducir que:

$$q^{-1} = \frac{q^*}{||q||^2}$$; siendo: $qq^{-1} = q^{-1}q = < 1, 0 > = 1.$ (Conmutativo por ser vectores antiparalelos)

Un cuaternión se puede normalizar, igual que un vector, dividiéndolo por su módulo:

$$q_n = \frac{q}{||q||}$$

Para un cuaternión unitario: $q_n^{-1} = q_n^*$.

Todo cuaternión unitario se puede escribir como: $q_n = < \cos(\alpha); \mathbf{u}_n \operatorname{sen}(\alpha) >$; donde $\mathbf{u}_n$ es, a su vez, un vector unitario:

$$q_n = < \cos(\alpha), \mathbf{u}_n \operatorname{sen}(\alpha) >;\ \mathbf{u}_n^2 = 1 \ \Rightarrow\ ||q_n|| = \sqrt{\cos^2(\alpha) + \operatorname{sen}^2(\alpha)\, \mathbf{u}_n^2} = 1$$

Rotación con cuaterniones:

La siguiente expresión combina el vector **v** con el cuaternión r:

$$\underline{\mathbf{v}} = r\,\mathbf{v}\,r^{-1} = r < 0, \mathbf{v} > r^{-1}$$

En primer lugar veamos que el resultado es el mismo si normalizamos el cuaternión:

$$r^{-1} = (\|r\|r_n)^{-1} = \|r\|^{-1} r_n^{-1} \quad \Rightarrow \quad (\|r\|r_n)\,\mathbf{v}\,(\|r\|r_n)^{-1} = r_n\,\mathbf{v}\,r_n^{-1}.$$

Esto simplifica la tarea, pues $r_n^{-1} = r_n^*$. Usaremos cuaterniones unitarios (sin subíndice $_n$).

En segundo lugar, notemos que nos hemos adelantado al llamar $\underline{\mathbf{v}}$ al resultado, pues aún no sabemos qué tipo de cuaternión es. Enseguida veremos que, efectivamente, es un vector.

Apliquemos esa expresión a un vector $\mathbf{v}_{//}$ paralelo u opuesto a **u** ($\mathbf{v}_{//} = \gamma\,\mathbf{u}$). Usando la fórmula del producto y siendo $\mathbf{u}^2 = 1$:

$$\begin{aligned}
\underline{\mathbf{v}}_{//} &= < \cos(\alpha), \mathbf{u}\,\mathrm{sen}(\alpha) > < 0, \gamma\,\mathbf{u} > < \cos(\alpha), -\mathbf{u}\,\mathrm{sen}(\alpha) > = \\
&= < -\gamma\,\mathrm{sen}(\alpha), \gamma\cos(\alpha)\,\mathbf{u} > < \cos(\alpha), -\mathbf{u}\,\mathrm{sen}(\alpha) > = \\
&= < [\gamma\,\mathrm{sen}(\alpha)\cos(\alpha) - \gamma\,\mathrm{sen}(\alpha)\cos(\alpha)], \gamma\,\mathbf{u}\,[\cos^2(\alpha) + \mathrm{sen}^2(\alpha)] > = < 0, \gamma\,\mathbf{u} > = \\
&= \mathbf{v}_{//}.
\end{aligned}$$

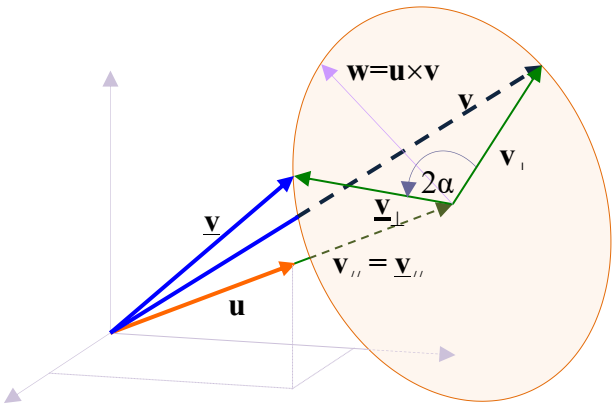
Es decir que un vector paralelo a **u** no cambia con esa expresión.

Ahora apliquemos la misma expresión a un vector $\mathbf{v}_\perp$ perpendicular a **u**. Llamando **w** al vector $\mathbf{u} \times \mathbf{v}_\perp$ que es perpendicular tanto a **u** como a $\mathbf{v}_\perp$ y notando además que $\mathbf{u} \times \mathbf{w} = -\mathbf{v}_\perp$:

$$\begin{aligned}
\underline{\mathbf{v}}_\perp &= < \cos(\alpha), \mathbf{u}\,\mathrm{sen}(\alpha) > < 0, \mathbf{v}_\perp > < \cos(\alpha), -\mathbf{u}\,\mathrm{sen}(\alpha) > = \\
&= < 0, \cos(\alpha)\,\mathbf{v}_\perp + \mathrm{sen}(\alpha)\,\mathbf{w} > < \cos(\alpha), -\mathbf{u}\,\mathrm{sen}(\alpha) > = \\
&= < 0, \cos^2(\alpha)\,\mathbf{v}_\perp + \mathrm{sen}(\alpha)\cos(\alpha)\,\mathbf{w} + \mathrm{sen}(\alpha)\cos(\alpha)\,\mathbf{w} - \mathrm{sen}^2(\alpha)\,\mathbf{v}_\perp > = \\
&= < 0, \{[\cos^2(\alpha) - \mathrm{sen}^2(\alpha)]\,\mathbf{v}_\perp + 2\,\mathrm{sen}(\alpha)\cos(\alpha)\,\mathbf{w}\} > = \\
&= < 0, \cos(2\alpha)\,\mathbf{v}_\perp + \mathrm{sen}(2\alpha)\,\mathbf{w} >.
\end{aligned}$$

La expresión aplicada a un vector $\mathbf{v}_\perp$, perpendicular a **u**, es un giro de ángulo 2α y eje **u**.

Dado que todo vector puede representarse como suma de un vector paralelo a **u** y uno perpendicular; la fórmula anterior, efectivamente transforma vectores en vectores y, en particular, es un giro de ángulo 2α alrededor del eje **u**.



Dado que el producto es asociativo, la composición de rotaciones se puede hacer mediante el producto de cuaterniones:

$$q_1\,q_2\,q_2^{-1}\,q_1^{-1} = 1 \;\Rightarrow\; (q_1\,q_2)^{-1} = q_2^{-1}\,q_1^{-1}$$
$$q_1\,(q_2\,\mathbf{v}\,q_2^{-1})\,q_1^{-1} = (q_1\,q_2)\,\mathbf{v}\,(q_1\,q_2)^{-1}.$$

Para armar la matriz de rotación que produce el mismo resultado que la aplicación de un dado cuaternión q, basta recordar que sus columnas son los versores base transformados: $\{q\,\mathbf{e}_x\,q^{-1}, q\,\mathbf{e}_y\,q^{-1}, q\,\mathbf{e}_z\,q^{-1}\}$. Encontrar el cuaternión equivalente a una dada matriz de rotación es más difícil, pero se puede deducir resolviendo un sistema de ecuaciones o utilizando las fórmulas que aparecen en la bibliografía y en *web*. El problema es la falta de unicidad, puesto que dos cuaterniones unitarios opuestos producen la misma rotación (analizar).

La interpolación de rotaciones se realiza mediante *slerp* de cuaterniones en S_3 (esfera unitaria en R^4) con la única advertencia de que para calcular el ángulo entre los vectores se utiliza el producto escalar estándar de R^4 y no el producto de cuaterniones.

Con estos métodos también se puede trazar una curva cualquiera en la superficie de S_3 (una spline definida por secuencias de interpolaciones esféricas) y realizar una animación suave de los movimientos de un objeto con giros complejos (trompo, helicóptero) y de hecho así se utiliza en la práctica, fundamentalmente en juegos interactivos y animaciones.

Nestor Calvo
Computación Grafica – FICH
agosto de 2018