ZH3: Quadratic Zonal Harmonics

THOMAS ROUGHTON, Activision Publishing, New Zealand
PETER-PIKE SLOAN, Activision Publishing, USA
ARI SILVENNOINEN, Activision Publishing, USA
MICHAL IWANICKI, Activision Publishing, USA
PETER SHIRLEY, Activision Publishing, USA

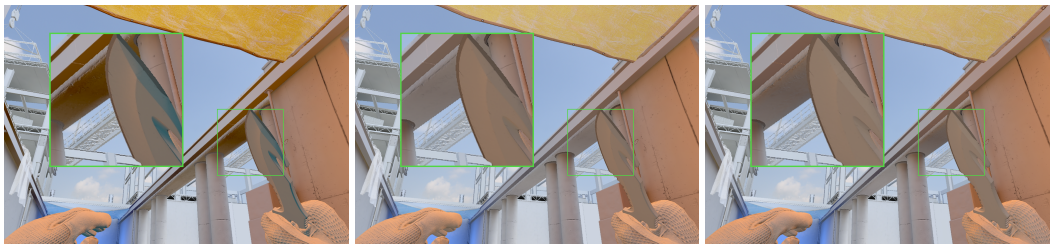


Fig. 1. Indirect diffuse lighting from linear SH (left), hallucinated ZH3 (center, our contribution), and quadratic SH (right) on a production map. Linear SH uses minimal storage but can have significant visual issues; in this example, linear SH has negative lobes and color shifting under the railing and on the held knife, and does not accurately represent the transmissive lighting on the tarpaulin. Quadratic SH is significantly more accurate but also requires over double the storage per sample, making it expensive for spatially dense data. Hallucinated ZH3 resolves the issues of linear SH while using the same storage and minimal extra computation, producing a much closer match to the quadratic SH reference. © Activision Publishing, Inc.

Spherical Harmonics (SH) have been used widely to represent lighting in games and film. While the quadratic (SH3) and higher order spherical harmonics represent irradiance well, they are expensive to store and evaluate, requiring 27 coefficients per sample. Linear SH (SH2), requiring only 12 coefficients, are sometimes used, but they do not represent irradiance signals accurately and can have challenges with negative reconstruction. We introduce a new representation (ZH3) that augments linear SH with just the zonal coefficient of quadratic SH, yielding significant visual improvement with just 15 coefficients, and discuss how solving for a luminance zonal axis can significantly improve reconstruction accuracy and reduce color artifacts. We also discuss how, rather than storing the ZH3 coefficients explicitly, we can hallucinate them from the linear SH, improving reconstruction accuracy over linear SH at minimal extra cost.

CCS Concepts: • **Computing methodologies** → **Rendering**.

Additional Key Words and Phrases: spherical harmonics, global illumination

ACM Reference Format:
Thomas Roughton, Peter-Pike Sloan, Ari Silvennoinen, Michal Iwanicki, and Peter Shirley. 2024. ZH3: Quadratic Zonal Harmonics. In . ACM, New York, NY, USA, Article 11, 15 pages. https://doi.org/10.1145/3651294

1 INTRODUCTION

Interactive applications require efficient, compact representations of spherical functions. Spherical harmonics (SH) are often used to represent radiance, irradiance, and transfer vectors, and provide high reconstruction quality at relatively low storage and evaluation cost. Other representations such

Symposium on Interactive 3D Graphics and Games, May 2024, Philadelphia, USA

as spherical Gaussians [Green et al. 2006; Neubelt and Pettineo 2015; Tsai and Shih 2006], wavelets [Ng et al. 2003], and Ambient Dice [Iwanicki and Sloan 2017] can also provide high quality, but are generally heavier in both storage and evaluation than low-order spherical harmonics, making them impractical for low-end platforms like mobile phones. Lighter spherical radial basis functions have also been used [McTaggart 2004] but have known limitations [Sloan 2008].

Spherical harmonics used for irradiance are generally either of the linear (SH2) or quadratic (SH3) orders. Linear spherical harmonics use only 12 coefficients and four basis functions but can be inaccurate, with the reconstructed lighting being low-frequency and prone to negative lobes, while the quadratic spherical harmonics accurately represent irradiance but require 27 coefficients and the evaluation of nine basis functions.

In this paper, we aim to approach the quality of quadratic SH at minimal extra cost over linear SH. We do this by considering a subset of spherical harmonics: the *zonal harmonics* (ZH). In prior work, zonal harmonics have been used to factor SH for fast rotations [Nowrouzezahrai et al. 2012], prefilter environments maps [Soler et al. 2015], and represent deformable transfer [Sloan et al. 2005]. Our contribution focuses on irradiance reconstruction: specifically, we show that adding the quadratic zonal (ZH3) basis function to linear SH can achieve much of the appearance of quadratic SH. We investigate explicitly solving for and storing the ZH3 coefficient, adding one value per color channel over linear SH. We also show how the ZH3 coefficient can be estimated from the linear SH rather than stored, building on prior work [Joseph 2015] that hallucinates higher frequencies of lighting based on the ratio between the linear and constant bands.

2 SPHERICAL HARMONICS

The real spherical harmonics are a set of orthonormal basis functions defined by

$$Y_l^m(\theta,\phi) = \begin{cases} \sqrt{2}K_l^m \cos(m\phi)P_l^m(\cos\theta) & m > 0 \\ \sqrt{2}K_l^m \sin(|m|\phi)P_l^{|m|}(\cos\theta) & m < 0 \\ K_l^m P_l^m(\cos\theta) & m = 0, \end{cases} \tag{1}$$

where (θ,ϕ) is a vector in spherical coordinates, P_l^m are the associated Legendre polynomials, and K_l^m are the normalization constants

$$K_l^m = \sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}}.$$

The band index l, where l is a non-negative integer, and function index m, where m is an integer in $[-l,l]$ for band l, uniquely identify individual spherical harmonic basis functions. A spherical harmonic of order O (notated here as SHO spherical harmonics) consists of the first O bands and O^2 basis functions; that is, SHO includes all basis functions whose l is between 0 and $O-1$. Spherical harmonics of a certain order can also be identified by their highest polynomial degree; for example, order 2/SH2, or linear, spherical harmonics consists of only linear polynomial basis functions, while order 3/SH3, or quadratic, spherical harmonics include quadratic polynomials.

Since spherical harmonics are an orthonormal basis and least-squares encoding is therefore expressed by projection, a function $f(s)$ approximated as spherical harmonics has basis coefficients given by the projection $f_l^m = \int_{S^2} f(s)\mathrm{Y}_l^m(s)\mathrm{d}s$. Reconstruction is given by $f(s) = \sum f_l^m \mathrm{Y}_l^m(s)$, or, equivalently, $f(s) = f \cdot \mathrm{Y}(s)$. Spherical harmonics are closed under rotation and can accurately represent smooth functions using a small number of bands.

The spherical coordinate form defined in Equation 1 is convenient for symbolic computations and evaluating analytic integrals, but is expensive to evaluate at run-time; in practice, SH are often efficiently evaluated as polynomials of Cartesian coordinates on the unit sphere [Sloan 2013].

2.1 Zonal Harmonics

Zonal harmonics (ZH) are the subset of spherical harmonic functions for which only the zonal ($m = 0$) coefficient for each band is non-zero. Any function that has circular symmetry in z projects into the zonal harmonics, and any spherical radial basis function (SRBF) can therefore be expressed as a zonal harmonic oriented along the basis function's axis.[1]

An arbitrary SH function f and zonal SH function h can be convolved in closed form using the following equation [Sloan et al. 2005]:

$$(f * h)_l^m = \sqrt{\frac{4\pi}{2l+1}} f_l^m h_l^0 = \frac{f_l^m h_l^0}{K_l^0}. \tag{2}$$

ZH coefficients, given by $k_l = h_l^0 / K_l^0$, are per-band scale factors applied when computing the convolution of an SH with a zonal harmonic. The SH expansion of a ZH is given by applying the convolution theorem to a delta function; you evaluate the SH in the direction of the zonal axis and then multiply the ZH coefficients per-band.

Represented in the zonal frame, a zonal harmonic function will have zero for all non-zonal ($m \neq 0$) coefficients; as such, the integral of the product of two ZH functions can always be computed in O operations, rather than the O^2 operations required when using expansion to SH [Dubouchet et al. 2019]. In a coordinate system where the first ZH is aligned with $z = (0, 0, 1)$ and the second with $(\sin\theta, 0, \cos\theta)$ where $\cos\theta$ is the dot product of the two ZH axes, the convolution is given by

$$\int_{S^2} f(s)g(s)\mathrm{d}s = \sum_l K_l^0 \mathrm{Y}_l^0(\theta, 0)\, f_l g_l. \tag{3}$$

2.2 Linear Spherical Harmonics (SH2)

The linear spherical harmonics have only four basis functions: the constant DC term f_0 and a scaled function corresponding to each of the cardinal axes $-y, z, -x$:

$$\mathrm{Y}(x, y, z) = \left[\frac{1}{2\sqrt{\pi}} \quad -\sqrt{\frac{3}{4\pi}}y \quad \sqrt{\frac{3}{4\pi}}z \quad -\sqrt{\frac{3}{4\pi}}x\right]. \tag{4}$$

Storing a coefficient per basis function per color channel, linear SH requires twelve coefficients for RGB colors.

SH2 are always a zonal harmonic when expressed in a coordinate frame oriented along the "optimal linear direction" $(x, y, z) = \frac{(-f_1^1, -f_1^{-1}, f_1^0)}{\|f_1\|}$ [Sloan et al. 2005]. In this coordinate frame, the zonal SH coefficient l_1^0 is given by the length of the L1 ($l = 1$) band $\|f_1\|$, and the l_1^{-1} and l_1^1 coefficients are zero.

[1] Common examples of SRBFs include the normalized cosine lobe for irradiance convolution, hemispheres, and cones.

2.3 Quadratic Spherical Harmonics (SH3)

The quadratic spherical harmonics add an additional five basis functions:

$$\mathrm{Y}_2(x,y,z) = \left[\sqrt{\tfrac{15}{4\pi}}xy \quad -\sqrt{\tfrac{15}{4\pi}}yz \quad \sqrt{\tfrac{5}{16\pi}}(3z^2-1) \quad -\sqrt{\tfrac{15}{4\pi}}xz \quad \sqrt{\tfrac{15}{16\pi}}(x^2-y^2)\right]. \tag{5}$$

Quadratic spherical harmonics represent irradiance with an average error of less than 3% [Ramamoorthi and Hanrahan 2001] and are often used as a reference for compact environment lighting.

3 ZONAL QUADRATIC SPHERICAL HARMONICS (ZH3)

Extending from linear to quadratic spherical harmonics requires adding an additional five coefficients per color channel, more than doubling the storage and evaluation cost; this can be a particular issue for spatially dense data. We introduce ZH3, which extends linear SH by adding just the quadratic zonal coefficient expressed in the linear SH coordinate frame. ZH3 can exactly represent zonal SH3 signals (such as the SH3 representation of a point or sphere light) and significantly improves reconstruction on a range of other signals (Figures 2 and 3). When compared with quadratic SH, ZH3 often delivers a very similar appearance while requiring four fewer coefficients per channel; furthermore, using a ratio-based encoding scheme (Appendix A.1), RGB ZH3 requires only 16 bytes of storage compared with 28 for quadratic SH.

3.1 Solving for ZH3

The quadratic ZH coefficient k_2 in a given zonal frame can be found through a least-squares fit to the quadratic ($l = 2$) band. Given a direction $\mathbf{d}_{lin}$ (taken to be the optimal linear direction from the linear spherical harmonic), the five quadratic basis functions evaluated in that direction $q = \mathrm{Y}_2(\mathbf{d}_{lin})$, and the target basis coefficients f_2, the squared error E is given by

$$E = (k_2 q - f_2) \cdot (k_2 q - f_2) \tag{6}$$

$$= (k_2)^2 (q \cdot q) - 2k_2 (q \cdot f_2) + f_2 \cdot f_2. \tag{7}$$

Differentiating with respect to k_2, you get

$$\frac{\mathrm{d}E}{\mathrm{d}k_2} = 2k_2 (q \cdot q) - 2q \cdot f_2, \tag{8}$$

where $q \cdot q = \frac{5}{4\pi}$. Solving for zero results in

$$k_2 = \frac{4\pi}{5} q \cdot f_2. \tag{9}$$

Equivalently, the SH l_2^0 coefficient in this coordinate frame is given by

$$l_2^0 = K_2^0 k_2 = \sqrt{\frac{4\pi}{5}} q \cdot f_2. \tag{10}$$

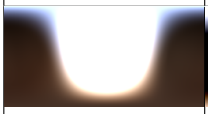
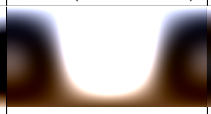
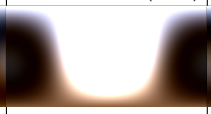
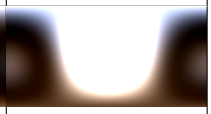
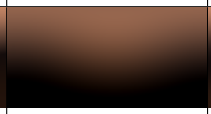
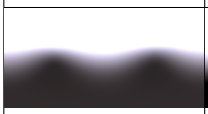
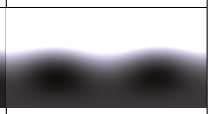
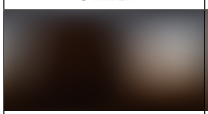
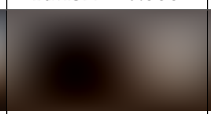
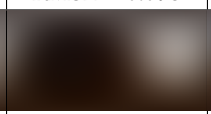
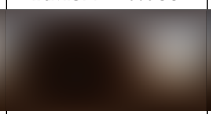
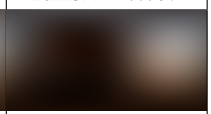
Reference	SH2	ZH3 (Lum. Axis)	Hallucinate (C.F.)	SH3
				
Ennis	RMSE = 0.556	RMSE = 0.110	RMSE = 0.159	RMSE = 0.095
				
Grace	RMSE = 0.0896	RMSE = 0.0140	RMSE = 0.0420	RMSE = 0.0132
				
Pisa	RMSE = 0.0244	RMSE = 0.0204	RMSE = 0.0214	RMSE = 0.0051
				
Uffizi	RMSE = 0.333	RMSE = 0.078	RMSE = 0.085	RMSE = 0.039
				
Wells	RMSE = 0.0552	RMSE = 0.0344	RMSE = 0.0343	RMSE = 0.0076

Fig. 2. Comparison of irradiance reconstruction for SH2, ZH3 (using a solved-for shared luminance axis), curve-fit ZH3 hallucinate, and SH3 on a range of environment maps [Debevec 2001; Vogl 2010].

3.2 Luminance Blend

Solving directly for the zonal coefficient along each color channel's zonal axis can lead to undesirable artifacts and color shifts if the zonal axes are not aligned (Figure 4). These artifacts can be avoided if we loosen the zonal constraint and introduce a new, shared axis, given by the optimal linear direction of the luminance spherical harmonic (a weighted average of the RGB SH). The new axis is used exclusively for the zonal L2 (quadratic) term, and we solve for and evaluate the L2 zonal coefficient for each channel along it. This carries the additional benefit of reducing the runtime cost of evaluation, since only one normalized axis need be computed rather than three. In some cases, this may also lower the error, since the optimal axis for the zonal L2 basis is not necessarily aligned with the L1 band's axis; the shared luminance axis may be a better fit.

3.3 Solving for the ZH3 Axis

A key property of ZH3 is that the zonal axis is implicitly determined from the linear spherical harmonic. However, the zonal axis given by the least-squares linear spherical harmonic may not result in the lowest error representation of the overall lighting, and it is often beneficial to instead solve for the axis that results in the lowest reconstruction error when considering both the L1/linear and L2/quadratic zonal terms (Figure 5). Put another way, if there is a direction that is well represented by an L2 zonal lobe, and the contribution of the L2 band is significant, then

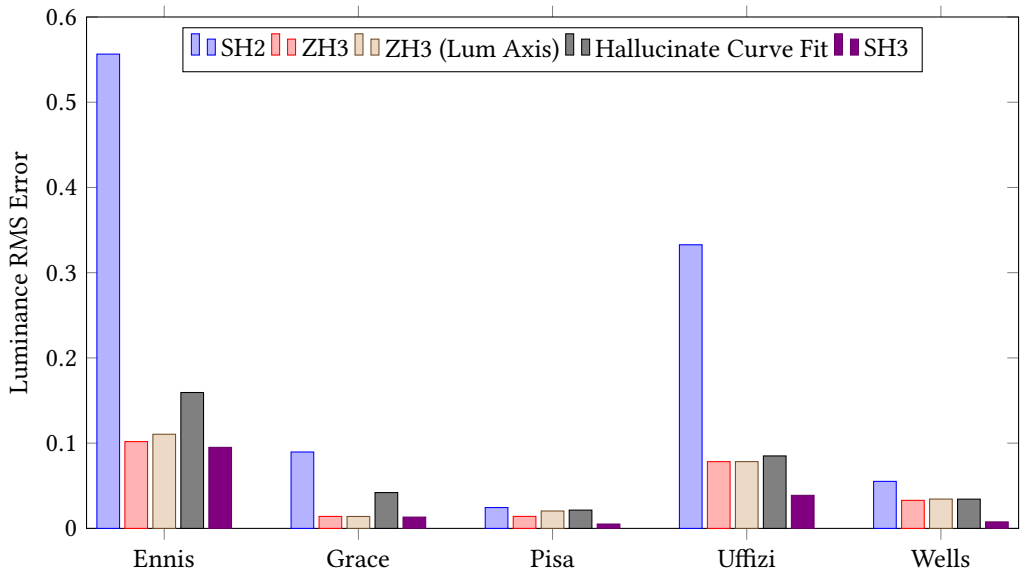


Fig. 3. RMS error for irradiance reconstruction on a range of environment maps [Debevec 2001; Vogl 2010].

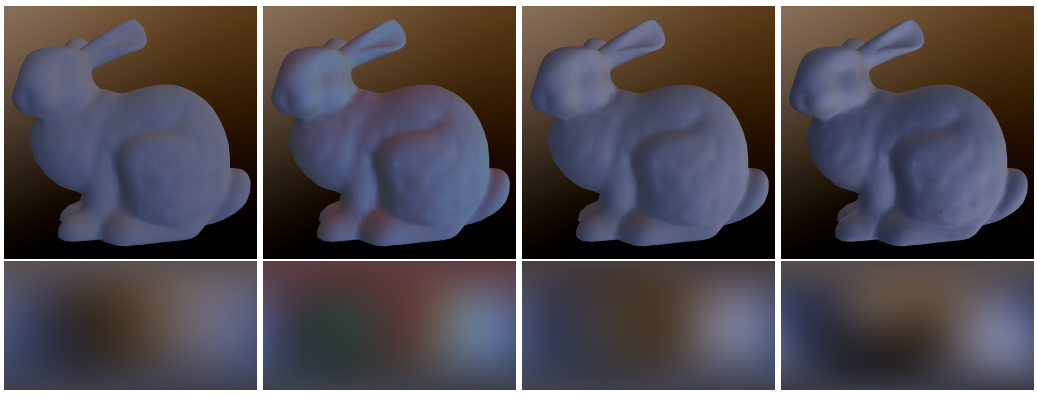
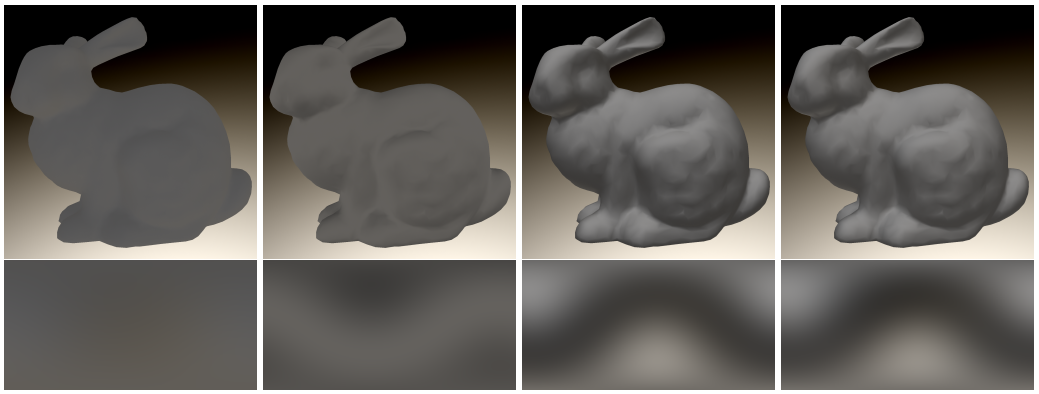


Fig. 4. Irradiance reconstruction from SH2, ZH3 without and with a shared luminance axis, and SH3, on a light probe from a production map. The shared luminance axis resolves color artifacts in the per-channel axis version and decreases the error in the red and green channels, resulting in a closer appearance to SH3.

moving the L1 axis toward the most favored direction for the L2 band can result in lower error overall, even though it will increase the error for the L1 band.

Following the method of [Sloan et al. 2005], we perform a BFGS-based search for the optimal direction. The solve parameters are the per-channel L1 SH coefficients, with derivatives taken with respect to a spherical harmonic error function E:

$$E = \|f_1 - t_1\|^2 + \|\frac{4\pi}{5}\left(\mathrm{Y}_2(\mathrm{A}(f)) \cdot t_2\right)\mathrm{Y}_2(\mathrm{A}(f)) - t_2\|^2, \tag{11}$$



0.347, 0.343, 0.337	0.323, 0.320, 0.317	0.061, 0.060, 0.060	*RGB RMSE*
(a) SH2	(b) ZH3, Opt. Lin. Axis	(c) ZH3, Shared Axis Solve	(d) SH3 (Reference)

Fig. 5. Irradiance reconstruction from SH2, ZH3 with an axis given by the unmodified linear SH, ZH3 with a solved-for zonal axis, and SH3, on a light probe from a production map. Solving for the zonal axis produces a major improvement in appearance and accuracy in this case; since the contribution of the L1 band is relatively small, refitting it based on the L2 band is highly beneficial.

where f is the linear SH being solved for, $\mathrm{A}(f)$ is the axis from the linear spherical harmonic (e.g. the optimal linear direction), t is the target irradiance SH, and $\frac{4\pi}{5}\left(\mathrm{Y}_2(\mathrm{Axis}(f)) \cdot t_2\right)$ is the ZH3 coefficient for that axis. In other words, the error is the squared difference between the target and the SH expansion of the ZH3, where the k_2 coefficient is given by a least-squares fit to t (Equation 9) along the axis given by the linear SH f. The L0 coefficient is invariant and is given by $f_0 = t_0$.

For separate per-channel axis solves, it is sufficient to solve only for the axis (rather than the linear SH coefficients) and find the least-squares zonal coefficients for both L1 (by simply projecting onto the axis) and L2 (from Equation 9) from that. However, when solving for a shared luminance axis, the zonal L1 coefficients also determine the luminance axis for the L2 band; a greater magnitude for the L1 band for a channel will weight the luminance axis towards that channel. Given that, we choose to parameterize the error over the full L1 band coefficients, implicitly determining the luminance axis, rather than over the shared axis and per-channel zonal L1 coefficients. We provide a full expansion of the derivatives and axis function in Appendix C.

3.4 Hallucinating the ZH3 Coefficient

Rather than storing the ZH3 coefficient, an alternative approach is to instead *hallucinate* it from the linear spherical harmonic, effectively modifying the reconstruction algorithm (Listing 1). Modifying irradiance reconstruction from linear SH is not a novel idea; linear SH are prone to negative lobes, color shifting, and fail to accurately represent irradiance from directional or point lights, and so prior work [Joseph 2015] introduced the Geomerics algorithm to address these issues. The Geomerics algorithm always preserves the first moment (DC) of lighting and effectively hallucinates higher-order terms from the ratio of the linear SH coefficients to DC; however, it doesn't preserve the second moment (linear SH) and can be overly bright and flat when the ratio is low [Sloan and Silvennoinen 2020] (Figure 6). If, instead, we hallucinate a ZH3 coefficient, we preserve the first two moments of lighting,[2] and in practice can match the reference more accurately.

[2]The idea of preserving multiple moments has been used in applied physics [Wyman et al. 1989] to approximate scattering parameters and presented in the graphics literature [Zhao et al. 2014].

0.133, 0.140, 0.182	0.047, 0.058, 0.101	*RGB RMSE*
(a) Geomerics	(b) ZH3 Curve-Fit Hallucinate	(c) Reference

Fig. 6. Geomerics tends to over-brighten the lighting, whereas ZH3 hallucinate more accurately reconstructs the overall tonality of the reference. The lighting environment is an SH3 probe taken from a production map.

Hallucinating higher-order terms can be easily motivated by real-world data. Linear SH is the truncation of a signal, and in almost all cases the average truncated data is not best represented by zeroes; that is, the average higher-order coefficients for real-world lighting are non-zero when the linear band is non-zero (Figure 7). Our approach is to predict the ZH3 coefficient from the ratio of linear to DC; specifically, given a zonal axis from the linear SH, a DC/f_0 coefficient, and the f_1^0 coefficient for the zonal term along that axis (i.e. the length of the L1 band vector), we consider how best to hallucinate a higher-order term or terms from that $\frac{f_1^0}{f_0}$ ratio.

3.4.1 Ambient & Directional. The most trivial model is to consider the L1 SH as the result of an ambient and directional light, where the direction is given by the optimal linear direction from the L1 SH. The ratio of zonal L1 term to DC for a directional light is given by $\sqrt{3}$; therefore, the intensities of the ambient and directional lights are given by

$$I_{amb} = 2\sqrt{\pi}\left(f_0 - \frac{\|f_1\|}{\sqrt{3}}\right), \tag{12}$$

$$I_{dir} = \sqrt{\frac{4\pi}{3}}\|f_1\|. \tag{13}$$

If this is treated as an analytic ambient and directional light, this carries the downside of there being no lighting variation per color channel in the hemisphere opposite the directional light, since that lighting is entirely provided by the ambient term. However, if we restrict ourselves to inferring the f_2^0 coefficient in the zonal frame from the directional light, we gain lighting variation at the cost of possible ringing. The hallucinated coefficient in the zonal frame is given by

$$f_2^0 = \frac{\sqrt{15}}{3}\|f_1\|. \tag{14}$$

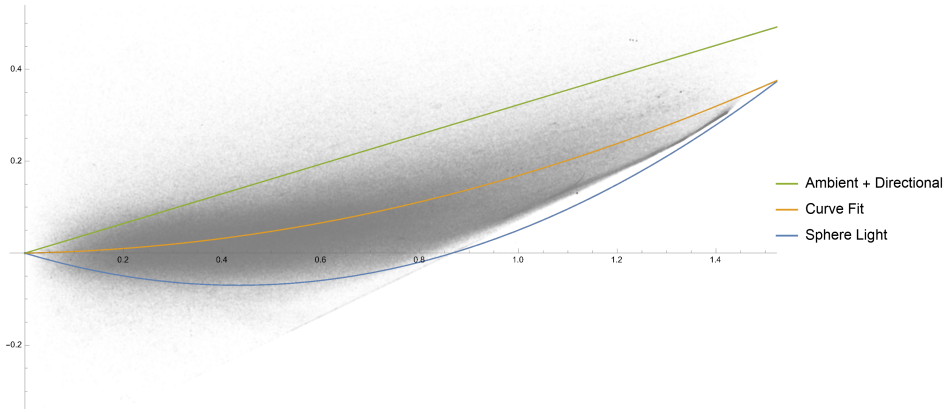


Fig. 7. Ratio of L1 to L0 lighting coefficients (horizontal axis) against the ZH3 irradiance coefficient (vertical axis). The gray points are real-world values from a set of production maps; the curves are three discussed hallucinate models. There is a clear overall trend, indicating that the average ZH3 coefficient is non-zero and that the ratio is a reasonable means to estimate it.

Adding this hallucinated f_2^0 coefficient results in a decrease in error of over 17% on average on a set of production maps compared to linear SH. Since our reference is quadratic SH, it is difficult to determine whether an analytic ambient and directional light model would result in lower error than the SH representation; however, this is an interesting avenue for future work.

3.4.2 Spherical Lights. Spherical lights are ZH functions with closed-form ZH coefficients. They can be represented using trigonometric functions [Sloan 2008] of σ, the angle of the opening, or as a function of radius r of a sphere and distance to the center d [Yuan et al. 2012]. These equations work for $\sigma \leq \frac{\pi}{2}$ and represent constant illumination from a cone that subtends a given angle:

$$\begin{aligned} f_0 &= \sqrt{\pi}(1 - \cos\sigma), \\ f_1^0 &= \tfrac{\sqrt{3\pi}}{2} \sin^2\sigma, \\ f_2^0 &= \tfrac{\sqrt{5\pi}}{2} \cos\sigma \sin^2\sigma. \end{aligned} \tag{15}$$

The ZH coefficients (including per-band scaling coefficients) are

$$\begin{aligned} k_0 &= 2\pi(1 - \cos\sigma), \\ k_1 &= \pi \sin\sigma^2, \\ k_2 &= \pi \cos\sigma \sin\sigma^2. \end{aligned} \tag{16}$$

The ratio $p = \frac{f_1^0}{f_0}$ of the coefficients for a spherical light in a zonal coordinate frame is

$$p = \frac{\frac{\sqrt{3\pi}}{2} \sin^2\sigma}{\sqrt{\pi}(1 - \cos\sigma)} = \frac{\sqrt{3}}{2} \frac{\sin^2\sigma}{(1 - \cos\sigma)}. \tag{17}$$

Using the trigonometric identity $\sin^2\sigma + \cos^2\sigma = 1$ and the fact that $1 - x^2 = (1 - x)(1 + x)$, the ratio simplifies to

$$p = \frac{\sqrt{3}}{2}(1 + \cos\sigma), \tag{18}$$

yielding an expression for $\cos\sigma$:

$$\cos\sigma = \frac{2\sqrt{3}}{3} p - 1. \tag{19}$$

This expression can either be used to directly evaluate the irradiance from an analytic spherical area light [Snyder 1996] or to hallucinate the zonal L2 coefficient. In practice, adding this hallucinated f_2^0 coefficient results in almost identical error (a less than 0.1% decrease) on our test maps compared to linear SH, although the appearance is subjectively improved due to the added detail.

3.4.3 Curve Fit. It can be derived that the hallucinated ZH3 coefficient from the sphere area light is simply a quadratic in the ratio $p = \frac{f_1^0}{f_0}$

$$f_2^0 = \left(\frac{2\sqrt{5}}{3}p^2 - \sqrt{\frac{5}{3}}p\right) f_0, \tag{20}$$

where f_2^0 is the computed *radiance* zonal L2 coefficient for the zonal axis.

While the sphere area light model isn't a good fit for our real-world data, it does raise whether there are coefficients for this quadratic that better fit actual lighting data. We compute a quadratic curve fit to the radiance coefficients for our input lighting environments:

$$f_2^0 = \left(0.6p^2 + 0.08p\right) f_0. \tag{21}$$

Using these parameters to hallucinate the ZH3 coefficient reduced error by over 23% on average on a range of test maps compared to linear SH.

3.4.4 Luminance Zonal Axis. As with a stored ZH3 coefficient, color fringing may occur if the zonal axes diverge, although this is more rare than with explicit ZH3. The strategy of using a shared luminance axis (Section 3.2) is also applicable for hallucinated zonal coefficients; in that case, the ratios are computed using the projection of the L1 coefficients onto the new axis (equivalent to computing the f_1^0 coefficient for a coordinate system aligned along that axis) (Listing 1). Using the luminance axis also yields a reduction of error of over 23% on average compared to linear SH, although the reduction is slightly less (by 0.2%) than for per-channel axes.

Listing 1. Irradiance reconstruction from linear SH using curve-fit ZH3 coefficients.

```
float3 SHHallucinateZH3Irradiance(float3 sh[4], float3 direction) {
    // Use the zonal axis from the luminance SH.
    const float3 lumCoeffs = float3(0.2126f, 0.7152f, 0.0722f); // sRGB luminance.
    float3 zonalAxis = normalize(float3(-dot(sh[3], lumCoeffs), -dot(sh[1], lumCoeffs), dot(sh[2], lumCoeffs)));

    float3 ratio = 0.0;
    ratio.r = abs(dot(float3(-sh[3].r, -sh[1].r, sh[2].r), zonalAxis));
    ratio.g = abs(dot(float3(-sh[3].g, -sh[1].g, sh[2].g), zonalAxis));
    ratio.b = abs(dot(float3(-sh[3].b, -sh[1].b, sh[2].b), zonalAxis));
    ratio /= sh[0];
    float3 zonalL2Coeff = sh[0] * (0.08f * ratio + 0.6f * ratio * ratio); // Curve-fit; Section 3.4.3

    float fZ = dot(zonalAxis, direction);
    float zhDir = sqrt(5.0f / (16.0f * PI)) * (3.0f * fZ * fZ - 1.0f);

    // Convolve sh with the normalized cosine kernel (multiply the L1 band by the zonal scale 2/3), then dot with
    // SH(direction) for linear SH (Equation 5).
    float3 result = SHLinearEvaluateIrradiance(sh, direction);

    // Add irradiance from the ZH3 term. zonalL2Coeff is the ZH3 coefficient for a radiance signal, so we need to
    // multiply by 1/4 (the L2 zonal scale for a normalized clamped cosine kernel) to evaluate irradiance.
    result += 0.25f * zonalL2Coeff * zhDir;
    return result;
}
```

4 DISCUSSION

Irradiance reconstruction using hallucinated ZH3 coefficients has been shipped in multiple commercial video games, with the motivation initially being to improve lighting through translucent surfaces over linear SH. While the shipped implementation uses per-channel zonal axes and a sphere-light-derived model, we intend to switch to using a luminance axis (Section 3.4.4) and the curve fit (Section 3.4.3) in the future. We have not shipped a stored ZH3 representation but are investigating doing so: we currently use full quadratic SH as a world-space lighting representation, where each SH probe uses 28 bytes of memory (Appendix A.1), and switching to stored ZH3 would reduce our memory usage by almost half to 16 bytes per probe, which is particularly attractive for mobile platforms.

It is worth considering the trade-offs between stored and hallucinated ZH3 coefficients. When comparing the quantized representation from a performance standpoint, there's very little to separate them; the storage requirements of the ZH3 coefficients are minimal at only three bytes, fitting naturally into a 16 byte alignment, and, compared to decoding the quantized coefficients, hallucinating adds a negligible 12 instructions/18 cycles on the AMD RDNA™ GPU architecture [Advanced Micro Devices, Inc. 2023]. Stored ZH3 coefficients have expectedly lower error; on our test sets, solved-for ZH3 coefficients have a 27% reduction in error compared to curve-fit hallucinated coefficients (44% compared to linear SH), making them the obvious choice for point-sampled representations.

The situation is more complicated when considering interpolation properties. The ZH3 coefficient is stored for the linear axis, and linear interpolation of the linear SH coefficients results in nonlinear interpolation of the ZH3 coefficient. When interpolating coefficients of a non-linear model, reconstruction artifacts can result if the data is not massaged [Iwanicki 2013], the parameterization changed to filter linear quantities [Sloan and Silvennoinen 2018] or some of the non-linear parameters fixed [Neubelt and Pettineo 2015]; thus, although ZH3 is mostly well behaved if the zonal axis changes slowly, for sparse data ZH3 interpolation should be done by first expanding to quadratic SH and then blending the results. This either carries a high per-pixel cost (if each probe is fetched, expanded, and blended per-pixel) or more than doubles the memory and bandwidth requirements relative to linear SH (if the ZH3 are decoded at runtime to quadratic SH stored in hardware-interpolable textures). By contrast, if linear SH is interpolated at runtime and the ZH3 term is hallucinated from the ratio of linear to DC, interpolation of perpendicular linear SH will naturally reduce that ratio, mimicking the results of expansion to SH3.[3] While direct performance comparisons are difficult,[4] our heuristic is that hallucinated coefficients should be used where linear SH would otherwise be used (improving reconstruction quality at low cost), and that stored coefficients may be considered as a smaller on-disk representation where quadratic SH would otherwise be used.

Rather than ZH3, we initially considered solving for a mixture coefficient between different lighting models, or coordinates in the latent space of a neural network, but simply solving for or hallucinating the quadratic ZH coefficient has a significant benefit: it simplifies the shaders while still preserving all of the moments we care about. Storing linear SH and reconstructing the ZH axis uses fewer coefficients and keeps a zonal (SRBF) structure.

[3]The least-squares fit ZH3 coefficient for aligned linear axes can be exactly linearly blended, while for the average of equal-scale perpendicular linear SH the ZH3 coefficient should be scaled by $\frac{1}{4}$; this means linear interpolation of the ZH3 coefficient will overshoot for perpendicular linear SH. Interpolation is further complicated by antipodal linear SH undergoing destructive interference where antipodal ZH3 constructively interfere.

[4]The performance impact depends on the GPU workload, but as a very rough estimate we see a 0.1ms performance difference on a 4K opaque pass between hallucinated ZH3 and interpolated SH3 (from stored ZH3) on an NVIDIA RTX 2080 Ti.

While using a shared luminance axis for the zonal term resolves false color shifting artifacts and improves on the visual results, the motivation for *why* this is the case is not entirely clear. One intuition is that when using the luminance axis we are treating the zonal axis as a colored light, and so any color shifts will be correlated between color channels. Another is that, when the linear axes diverge (e.g. in the case of directional lights oriented along perpendicular axes), the least-squares fit for the ZH3 coefficient will generally be reduced, meaning the projection onto the luminance axis acts as some form of windowing. Investigating this more rigorously is an interesting avenue for future work.

5 CONCLUSIONS AND FUTURE WORK

We introduced a novel ZH3 format that bridges the gap between linear and quadratic SH, significantly improving visual quality over linear SH with minimal extra storage or computation. We discussed how to solve for ZH3, including how the linear SH parameters can be adjusted to form a better zonal axis, and showed how using a shared luminance axis can minimize color artifacts. We also used ZH3 to hallucinate higher frequencies from linear SH, improving irradiance reconstruction compared to prior work in both accuracy and appearance.

For future work, further exploring ZH3 interpolation and the behavior of the shared luminance axis are interesting directions. Additionally, our choice of the zonal quadratic SH basis function for the zonal axis is not necessarily optimal; it would be worth investigating other basis functions along that axis. Finally, the reasoning for hallucinating the ZH3 coefficient could equally be applied to higher-order terms and other basis functions; even though higher-order SH terms have increasingly minimal contribution to the irradiance, a more thorough investigation of this may be worthwhile.

ACKNOWLEDGMENTS

Our technical contributions are shown in the best possible light due to the hard work of an incredibly talented group of artists, lighters, and engineers. While there are too many to mention here, We would like to particularly thank Michal Drobot, Michael Vance, and Jennifer Velazquez for their support. Thank you also to the reviewers for their constructive feedback and diligent error-spotting.

This paper is dedicated to the memory of Graham Madarasz who contributed to an early implementation of ZH3.

REFERENCES

Advanced Micro Devices, Inc. 2023. "RDNA3" Instruction Set Architecture Reference Guide. (Feb 2023). https://www.amd.com/content/dam/amd/en/documents/radeon-tech-docs/instruction-set-architectures/rdna3-shader-instruction-set-architecture-feb-2023_0.pdf

Paul Debevec. 2001. Light Probe Image Gallery. In *Proceedings of SIGGRAPH*, Vol. 98. https://vgl.ict.usc.edu/Data/HighResProbes/

Adrien Dubouchet, Peter-Pike Sloan, Wojciech Jarosz, and Derek Nowrouzezahrai. 2019. Impulse Responses for Precomputing Light from Volumetric Media. In *Eurographics Symposium on Rendering - DL-only and Industry Track*, Tamy Boubekeur and Pradeep Sen (Eds.). The Eurographics Association.

Paul Green, Jan Kautz, Wojciech Matusik, and Frédo Durand. 2006. View-Dependent Precomputed Light Transport Using Nonlinear Gaussian Function Approximations. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games* (Redwood City, California) *(I3D '06)*. Association for Computing Machinery, New York, NY, USA, 7–14.

Michał Iwanicki. 2013. Lighting Technology of the Last of Us. In *ACM SIGGRAPH 2013 Talks*.

Michał Iwanicki and Peter-Pike Sloan. 2017. Ambient Dice. In *"Eurographics Symposium on Rendering - Experimental Ideas & Implementations"*, Matthias Zwicker and Pedro Sander (Eds.). The Eurographics Association.

William Joseph. 2015. Reconstructing diffuse lighting from spherical harmonic data. In *CEDEC*. https://cedil.cesa.or.jp/cedil_sessions/view/1329

Gary McTaggart. 2004. Half-Life 2 source shading. In *Game Developers Conference*.

David Neubelt and Matt Pettineo. 2015. Advanced Lighting R&D at Ready At Dawn Studios. In *SIGGRAPH 2015 Course: Physically Based Shading in Theory and Practice*.

Ren Ng, Ravi Ramamoorthi, and Pat Hanrahan. 2003. All-Frequency Shadows Using Non-Linear Wavelet Lighting Approximation. *ACM Trans. Graph.* 22, 3 (July 2003), 376–381.

Derek Nowrouzezahrai, Patricio D. Simari, and Eugene Fiume. 2012. Sparse zonal harmonic factorization for efficient SH rotation. *ACM Trans. Graph.* 31, 3 (2012), 23:1–23:9.

Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *SIGGRAPH 2001 Conference Proceedings, August 12–17, 2001, Los Angeles, CA*, ACM (Ed.). ACM Press, pub-ACM:adr, 497–500.

Peter-Pike Sloan. 2008. Stupid Spherical Harmonics (SH) Tricks. In *Game Developers Conference*.

Peter-Pike Sloan. 2013. Efficient Spherical Harmonic Evaluation. *Journal of Computer Graphics Techniques (JCGT)* 2, 2 (8 September 2013).

Peter-Pike Sloan. 2017. Deringing Spherical Harmonics. In *SIGGRAPH Asia 2017 Technical Briefs*.

Peter-Pike Sloan, Ben Luna, and John Snyder. 2005. Local, Deformable Precomputed Radiance Transfer. *ACM Trans. Graph.* 24, 3 (July 2005), 9 pages.

Peter-Pike Sloan and Ari Silvennoinen. 2018. Directional Lightmap Encoding Insights. In *SIGGRAPH Asia 2018 Technical Briefs*.

Peter-Pike Sloan and Ari Silvennoinen. 2020. Precomputed Lighting Advances in Call of Duty: Modern Warfare. In *SIGGRAPH Course: Advances in Real-Time Rendering in Games*.

John Snyder. 1996. *Area Light Sources for Real-Time Graphics.* Technical Report MSR-TR-96-11. Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/1996/03/arealights.pdf

Cyril Soler, Mahdi M. Bagher, and Derek Nowrouzezahrai. 2015. Efficient and Accurate Spherical Kernel Integrals Using Isotropic Decomposition. *ACM Trans. Graph.* 34, 5 (2015), 161:1–161:14.

Yu-Ting Tsai and Zen-Chung Shih. 2006. All-Frequency Precomputed Radiance Transfer Using Spherical Radial Basis Functions and Clustered Tensor Approximation. *ACM Trans. Graph.* 25, 3 (July 2006), 967–976.

Bernhard Vogl. 2010. Light probes. https://dativ.at/lightprobes/.

Tyler Wiederien and Peter-Pike Sloan. 2022. Tighter Spherical Harmonic Quantization Bound. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games - Posters*.

Douglas R Wyman, Michael S Patterson, and Brian C Wilson. 1989. Similarity relations for anisotropic scattering in Monte Carlo simulations of deeply penetrating neutral particles. *J. Comput. Phys.* 81, 1 (1989), 137–150.

Hong Yuan, Derek Nowrouzezahrai, and Peter-Pike Sloan. 2012. Irradiance Rigs. *Journal of Graphics Tools* 16, 1 (2012).

Shuang Zhao, Ravi Ramamoorthi, and Kavita Bala. 2014. High-order Similarity Relations in Radiative Transfer. *ACM Trans. Graph.* 33, 4, Article 104 (July 2014), 12 pages.

A SH COEFFICIENT BOUNDS FOR POSITIVE FUNCTIONS

Many signals that are dealt with in graphics (such as radiance, irradiance, and visibility) come from strictly positive inputs. Monte-Carlo integration with projection into spherical harmonics involves summing delta functions (the value in some sampled direction) with strictly positive weights; this generalises to most integrals of positive functions. For spherical harmonics resulting from the projection of a positive function, an upper bound on the ratio of the higher order coefficients divided by the DC term f_0 can be derived by looking at the coefficient ratios of a SH delta function [Wiederien and Sloan 2022].[5] The DC term f_0 only grows when scaled by positive inputs, while the higher frequencies can have either constructive or destructive interference. Assuming only constructive interference, there is an upper bound ratio of $\frac{\sqrt{2L+1}}{f_0}$ for the length of band L.[6] The bound only exists because DC is a constant strictly positive basis function and so only has perfectly constructive interference; projecting a negative value results in destructive interference, resulting in no general upper bound.

Tighter bounds also exist for individual basis functions, and can be derived by looking at the SH projection of a delta function. For example, $f_2^0 = \sqrt{\frac{5}{16\pi}}(3z^2 - 1)$ has a maximum value of $\sqrt{\frac{5}{4\pi}}$ when $z = 1$ and a minimum value of $-\sqrt{\frac{5}{16\pi}}$ when $z = 0$; divided by the DC basis function $Y_0 = \frac{1}{2\sqrt{\pi}}$,

[5]Because SH are closed under rotation, a rotation will never cause this ratio to change.

[6]This same analysis can be done using the Fourier basis on the circle.

the bounds are $[-\frac{\sqrt{5}}{2}, \sqrt{5}]$. Using a similar analysis, it can be shown that the bounds are $[-\frac{\sqrt{15}}{2}, \frac{\sqrt{15}}{2}]$ for all other basis functions in the L2 band.

A.1 Encoding and quantization

This bound can be used to reduce the required storage for lighting signals. The DC coefficient is stored in a high dynamic range (HDR) texture (11f/11f/10f, which can only encode positive values, is a good candidate), and the values for higher-order bands are divided by the DC coefficient and remapped according to the lower and upper bounds, yielding a $[-1, 1]$ signal that can be stored in low dynamic range (LDR) textures. If the input signal is known to represent irradiance, meaning the L1 band has been scaled by the zonal irradiance factor $\frac{2}{3}$ and the L2 band by $\frac{1}{4}$, these scale factors should be included in the upper bounds. For lighting that is not going to be interpolated, using a square root to encode the magnitude of the value and simply squaring after decode and preserving the sign reduced the mean absolute error by half on an irradiance volume over a shipping level.

It is worth reasoning about what interpolation does to this non-linearly encoded signal, noting that DC itself, the average value over the sphere, is interpolated linearly and is therefore exact. If DC is constant between texels, there is no interpolation error. If the function is a zonal harmonic (such as the projection of a light source where the direction to the light is not changing) the error will be minor, as resulting from the non-uniform scaling of the vector. If both the direction and DC are changing, linear interpolation would change directions slower compared to the normalized result, but the actual arc traced out would be the same. In practice we have not noticed any visual artifacts when switching to this encoding.

B DERINGING ZH3

We want to guarantee non-negative irradiance reconstruction with ZH3 coefficients. For this to hold, we require that for $z \in [-1, 1]$ in the zonal frame,

$$\frac{1}{2\sqrt{\pi}} f_0 + \frac{2}{3}\sqrt{\frac{3}{4\pi}} f_1^0 z + \frac{1}{4}\sqrt{\frac{5}{16\pi}} f_2^0 (3z^2 - 1) \geq 0, \tag{22}$$

or, more simply,

$$1 + az + b(3z^2 - 1) \geq 0, \tag{23}$$

where $a = \frac{2f_1^0}{\sqrt{3}f_0}$ and $b = \frac{\sqrt{5}f_2^0}{8f_0}$. The bounds on a and b reduce to

$$a \leq \sqrt{3}, \max\left(\frac{az - 1}{3z^2 - 1}\right) \leq b \leq \frac{\sqrt{5}}{4} \text{ for } z \in (\frac{1}{\sqrt{3}}, 1]. \tag{24}$$

If $a > \sqrt{3}$ there are no values of b that will guarantee non-negativity across the domain, since a large enough value for b will introduce negative values around $z = 0$ while too small will not compensate for the negative values from the L1 lobe at $z = 1$.

The maximum value of $\frac{az-1}{3z^2-1}$ in $(\frac{1}{\sqrt{3}}, 1]$ is given by $\frac{a-1}{2}$, since for our range the function maximum is always at $z = 1$; therefore, we have

$$\frac{a - 1}{2} \leq b \leq \frac{\sqrt{5}}{4}. \tag{25}$$

Equivalently, if we substitute back in for our SH coefficients, we have

$$\frac{2\sqrt{15}}{3} f_1^0 - \frac{4}{\sqrt{5}} f_0 \leq f_2^0 \leq 2f_0. \tag{26}$$

f_2^0 and f_1^0 are here radiance coefficients; for irradiance reconstruction, the bounds on f_2^0 are multiplied by the zonal cosine convolution scale $\frac{1}{4}$.

This analysis only applies when the zonal axes are aligned for both linear and ZH3; for diverging axes more general windowing approaches [Sloan 2017] can be used.

C DERIVATIVES FOR GRADIENT-BASED ZH3 SOLVES

The error function derivative is required in the use of BFGS or related gradient-based search. The analytic derivatives of the ZH3 error function E defined in Equation 11 are as follows:

$$\frac{\mathrm{d}E}{\mathrm{d}f} = 2\Bigg((f - t_1) + \left(\frac{4\pi}{5}\left(t_2^T \mathrm{Y}_2(\mathrm{A}(f))\right)\mathrm{Y}_2(\mathrm{A}(f)) - t_2\right)^T \left(\frac{4\pi}{5}\left(\mathrm{Y}_2(\mathrm{A}(f))t_2^T + t_2^T\mathrm{Y}_2(\mathrm{A}(f))\right)\mathbf{J}_{\mathrm{Y}_2}\mathbf{J}_{\mathrm{A}}\right)\Bigg), \tag{27}$$

where $\mathbf{J}_{\mathrm{A}}$ is the Jacobian matrix of the axis function $\mathrm{A}(f)$ and $\mathbf{J}_{\mathrm{Y}_2}$ is the Jacobian matrix of Y_2. $\mathrm{A}(f)$ must return a unit vector.

The Jacobian matrix of Y_2 is given by:

$$\mathbf{J}_{Y_2}^T = \begin{bmatrix} \frac{\delta Y_2}{\delta x} \\ \frac{\delta Y_2}{\delta y} \\ \frac{\delta Y_2}{\delta z} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{15}{4\pi}}y & 0 & 0 & -\sqrt{\frac{15}{4\pi}}z & \sqrt{\frac{15}{4\pi}}x \\ \sqrt{\frac{15}{4\pi}}x & -\sqrt{\frac{15}{4\pi}}z & 0 & 0 & -\sqrt{\frac{15}{4\pi}}y \\ 0 & -\sqrt{\frac{15}{4\pi}}y & \frac{3\sqrt{5}}{\sqrt{4\pi}}z & -\sqrt{\frac{15}{4\pi}}x & 0 \end{bmatrix}. \tag{28}$$

We define $\mathrm{A}(f)$ to be a function mapping from an RGB linear spherical harmonic (represented as a 4×3 matrix) to a 3D unit vector:

$$\mathrm{A}(f) = \mathrm{OptLin}(\mathrm{LumSH}(f)), \tag{29}$$

where

$$\mathrm{LumSH}(f) = \begin{bmatrix} f_1^{-1} \\ f_1^0 \\ f_1^1 \end{bmatrix} \cdot \begin{bmatrix} w_r \\ w_g \\ w_b \end{bmatrix}, \tag{30}$$

$$\mathrm{normalize}(x) = \frac{x}{\|x\|}, \tag{31}$$

$$\mathrm{OptLin}(f) = \mathrm{normalize}([-f_1^1, -f_1^{-1}, f_1^0]). \tag{32}$$

For a fully shared luminance axis, w is defined to be e.g. $w = [0.2126, 0.7152, 0.0722]$ (for sRGB); for per-channel zonal axes w can be set to e.g. $w_r = 1, w_g = 0, w_b = 0$ for red.

The Jacobian matrix of $\mathrm{A}(f)$ is given by

$$\mathbf{J}_{\mathrm{A}} = \mathbf{J}_{\mathrm{OptLin}} \cdot \mathbf{J}_{\mathrm{LumSH}}, \tag{33}$$

$$\mathbf{J}_{\mathrm{LumSH}} = \begin{bmatrix} w_r & w_g & w_b & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_r & w_g & w_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & w_r & w_g & w_b \end{bmatrix}, \tag{34}$$

$$\mathbf{J}_{\mathrm{OptLin}} = \mathbf{J}_{\mathrm{normalize}}(\mathrm{LumSH}(f)) \cdot \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \tag{35}$$

$$\mathbf{J}_{\mathrm{normalize}}(\mathbf{d}) = \|\mathbf{d}\|^{-3} \begin{bmatrix} \mathbf{d}_y^2 + \mathbf{d}_z^2 & -\mathbf{d}_x\mathbf{d}_y & -\mathbf{d}_x\mathbf{d}_z \\ -\mathbf{d}_x\mathbf{d}_y & \mathbf{d}_x^2 + \mathbf{d}_z^2 & -\mathbf{d}_y\mathbf{d}_z \\ -\mathbf{d}_x\mathbf{d}_z & -\mathbf{d}_y\mathbf{d}_z & \mathbf{d}_x^2 + \mathbf{d}_y^2 \end{bmatrix}. \tag{36}$$