Useful Results in Spherical Harmonics (mainly 2-band)

Introduction

(Real) spherical harmonics are a set of basis functions defined over the sphere of directions, usually denoted $\{Y_l^m\}$, where $l \geq 0$ is the band, and m is the index within the band ($-l \leq m \leq l$).

A function defined on the sphere can be approximated by a weighted sum of these basis functions up to some number of bands. Using more bands leading to a more accurate approximation, as one would expect.

It is often convenient to flatten out this 2-index numbering scheme into a 1-indexed one, replacing (m,l) with the single index i, as follows:

$$Y_l^m(\theta, \phi) = Y_i(\theta, \phi) \text{ where } i = l(l+1) + m$$

The functions $\{Y_i\}$ satisfy the orthonormal property

$$\int_S Y_i Y_j dS = \delta_{ij}$$

where the integral is evaluated is over the full sphere S, and δ_{ij} is the Kronecker delta.

Parametrization

Here's the spherical polar coordinate system we'll use (I'll get round to a proper diagram eventually):

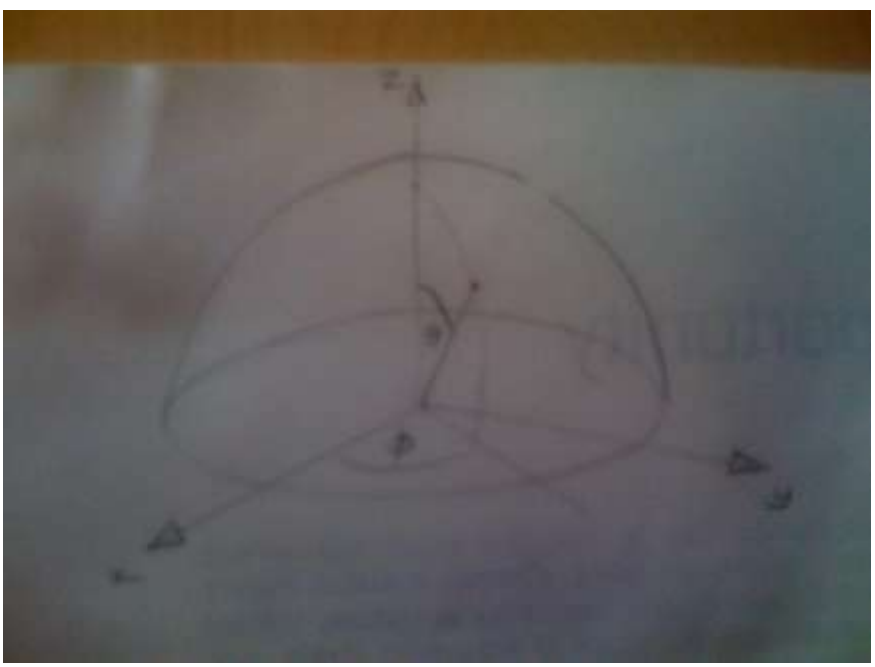

Figure 1: coordinate system for spherical harmonics

θ measures the angle from the z-axis, and ϕ the angle from the x-axis measured anticlockwise in the xy-plane. So a point (x,y,z) on the unit sphere parameterizes as

$$\begin{aligned} x &= \sin\theta\cos\phi \\ y &= \sin\theta\sin\phi \\ z &= \cos\theta \end{aligned}$$

Basis functions

Following Green, here is the orthonormal basis for the first 2 SH bands (the only difference from other sources being a change of sign in a couple of the functions):

$$Y_0^0(\theta,\phi) = \frac{1}{2\sqrt{\pi}} \approx 0.265079452$$

$$Y_1^{-1}(\theta,\phi) = \frac{\sqrt{3}}{2\sqrt{\pi}} y = \frac{\sqrt{3}}{2\sqrt{\pi}} \sin\theta\sin\phi \approx 0.488602512\, y$$

$$Y_1^0(\theta,\phi) = \frac{\sqrt{3}}{2\sqrt{\pi}} z = \frac{\sqrt{3}}{2\sqrt{\pi}} \cos\theta \approx 0.488602512\, z$$

$$Y_1^1(\theta,\phi) = \frac{\sqrt{3}}{2\sqrt{\pi}} x = \frac{\sqrt{3}}{2\sqrt{\pi}} \sin\theta\cos\phi \approx 0.488602512\, x$$

We'll give these basis functions additional names as follows:

$$Y_x = Y_1^1(\theta,\phi)$$

$$Y_y = Y_1^{-1}(\theta,\phi)$$

$$Y_z = Y_1^0(\theta,\phi)$$

$$Y_w = Y_0^0(\theta,\phi)$$

so that when useful we can store them in a 4-component vector as (Y_x, Y_y, Y_z, Y_w).
We might also prefer to use the basis (x,y,z,1), which is just a componentwise scaling of (Y_x, Y_y, Y_z, Y_w). The former has the advantage of being trivial to compute, but the disadvantage that computing certain results (such as the integral of a product) requires more work since (x,y,z,1), while still an orthogonal set of basis functions, is no longer orthonormal.

Function projection

To project a function A(θ,ϕ) onto our SH basis, we compute a coefficient vector (A_x, A_y, A_z, A_w), where

$$A_x = \int_0^{2\pi} \int_0^{\pi} A(\theta, \phi) Y_x(\theta, \phi) \sin\theta \, d\theta \;\, d\phi$$

and similarly for A_y, A_z, A_w .
The presence of $\sin\theta$ in this equation is due to the particular parametrization we're using: $\sin\theta \, d\theta \;\, d\phi$ represents the differential area of an elemental patch centred on coordinates (θ,ϕ).

The projected function, $\tilde{A} = A_x Y_x + A_y Y_y + A_z Y_z + A_w Y_w$, represents an approximation to the original function, i.e. $\tilde{A} \approx$ A.

Integral of a product of functions

A fundamentally useful result in SH theory is that the integral of the product of two projected functions $\tilde{A}$ and $\tilde{B}$ over the full sphere is simply the dot product of their vectors of SH coefficients,

$$\int_0^{2\pi} \int_0^{\pi} \tilde{A}(\theta, \phi) \tilde{B}(\theta, \phi) \sin\theta \, d\theta \;\, d\phi = A_x B_x + A_y B_y + A_z B_z + A_w B_w$$

This is easily shown using the orthonormal property of the basis functions.

Projection of a product of functions

Given two projected functions $\tilde{A}$ and $\tilde{B}$, the projection of their product has SH coefficients

$$\frac{1}{2\sqrt{\pi}} (A_x B_w + A_w B_x \quad A_y B_w + A_w B_y \quad A_z B_w + A_w B_z \quad A_x B_x + A_y B_y + A_z B_z + A_w B_w)$$

An alternative form for the projection of $\tilde{A}\tilde{B}$ is

$$\frac{1}{2\sqrt{\pi}} (A_x \quad A_y \quad A_z \quad A_w) \begin{pmatrix} B_w & 0 & 0 & B_x \\ 0 & B_w & 0 & B_y \\ 0 & 0 & B_w & B_z \\ B_x & B_y & B_z & B_w \end{pmatrix}$$

In this form, the 4x4 matrix acts as a *transfer matrix* which operates on the vector of $\tilde{A}$'s coefficients (see Green, p.19).

Use of Mathematica to evaluate the integrals

Both of the above results for products can be shown in Mathematica using the following script:

```
Yx[t_,p_] := Sqrt[3]/(2*Sqrt[Pi])*Sin[t]*Cos[p]
Yy[t_,p_] := Sqrt[3]/(2*Sqrt[Pi])*Sin[t]*Sin[p]
Yz[t_,p_] := Sqrt[3]/(2*Sqrt[Pi])*Cos[t]
Yw[t_,p_] := 1/(2*Sqrt[Pi])
A[t_,p_] := Ax*Yx[t,p] + Ay*Yy[t,p] + Az*Yz[t,p] + Aw*Yw[t,p]
B[t_,p_] := Bx*Yx[t,p] + By*Yy[t,p] + Bz*Yz[t,p] + Bw*Yw[t,p]
Integrate[A[t,p]*B[t,p]*Sin[t], {t, 0, Pi}, {p, 0, 2*Pi}]
Integrate[A[t,p]*B[t,p]*Yx[t,p]*Sin[t], {t, 0, Pi}, {p, 0, 2*Pi}]
Integrate[A[t,p]*B[t,p]*Yy[t,p]*Sin[t], {t, 0, Pi}, {p, 0, 2*Pi}]
Integrate[A[t,p]*B[t,p]*Yz[t,p]*Sin[t], {t, 0, Pi}, {p, 0, 2*Pi}]
Integrate[A[t,p]*B[t,p]*Yw[t,p]*Sin[t], {t, 0, Pi}, {p, 0, 2*Pi}]
```

Rotation of a function

A handy property of spherical harmonics is their rotational invariance: Given a function G which is a rotated copy of a function F, then for any direction ω it is true that $\tilde{G}(\omega) = \tilde{F}\big(R(\omega)\big)$ where R describes the rotation.

In general, the coefficients of $\tilde{G}$ can be derived from those of $\tilde{F}$ by a purely linear transformation, i.e. by applying an $n^2 \times n^2$ matrix where n is the number of bands under consideration. The derivation of the matrix elements is difficult in the general case, but in our 2-band case the 4x4 matrix has a particularly simple form: it is just the rotation matrix expressed in 4x4 homogeneous form:

$$(G_x \quad G_y \quad G_z \quad G_w) = (F_x \quad F_y \quad F_z \quad F_w)\begin{pmatrix} r_{xx} & r_{xy} & r_{zx} & 0 \\ r_{yx} & r_{yy} & r_{yz} & 0 \\ r_{zx} & r_{zy} & r_{zz} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Projection of a cosine lobe

A cosine lobe pointing in the direction of the positive z-axis can be described by the function

$$C(\theta) = \max\{\cos\theta, 0\}$$

Because C is independent of ϕ, it can be described purely in terms of *zonal harmonics* Y_l^0. In our case where we're only using 2 bands, the zonal harmonics are $Y_0^0 = Y_w$ and $Y_1^0 = Y_z$, and the projection of C has coefficients $\left(0 \quad 0 \quad \sqrt{\frac{\pi}{3}} \quad \frac{\sqrt{\pi}}{2}\right)$.

Now let C' be a rotated copy of C, such that the peak of the lobe of C' points in the direction of the unit vector (n_x, n_y, n_z) . The matrix for such a rotation, in row-major form, will have the given direction vector as its 3rd row and unknown elements in its 1st and 2nd rows. When we expand this to a 4x4 homogeneous matrix and apply it to the coefficients of $\tilde{C}$, we obtain the coefficients of the projection of the rotated lobe $\tilde{C}\,'$:

$$\left(\sqrt{\frac{\pi}{3}}n_x \quad \sqrt{\frac{\pi}{3}}n_y \quad \sqrt{\frac{\pi}{3}}n_z \quad \frac{\sqrt{\pi}}{2}\right)$$

This simple process can be applied to any function expressible in terms of 2-band zonal harmonics.

<u>3-band projection of a cosine lobe</u>

Should it be necessary to use 3 bands, the 3rd zonal harmonic is

$$Y_2^0(\theta) = \frac{1}{4}\sqrt{\frac{5}{\pi}}(3\cos^2\theta - 1)$$

and in the expansion of a cosine lobe it takes the coefficient $\frac{\sqrt{5\pi}}{8}$. Thus

$$C(\theta) \approx \frac{1}{4} + \frac{1}{2}\cos\theta + \frac{5}{32}(3\cos^2\theta - 1)$$

The figure on the next page shows both the 2- and 3-band approximations, superimposed on a plot of a true clamped cosine lobe.

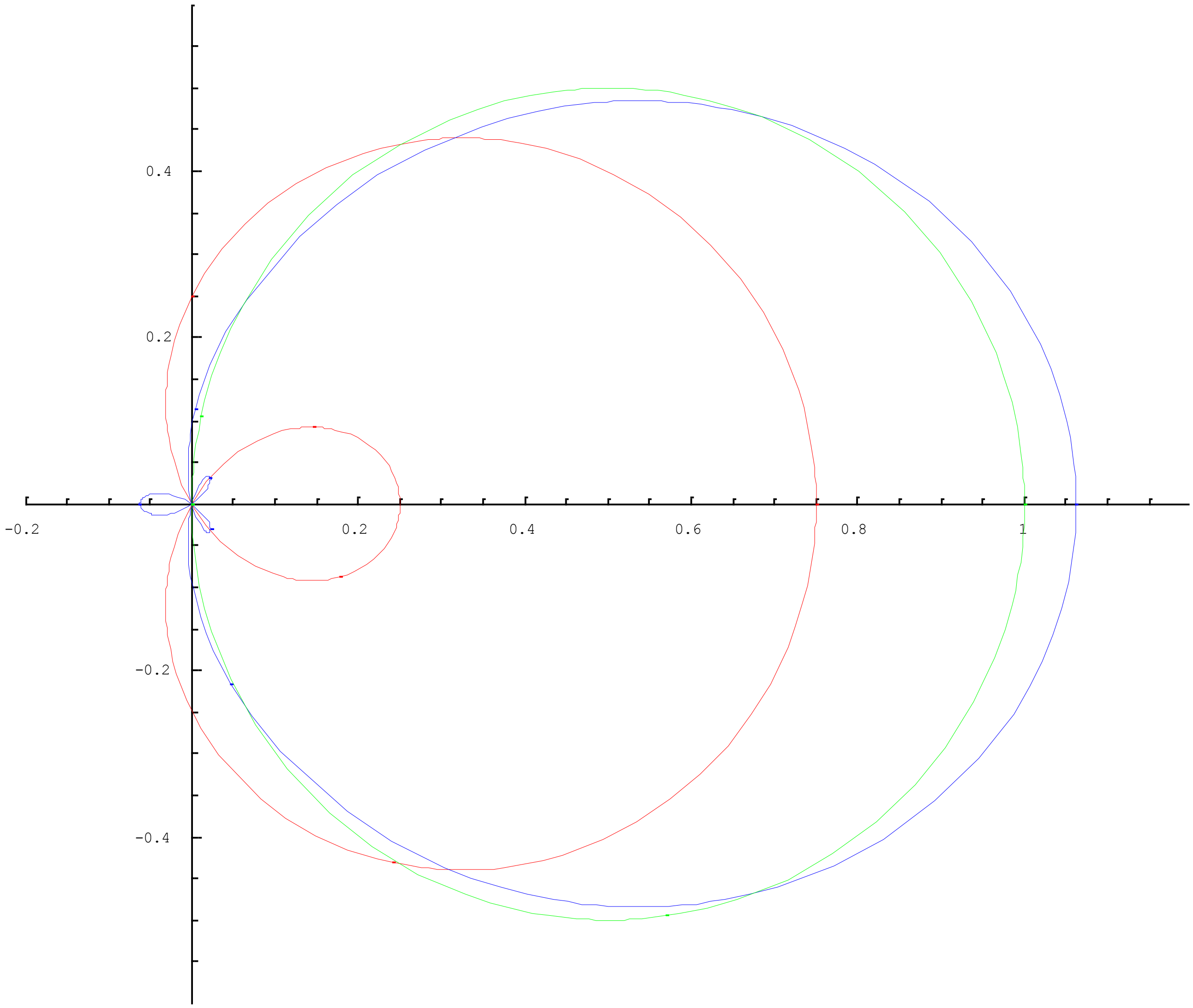

Figure 2: 2-band (red) and 3-band (blue) projections of a clamped cosine lobe (green)

Perfect reconstruction of a cosine lobe

Here's a neat trick: If it is known that a set of 2-band SH coefficients represents a pure clamped cosine lobe, the lobe can be exactly reconstructed by simply clamping a 3-component dot product and scaling appropriately.

Furthermore, provided we never sum anything 'sharper' than a cosine lobe into the coefficients, which is guaranteed to be the case for diffusely reflected light, then we can automatically detect when coefficients represent a single pure lobe.

First, generate a parameter t which indicates how much 'like a lobe' the SH coefficients are:

$$t = \frac{\sqrt{3}}{2}\frac{\sqrt{{c_x}^2 + {c_y}^2 + {c_z}^2}}{c_w}$$

Then, use this to interpolate between the true SH coefficients and a new set modified in such a way that when the result is clamped to zero the output will be the clamped cosine lobe we seek to reconstruct. The modified coefficients are

$$\begin{pmatrix} 2c_x & 2c_y & 2c_z & 0 \end{pmatrix}$$

We may get the best results by reparametrizing a bit, e.g. use t^2 or t^3 in place of t.

The solid angle of a texel

Though not an SH result per se, this useful integral gives the solid angle of any texel of a cube map:

$$\int_{y_0}^{y_1}\int_{x_0}^{x_1}(1+x^2+y^2)^{-\frac{3}{2}}\,dx\,dy = \left[\left[\tan^{-1}\frac{xy}{\sqrt{1+x^2+y^2}}\right]_{x=x_0}^{x=x_1}\right]_{y=y_0}^{y=y_1}$$

$$= \tan^{-1}\frac{x_1y_1}{\sqrt{1+{x_1}^2+{y_1}^2}} - \tan^{-1}\frac{x_0y_1}{\sqrt{1+{x_0}^2+{y_1}^2}} - \tan^{-1}\frac{x_1y_0}{\sqrt{1+{x_1}^2+{y_0}^2}} + \tan^{-1}\frac{x_0y_0}{\sqrt{1+{x_0}^2+{y_0}^2}}$$

It may be useful to express each inverse tangent as an inverse sine, where each argument has a form which factorizes:

$$\tan^{-1}\frac{xy}{\sqrt{1+x^2+y^2}} = \sin^{-1}\left(\frac{x}{\sqrt{1+x^2}}\frac{y}{\sqrt{1+y^2}}\right)$$

Constructing SH coefficients via Monte Carlo integration

Given an unbiased (with respect to solid angle) set of samples of a spherical function F, we can use Monte Carlo integration to compute the coefficients of its SH projection $\tilde{F}$. Two convenient sample sets are those arising from the 6 axial directions, and those arising from the 8 corner directions $\left(\pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}, \pm\frac{1}{\sqrt{3}}\right)$.

Axial directions:

$$c_x = \sqrt{\frac{\pi}{3}}\big(F(1,0,0) - F(-1,0,0)\big)$$

$$c_y = \sqrt{\frac{\pi}{3}}\big(F(0,1,0) - F(0,-1,0)\big)$$

$$c_z = \sqrt{\frac{\pi}{3}}\big(F(0,0,1) - F(0,0,-1)\big)$$

$$c_w = \frac{\sqrt{\pi}}{3}\sum F$$

Corner directions:

Let $F_{ijk} = F\left(\frac{1}{\sqrt{3}}(-1)^i,\ \frac{1}{\sqrt{3}}(-1)^j,\ \frac{1}{\sqrt{3}}(-1)^k\right)$. Then

$$c_x = \frac{\sqrt{\pi}}{4}\left(\sum_{j,k} F_{0jk} - \sum_{j,k} F_{1jk}\right)$$

$$c_y = \frac{\sqrt{\pi}}{4}\left(\sum_{i,k} F_{i0k} - \sum_{i,k} F_{i1k}\right)$$

$$c_z = \frac{\sqrt{\pi}}{4}\left(\sum_{i,j} F_{ij0} - \sum_{i,j} F_{ij1}\right)$$

$$c_w = \frac{\sqrt{\pi}}{4}\sum_{i,j,k} F_{ijk}$$

In both cases, if we start with a projected function, evaluate it in the given directions, and compute the resulting coefficients, we'll get back exactly the same function we started with. However, note that the converse is not true in general. i.e. given an *arbitrary* set of sampled values, if we compute the coefficients of the projection and evaluate the projected function in the given directions, we won't recover our starting sample values. (This is clearly true because both sampling arrangements contain more samples than there are 2-band coefficients.)