Spherical Harmonics for Lighting

Dr Graham Hazel
Geomerics, an ARM Company

Introduction

The Spherical Harmonic (SH) series of functions is the analogue of the Fourier Series for functions on the surface of a 2-sphere $\mathbb{S}^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$. The complete set of functions is an infinite-dimensional basis for functions on the sphere, but in practical use the series is truncated to give an approximation of an arbitrary function by a finite weighted sum of basis functions.

In computer graphics, spherical harmonics are used as a form of compression, which in turn greatly accelerates computations. For instance, the incoming light at a point in space is a spherical function (since it varies with direction), but using SH its approximation can be compactly represented by a handful of coefficients (the weights for the first few basis functions). This allows computations to be performed on a vector of these SH coefficients, rather than the spherical functions themselves.

Spherical harmonics are a good choice for this compressed representation because the SH approximation is invariant under rotation. That is, the SH approximation of a rotated function is the same as the rotated SH approximation of the original function.

However, the standard definition of the spherical harmonic basis functions is cumbersome and results in a number of superfluous constant factors. In the standard references these constants are often quoted as decimal expansions, making their origin even more mysterious. In this note we show that (for the lighting use case at least) most of the constants are unnecessary, and consequently arrive at a more easily comprehensible definition of spherical harmonics for graphics.

Spherical Harmonic Bands

The Spherical Harmonic Series is composed of separate "bands" of functions, usually denoted L0, L1, L2 etc. The L0 band is a single constant function; the L1 band consists of three linear functions; the L2 band contains five quadratic functions; and so on. For real-time lighting it is unusual to go past the L2 band (which requires nine total coefficients per channel), since the data and computation requirements become large, and the L2 approximation is already pretty accurate. However use cases in physics and astronomy can require up to L3000!

We will return to the bands later, but for now let's start with some definitions.

Truncated Weighted Sum

Given a basis $\mathcal{B} = \{B_i(\vec{\omega})_{i=0,1,...}\}$, a spherical function $R(\vec{\omega})$ can be written as a weighted sum

$$R(\vec{\omega}) = R_0 B_0(\vec{\omega}) + R_1 B_1(\vec{\omega}) + R_2 B_2(\vec{\omega}) + ...$$

where $R_0, R_1, ...$ are scalar coefficients.

Truncating the sum gives a finite approximation

$$R_{tr}(\vec{\omega}) \; = \; R_0 B_0(\vec{\omega}) + ... + R_n B_n(\vec{\omega})$$

where the vector of coefficients $(R_0, ..., R_n)$ fully describes the approximation.

Standard Inner Product

Recall that the standard definition of the inner product of spherical functions R, S is a spherical integral

$$\langle R, S\rangle_{std} = \oint R(\vec{\omega})\, S(\vec{\omega})\, d\Omega$$

where $d\Omega$ is the measure over the sphere and $\vec{\omega}$ is the variable of integration.

Therefore if the basis $\mathcal{B}$ is orthogonal we compute the coefficients R_i for a function R with integrals

$$R_i = \langle R, B_i\rangle_{std} = \oint R(\vec{\omega})\, B_i(\vec{\omega})\, d\Omega \qquad i = 0, 1, ...$$

Constant Basis Function

The standard definition of the first SH basis function – the single function in L0 – is

$$Y_0^0(\vec{\omega}) = \frac{1}{\sqrt{4\pi}}$$

This defintion is motivated by ensuring normalisation over the sphere, that is

$$\langle Y_0^0, Y_0^0\rangle_{std} \; = \; \oint (Y_0^0)^2 \, d\Omega \; = \; \oint \frac{1}{4\pi}\, d\Omega \; = \; 1$$

However for our use case this “extra” factor of $\sqrt{4\pi}$ is redundant. Consider what happens when we approximate a constant function $R(\vec{\boldsymbol{\omega}}) = \alpha$. First we compute the first SH coefficient

$$R_0 = \oint R(\vec{\boldsymbol{\omega}})\, Y_0^0 \, d\Omega \;=\; \oint \alpha \, \frac{1}{\sqrt{4\pi}} \, d\Omega \;=\; \sqrt{4\pi}\alpha$$

Now to construct the SH approximation of R we form the weighted sum of basis functions, that is we multiply the coefficient and the basis function

$$R_{sh}(\vec{\boldsymbol{\omega}}) = R_0 Y_0^0 = \frac{\sqrt{4\pi}\alpha}{\sqrt{4\pi}} = \alpha$$

As expected, the function R is reconstructed exactly, but the process is clearly more complicated than necessary: we gain a factor of $\sqrt{4\pi}$ in the SH coefficient only to cancel it with the same factor in the basis function.

The factor of 4π arises since it is the surface area of the unit sphere, but that isn’t relevant to our intended use case. It is much simpler to subsume this factor in a new definition of the inner product.

Normalised Inner Product

Redefine the inner product

$$\langle R, S \rangle \;=\; \frac{1}{4\pi} \oint R(\vec{\boldsymbol{\omega}})\, S(\vec{\boldsymbol{\omega}})\, d\Omega$$

The idea is to “normalise so the surface area of the sphere is one.” With this new definition, for $R(\vec{\boldsymbol{\omega}}) = \alpha$ we now obtain the more natural outcome that the corresponding SH coefficient $R_0 = \alpha$.

In fact this definition results in a 1-to-1 correspondence between the constructive approach of computing SH coefficients via sampling and the theoretical integral form

$$R_0 = \frac{1}{n} \sum_{i=1}^{n} R(\vec{\boldsymbol{\omega}}_i) \quad \mapsto \quad R_0 = \frac{1}{4\pi} \oint R(\vec{\boldsymbol{\omega}})\, d\Omega$$

where $\vec{\boldsymbol{\omega}}_1, ..., \vec{\boldsymbol{\omega}}_n$ are samples taken from a uniform spherical distribution.

This means that to compute the ith SH coefficient we simply sum the values of the ith basis function for n given sample directions and divide the result by n.

Notation

We will abuse the notation B_l^* for the basis functions, where the subscript l is the band, and the superscript is an index within the band. This is the similar to standard notation, except that we use B rather than Y to distinguish our renormalised basis functions.

L0 Band

With our renormalised inner product, we can define the first SH basis function

$$B_0^0(\vec{\boldsymbol{\omega}}) = 1$$

This is the only basis function which has a non-zero average value over the sphere. This means that the first SH coefficient for a function represents the average energy over the sphere – the subsequent coefficients and basis functions neither add nor remove energy (instead they simply "move it around").

L1 Band

Define

$$B_1^{-1,0,1} = x, y, z$$

Note that we have chosen functions which are not unit-length with respect to our inner product. This is deliberate. We choose to take care of the normalisation in the reconstruction step – that is, when we compute the approximation R_{sh}. One reason for this is that if R is radiance (incoming light) then for shading we need to convert it to irradiance (outgoing light) for a given normal direction. The reconstruction coefficients will get swept into this conversion.

In fact, it is convenient to think of the L1 band as a single vector $\vec{\boldsymbol{B}}_1 = (x, y, z)$ with corresponding SH coefficients $\vec{\boldsymbol{R}}_1$.

The direction of $\vec{\boldsymbol{R}}_1$ is the average direction of the function R. If R is radiance, then it is the average direction of the incoming light (weighted by intensity).

By construction, the length of $\vec{\boldsymbol{R}}_1$ will vary between 0 and R_0, the first SH coefficient. The ratio $|\vec{\boldsymbol{R}}_1|/R_0$ is an indication of "how directional" the function R is. If the ratio is one then R is completely directional – all of the energy is at a single point on the sphere. Conversely, if the ratio is zero then R is symmetrical, and the L1 band gives us no directional information at all.

L2 Band

There are six different quadratic combinations of the variables x, y and z: x^2, y^2, z^2, xy, yz and xz. However, since we are on the surface of a 2-sphere we know that $x^2 + y^2 + z^2 = 1$. Therefore there is one linear dependency between these six functions, and we end up with five basis functions in L2.

The simplest way to capture the L2 band is a 3x3 matrix:

$$B_2^{ij} = \omega_i \omega_j - \tfrac{1}{3}\delta_{ij} \qquad\qquad i, j \in \{1, 2, 3\}$$

with a corresponding matrix R_2^{ij} of SH coefficients, where $\omega_{1,2,3} = x, y, z$ and δ_{ij} is one if $i = j$ and zero otherwise. The negative one-third term is required to ensure each function has zero integral over the sphere, so is orthogonal to the constant basis function.

This 3x3 matrix R_2^{ij} is *symmetric*, meaning that $R_2^{ij} = R_2^{ji}$, and *traceless*, meaning that the diagonal elements sum to zero: $R_2^{00} + R_2^{11} + R_2^{33} = 0$. Therefore it does indeed have five degrees of freedom (and an actual implementation would not store nine coefficients!)

Whereas L0 and L1 have relatively simple physical meanings, it is harder to grasp L2 intuitively. The linear L1 band is “antipodal” in the sense that if you add weight in direction $+x$ then it must be balanced by negative weight in the opposite direction $-x$. For the quadratic L2 band, adding weight in the $+x$ direction must also add weight in the $-x$ direction since $x^2 = (-x)^2$. Instead, to ensure a net-zero overall contribution, the balancing negative weight is shared equally in the orthogonal y and z axes.

Although L2 is already a good approximation, in the case of CG lighting this property can have counter-intuitive effects: adding a light may cause “negative light” to appear in some orthogonal direction.

Reconstruction

With the above definitions, the SH approximation to the original function is:

$$R_{sh}(\vec{\boldsymbol{\omega}}) = R_0 + 3\vec{\boldsymbol{R}}_1 \cdot \vec{\boldsymbol{\omega}} + \tfrac{15}{2}\omega_i \omega_j R_2^{ij} + ...$$

This formulation requires only simple rational constants applied in the reconstruction step.

Radiance and Irradiance

As we have hinted already, in our use case the function R is radiance (incoming light at a point), while for shading we need to compute irradiance for a given normal direction.

Assuming the diffuse Lambertian BRDF, irradiance is a hemispherical integral:

$$I(\vec{\boldsymbol{n}}) = \frac{1}{\pi}\int_H \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, R(\vec{\boldsymbol{\omega}})\, d\Omega$$

where $\vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}$ is the geometry term.

Computing the SH approximation we can replace $R(\vec{\boldsymbol{\omega}})$ by $R_{sh}(\vec{\boldsymbol{\omega}})$ and compute the SH coefficients of I independently:

$$I_0 = \frac{1}{\pi}\int_H \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, R_0\, d\Omega = R_0$$

$$\vec{\boldsymbol{I}}_1 = \frac{1}{\pi}\int_H \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, 3\vec{\boldsymbol{R}}_1\cdot\vec{\boldsymbol{\omega}}\, d\Omega = 2\vec{\boldsymbol{R}}_1$$

$$I_2^{ij} = \frac{1}{\pi}\int_H \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, \tfrac{15}{2}\omega_i\omega_j R_2^{ij} = \tfrac{15}{8}R_2^{ij}$$

Putting it all together, we find that the irradiance reconstruction is:

$$I_{sh}(\vec{\boldsymbol{\omega}}) = R_0 + 2\vec{\boldsymbol{R}}_1\cdot\vec{\boldsymbol{\omega}} + \tfrac{15}{8}\omega_i\omega_j R_2^{ij} + ...$$

As expected, the radiance reconstruction constants have been absorbed into the irradiance computation, so we have only one set of constants to apply to convert measured radiance into irradiance.

This formulation makes it easy to see that this irradiance approximation can easily break down in extreme lighting conditions. The vector $\vec{\boldsymbol{R}}_1$ has maximum length R_0, so the coefficient of 2 in the L1 term can cause the absolute value of this term to exceed R_0 – which means in some direction the L1 expansion for irradiance will be negative. Since we know that irradiance is never negative this is clearly an undesirable feature, and causes clear visual artefacts.

For the L1 reconstruction of irradiance, it is possible to do better by introducing a non-linear reconstruction step which takes into account the real-world non-negativity constraint. See the Geomerics CEDEC 2015 talk "Reconstructing Diffuse Lighting from Spherical Harmonic Data" for more details. We are still working on the best formulation to improve L2 irradiance reconstruction in a similar way.

Odd Terms in Irradiance Expansion

Consider the integral for computing the L1 vector of coefficients $\vec{\boldsymbol{R}}_1$:

$$\vec{\boldsymbol{R}}_1 = \frac{1}{4\pi} \oint \vec{\boldsymbol{\omega}}\, R(\vec{\boldsymbol{\omega}})\, d\Omega$$

This looks somewhat similar to the definition of irradiance, and if we dot the whole expression by $\vec{\boldsymbol{n}}$ we get:

$$\vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{R}}_1 = \frac{1}{4\pi} \oint \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, R(\vec{\boldsymbol{\omega}})\, d\Omega = \frac{1}{4}\left(\frac{1}{\pi}\int_{H+} \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, R(\vec{\boldsymbol{\omega}})\, d\Omega + \frac{1}{\pi}\int_{H-} \vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{\omega}}\, R(\vec{\boldsymbol{\omega}})\, d\Omega\right)$$

which yields

$$\vec{\boldsymbol{n}}\cdot\vec{\boldsymbol{R}}_1 = \frac{1}{4}\left(I(\vec{\boldsymbol{n}}) - I(-\vec{\boldsymbol{n}})\right)$$

This is actually quite surprising – what it says is that the L1 band captures the antipodal difference in irradiance *perfectly.* If you know the *exact* irradiance in a direction, and you know the L1 vector, then you can compute the *exact* irradiance in the opposite direction. Or, to put it another way, the only odd terms in the irradiance expansion are the linear terms in the L1 band.

It's not clear how useful this observation is in practice, but it may give us a different direction in which we can refine the existing lighting reconstruction models.

Conclusion

We have described an alternative definition of the Spherical Harmonic series which is easier to work with in the context of CG lighting, not least because it removes as many constant factors of π (or its decimal expansion) as possible.

This in turn makes the physical interpretation of the coefficients easier, which opens up possibilities for improving the final reconstruction. Finding the optimal reconstruction for irradiance in general remains an area of active research.