Global Illumination

Diego Lopez

November 2021

1 Introduction

The problem of realistic image synthesis is one where we are given a scene and we are tasked with producing an image as it is seen by a camera in the scene. The scene contains all the information about the shape of objects and light sources, their location, their orientation and their material properties. The type of camera, its location and orientation also needs to be specified. The image we compute is typically expected to be a rectangular $n \times m \times 3$ multidimensional array of RBG colors.

Historically this problem was solved by gradually making a renderer more complicated every time we want to include a certain effect like refraction, reflections, shadows or penumbra. One feature of renderers that has been difficult to compute in this way is global illumination, that is the idea that a surface can be illuminated directly by a light source (called direct illumination) but it can also be illuminated by other surfaces which reflect part of their light towards that area (called indirect illumination).

Borrowed from heat transfer literature, Kajiya (1986) introduced the light transport equation (also known as the rendering equation) which mathematically models the idea of global illumination as an integral equation. However, work in the heat transfer literature did not write heat transfer equations in a way that was directly amenable to image synthesis, so the form in which the light transport equation was presented is slightly different. The light transport equation was again rewritten and we will give the modern form.

Additionally, the physical values we calculate when we solve the light transport equation need to be mapped to RGB colors, so we need to understand how human color perception maps physical values to colors.

2 Color

Light is the range of electromagnetic radiation that is visible to human beings. This range is from about 380 nm to 700 nm. Objects and light sources emit a distribution of eletromagnetic radiation which is described by a *spectral power distribution* (SPD). Roughly speaking, this measures the amount of electromagnetic energy at each wavelength. However, experimentally it seems like color perception has only three degrees of freedom, in contrast to the fact that SPDs are infinite dimensional. This is called the tristimulus theory of color perception. These three dimensions are given the name *XYZ color*. They are

defined so that they match closely with the three types of cone receptors in the human retina. Given an SPD, we can compute the XYZ color by performing an inner product with three *spectral matching curves* X, Y and Z. More precisely, given an SPD S, the corresponding XYZ color is given by

$$S \mapsto (x, y, z) \in \mathbb{R}^3$$

where

$$x = \int S(\lambda)X(\lambda)\,\mathrm{d}\lambda$$
$$y = \int S(\lambda)Y(\lambda)\,\mathrm{d}\lambda$$
$$z = \int S(\lambda)Z(\lambda)\,\mathrm{d}\lambda.$$

This makes two things clear. On the one hand, there are infinitely many SPDs which result in the same visual stimulus. On the other hand, the conversion from an SPD to XYZ color is a linear map.

However this is not the same color space that we usually use to display colors on a screen. This is because the SPDs produces by screens are in fact a three dimensional subspace and we use RBG as the basis, each component corresponding to each type of light source in a computer display. In addition, RBG is in fact not well-defined in most cases because different types of displays will have different emissive SPDs. More precisely, given an RGB color (r, g, b), the emissive SPD will be given by

$$(r, g, b) \mapsto rR + gG + bB$$

where R, G and B are the emissive SPDs of each of the three types of lights in the display. Then, for each type of display, we will have different SPDs R, G and B. At the end of the day, if we want to reproduce some visual stimulus as faithfully as possible, then after one simulates the final SPD S we must first know the emissive SPDs R, G and B so that the stimulus given by it most closely resembles the stimulus we would get from the SPD we computed. However, this information is not typically available. To remedy this problem, we could save images in XYZ color space or save the spectrum.

The spectrum, of course, is infinite dimensional. So, we must save some finite dimensional representation to perform computation. The two usual ways we go about this is by sampling the spectrum at n points or writing the spectrum as a sum of basis functions. Even though the RGB color space has a lot of limitations and issues, it is still a common way to store spectrum values as it does not take a lot of storage, computation is fast and is still capable of creating realisic looking images even if it is innacurate. However, if we wish to capture phenomena such as dispersion and refraction on e.g. glass, having an accurate representation of SPDs becomes more important.

3 Radiometry

We have not yet discussed what measurement we are making for our a given SPD S. In fact, SPDs can be given in any of many radiometric quantities. The quantity we will be

Figure 1: Light source is modelled as emitting energy through rays in every direction.

interested in is called radiance. First, we will discuss the model of light we will be using.

The way light travels and interacts with atoms is described by quantum electrodynamics. However, since objects in our scene are far from relativistic in scale, classical electromagnetism is able to approximate the behavior of light. Going one step further, there is another simplification we can do when the wavelenth of electromagnetic radiation is much smaller than the objects it interacts with: model it as traveling through rays. This model is called geometric optics and it is able to produce photorealistic images. Having lost the assumption of light as either particles or waves and instead modeling light as a continuous rather than discrete particles, our model will not exhibit certain phenomena such as dispersion and interference. However, for most scenes this phenomena does not affect the final image very much.

The fundamental unit of radiometry which is conserved over time is flux or radiant power Φ, measured in watts. This measures the amount of energy passing through a given surface per unit time. This energy can either be emitted by this surface or arriving to it. It is from the conservation of this quantity that we derive our equations. We will model light sources as surfaces which emit light at a constant flux.

Another quantity of interest is irradiance E. This measures the amount of flux per area going into a point on a surface, measured in watts per meter squared. It is related to flux by integration. Indeed, given a surface A, we have

$$\phi = \int_A E(x)\,\mathrm{d}x.$$

A very similar quantity is radiant exitance M, again measured in watts per meter squared. This is just like irradiance, except it measures the amount of flux leaving a point on a surface. It is related to flux once again by integration:

$$\phi = \int_A M(x)\,\mathrm{d}x.$$

The reason we have two symbols and names for similar quantities is that when light interacts with a surface, it may not reflect all of the energy it receives. This is the case for example

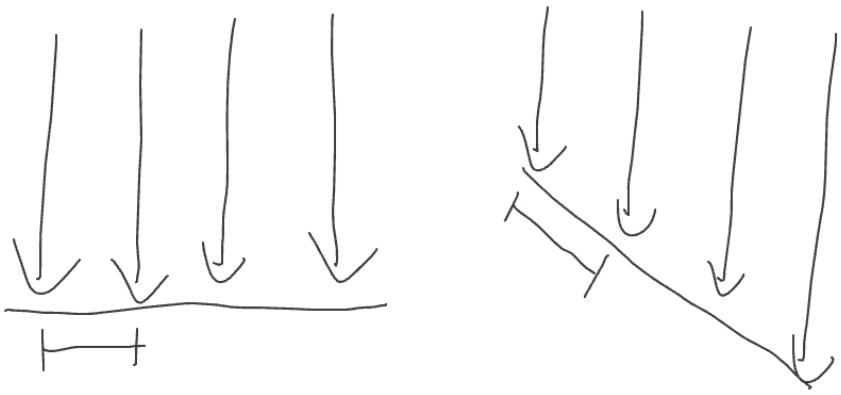

Figure 2: As we increase the angle, the irradiance due to verticle rays decreases.

Figure 3: The point x is unchanged but the orientation changes the irradiance.

with a pure black wall which does not reflect any energy and would thus have $M(x) = 0$ for all points x on the wall no matter the value of E on the wall. On the other hand, we may have some surface A in the air (as in, a surface which does not intersect any object in the scene) and in this case we would have $E(x) = M(x)$ for all $x \in A$ due to energy conservation.

In order to define radiance L, which is our ultimate quantity of interest, we must discuss two things. First, we need to introduce *Lambert's cosine law*, which states that the irradiance at a point of a surface is proportional to the cosine of the angle between the normal of the surface and the direction of the light. Roughly speaking, this is because as we increase this angle, rays will be more spread out on the surface and so the irradiance will decrease.

An important consequence of Lambert's cosine law is that the irradiance of a point in the air (as in a point which is not on any surface of any object in the scene) depends on the orientation of the surface.

Figure 4: A radian (left) and a steradian (right).

The second item we need to discuss is the concept of steradians. Just like a radian is a unit length on the circle, a steradian is a unit area on the sphere. So, for instance, the half circle is π radians while the half sphere is 2π steradians.

Now, we can define the concept of radiance L. Given a point x and a direction ω, the radiance $L(x, \omega)$ measures the flux per unit projected area per unit steradian. It is measured in watts per meter squared per steradian. We can integrate and multiply by the cosine of the angle between the normal and the ray to get back irradiance as follows, considering only the rays going into the point x:

$$E(x) = \int_{\Omega} L(x, -\omega) \cos\theta \, \mathrm{d}\omega$$

where Ω is the half sphere. We can likewise integrate to get radiant exitance by considering instead the rays going out of the point x:

$$M(x) = \int_{\Omega} L(x, \omega) \cos\theta \, \mathrm{d}\omega.$$

However, inverting the direction of ω can be a bit confusing and cumbersome, so we will use

$$L_o(x, \omega) = L(x, \omega)$$

and

$$L_i(x, \omega) = L(x, -\omega)$$

to denote rays coming into and out of the point x, respectively. To be clear, whenever we will be considering a point x on the surface of a object in the scene, we will always take ω so that it is pointing away from x. In this case, if we want to know the radiance going into x we will use $L_i(x, \omega)$ and if we want to know the radiance going out from x we will use $L_o(x, \omega)$.

The advantage of radiance is two-fold. On the one hand, given a point x and a ray ω, it is well defined. This is unlike irradiance which would depend on the orientation of the surface. This orientation invariance is given by the fact that we measured flux per

Figure 5: Incoming (left) and outgoing (right) radiance.

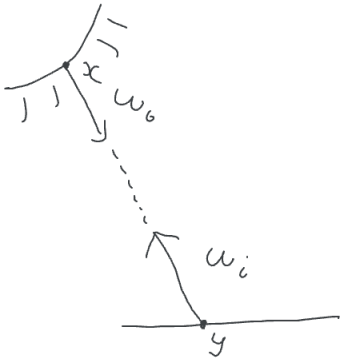

Figure 6: Radiance Invariance: $L_o(x, \omega_o) = L_i(y, \omega_i)$.

projected area to cancel the cosine falloff term. Moreover, the radiance is conserved along rays. More precisely, we have

$$L(x, \omega) = L(x + t\omega, \omega)$$

for all t smaller than the smallest value $t_0 > 0$ such that $x + t_0\omega$ is in another object in the scene. This is in stark contrast to other radiometric quantities which typically are proportional to the square of the distance to the light source due to energy conservation. In fact, since we defined radiance per unit steradian, this makes it so that invariance over rays is a direct consequence of energy conservation. Thus, computation can be drammatically reduced as the radiance coming out of the light source will be equal to the radiance of an object the light directly illuminates.

4 Light Reflection

When light hits an object, light gets reflected in different amounts in different directions. The simplest kind of reflection is when it is perfectly specular. This is the case when the object in question is a mirror. However, most objects are do not have a uniform surface.

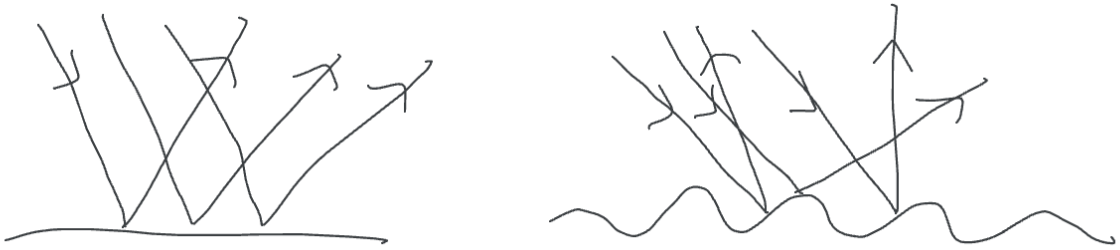

Figure 7: A model for light scattering for specular (left) and diffuse (right) reflection

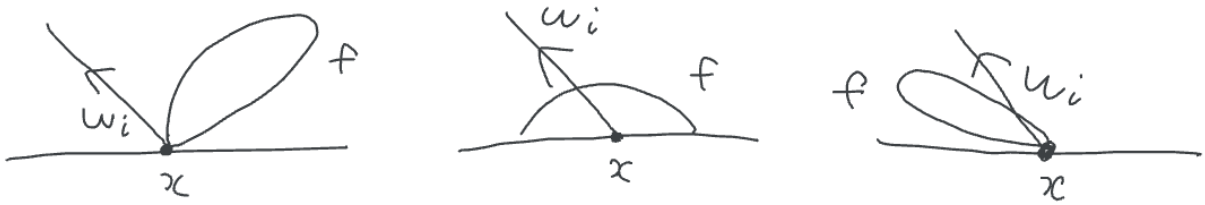

Figure 8: Examples of BRDFs for glossy (left), matte (middle) and retroreflexive (right) materials. Metal tends to be glossy, paint tends to be matte and bike reflectors tend to be retroreflexive.

These imperfections cause light to be reflected in all directions. We could model this phenomenon directly by modeling these surface imperfections and assuming the reflection is specular at this smaller scale.

We will instead model light in a different way: any time light hits a point on the surface, it gets reflection in all direction according to a distribution called a *bidirectional reflectance distribution function* or BRDF. Then, the material properties of the objects in the scene will be simply given by the BRDF for each object. More precisely, for the BRDF f, if we are given a point x, an incoming direction ω_i and an outgoing direction ω_o, the value $f(x, \omega_i, \omega_o)$ is the ratio of light that is scattered in direction ω_o when the point x is illuminated by a ray coming in direction ω_i.

Given a radiance value $L_i(x, \omega_i)$, consider the expression

$$L_i(x, \omega_i) f(x, \omega_i, \omega_o) \cos \theta_i$$

where ω_o is some outgoing direction and θ_i is the angle between the normal of the surface and the incoming direction ω_i (recall Lambert's cosine law). This expression does not measure how much radiance is leaving the surface in direction ω_o due to the incoming radiance in direction ω_i. It is instead a measure of "randiance density" or "differential radiance". The total outgoing radiance $L_o(x, \omega_o)$ is given by integrating over all possible

incoming directions:

$$L_o(x, \omega_o) = \int_\Omega L_i(x, \omega_i) f(x, \omega_i, \omega_o) \cos\theta_i \, \mathrm{d}\omega_i.$$

This is called the *reflection equation*.

There are two properties that physical BRDFs must satisfy. The first property is called *reciprocity*. It states that for every pair of directions ω and ω', we have that

$$f(x, \omega, \omega') = f(x, \omega', \omega).$$

This is a consequence of a more general fact called the Helmholtz reciprocity principle. The second property is conservation of energy. An object may absorb some of the energy from the radiance it receives, but it may not reflect more than it absorbs. More precisely, we must have that for all ω_o,

$$\int_\Omega f(x, \omega_i, \omega_o) \cos\theta_i \, \mathrm{d}\omega_i \leq 1.$$

5 Camera Models

So far we have specified what we mean by a scene. We have a collection of light sources which are given as surfaces which emit radiance. This means that when we specify a light source, we must specify the shape, orientation, as well as the radiance it emits for all points in the surface and for all outgoing directions. Moreover, the scene contain objects which interact with the light. For each point on the surface of an object, we must specify the BRDF on that point's surface. The remaining part of the scene that we need to model is the camera. The simplest model we can use is called the pinhole camera. It is specified by providing a point in the scene as well as a rectangle to serve as the film. Then, we trace a point on the rectangle towards the point. This segment specifies the direction of the radiance that the point on the film can see. That is, every point on the film plane measures exactly one radiance value. Computing all the radiance values on the film plane will result in the image of the scene.

There are also other camera models. For instance, there is the simple lens model. In this model, there is an aperture and a single lens in from of the film plane. Now, since light rays are refracted by the lens, a single point on the film plane is illuminated by infinitely many rays, so we must perform an additional integration. This in turn allows the thin lens model to capture depth of field. It is also possible to simulate the same lens systems found in real cameras to get even more photorealism. We can also animate the scene and additionally integrate over an exposure period and capture motion blur. We will use the pinhole camera model as it is a good enough approximation when all the objects are more or less the same distance to the camera.

6 Monte Carlo Integration

It turns out that the integrals that result from computing final radiance values are have a very high number of dimensions. This means that usual deterministic ways to compute

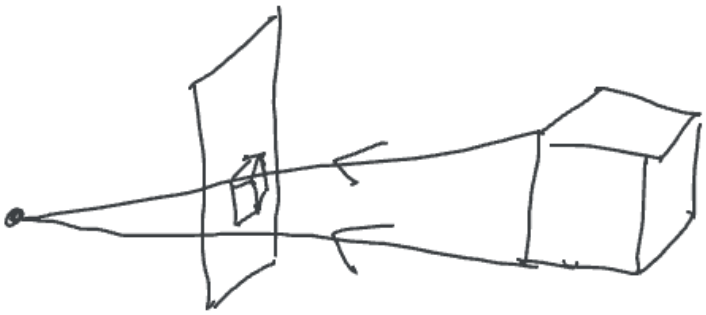

Figure 9: The pinhole camera model.

these integrals result in very slow runtimes. We can however use a numerical integration method known as Monte Carlo integration which gives us a faster convergence rate than any of the usual deterministic integration methods.

Consider a compact set $S \subseteq \mathbb{R}^n$ (S may be a manifold of lower dimension) and an integrable function $f : S \to \mathbb{R}$. We wish to estimate the integral

$$\int_S f(x)\,\mathrm{d}x,$$

using the usual measure on S. Consider a uniform and i.i.d. sample $X_1, \ldots, X_n \in S$. We define the *blind Monte Carlo estimator* as taking this sample and computing the value

$$F_n = \frac{\mu(S)}{N} \sum_{i=1}^{N} f(X_i),$$

where $\mu(S)$ is the measure of S. This is esimator for the integral is consistent and unbiased. However, there are two issues. The first is that what if S is a set with infinite measure. Then, it would not be possible to sample points uniformly on it. Secondly, the central limit theorem guarantees a convergence rate of only $O(\sqrt{n})$.

To remedy these issues, we introduce a simple *informed Monte Carlo estimator*. Consider a set $S \subseteq \mathbb{R}^n$ which may not be compact and an integrable function $f : S \to \mathbb{R}$. Let p be some probability density function of S so that $p(x) > 0$ whenever $f(x) \neq 0$. We once again wish to estimate the integral

$$\int_S f(x)\,\mathrm{d}x,$$

using the usual measure on S. Consider an i.i.d sample $X_1, \ldots, X_n \in S$ drawn according to the distribution p. Our estimator will take this sample and compute the value

$$F_n = \frac{1}{N} \sum_{i=1}^{n} \frac{f(X_i)}{p(X_i)}.$$

This estimator for the integral is again consistent and unbiased. We of course have the benefit that we can compute integrals on noncompact spaces, but in fact we also have the benefit that if p is a good approximation of $|f|$, then it turns out that the variance of the estimator decreases. However, bad choices of p could result in a variance larger than a uniform sampling. Thus, the idea is that by knowing more information about the domain in question, we can pick probability density functions which reduce the variance and in turn have smaller convergence times. It is important to point out however that while we can do many things to reduce variance, the convergence rate is still $O(\sqrt{n})$.

7 Light Transport

When light hits the camera film, some of the light is due to radiance coming directly from the light source which is already given as part of the description of the scene. However, there is also radiance due to the fact that the light source illuminates objects in the scene and so objects will reflect some of this light towards the points in the film plane. This kind of lighting is called direct lighting and it is can be easily computed using Monte Carlo integration. Of course, a given scene may have multiple light sources, but since integration is linear the total exitant radiance due to light sources will be a sum of each of the direct lighting due to each of the light sources. Given a light source, a point in the scene x and an outgoing direction ω_o, the radiance due to direct lighting is given by

$$L_d(x, \omega_o) = \int_\Omega L_e(x, \omega_i) f(x, \omega_i, \omega_o) \cos\theta_i \, \mathrm{d}\omega_i,$$

where $L_e(x, \omega_i)$ is the radiance emitted by the light source in question. Of course, $L_e(x, \omega_i)$ is zero for all direction ω_i which do not go towards the light source or are occluded by another object in the scene. This means that the computation of direct lighting can be sped up by only sampling direction which go towards the light source in question when performing Monte Carlo integration.

Of course, points which are not illuminated still emit radiance. This is because points which are directly illuminated emit radiance towards points which are not directly illuminated. However, those points in turn will emit more radiance towards all the points which are visible to them. How do we compute the final radiance values which we need in order to generate our image?

At the end of the day, we know that energy is conserved. That is, any energy coming into a point is equal to the energy absorbed and reflected. This gives us that for all points x in the scene, we have that the outgoing radiance in direction ω_o is given by

$$L_o(x, \omega_o) = L_e(x, \omega_o) + \int_\Omega L_i(x, \omega_i) f(x, \omega_i, \omega_o) \cos\theta_i \, \mathrm{d}\omega_i,$$

where $L_e(x, \omega_o)$ is the radiance emitted at x, which is nonzero if and only if x is on the surface of a light source. This is the light transport equation (LTE). It a Fredholm integral equation of the second kind, as we have an unknown function L inside and outside of the integrand. It is also inhomogeneous because of the $L_e(x, \omega_o)$ term.

The LTE has analytic solutions for very simple scenes with simple BRDFs and light sources and emitted radiance values for those light sources. This is a useful way to check if the renderer has be incorrectly implemented by comparing the numerical solution to the analytical solution.

In order to solve the LTE we will gradually rewrite the LTE in different forms which ultimately will be able to be write an estimator which is consistent and unbiased. We will first rewrite the LTE as an integral over all the surfaces in the scene instead of over all the directions of the halfsphere. To simplify the equations, we will introduce some notation to write radiance and BRDFs in terms of points instead of directions. The exitant radiance from p' to p will be written as

$$L(p' \to p) = L(p', \omega),$$

where ω is the unit vector from p' in the direction of p. The BRDF at p' due to radiance coming from p'' and leaving towards p will be written as

$$f(p'' \to p' \to p) = f(p', \omega_o, \omega_i),$$

where ω_o is the unit vector from p' in the direction of p and ω_i is the unit vector from p' in the direction of p''. We define the *visibility function* V which is given by

$$V(p \leftrightarrow p') = \begin{cases} 1 & \text{if } p \text{ and } p' \text{ are mutually visible} \\ 0 & \text{if } p \text{ and } p' \text{ are not mutually visible.} \end{cases}$$

We define a geometry term between pairs of points:

$$G(p \leftrightarrow p') = V(p \leftrightarrow p') \frac{\cos\theta \cos\theta'}{r^2}$$

where

- θ is the angle between the normal at the point p and the vector from p to p',
- θ' is the angle between the normal at p' and the vector from p' to p, and
- r is the distance from p to p'.

Then, after a change of coordinates transformation we have the *surface form of the LTE*:

$$L(p' \to p) = L_e(p' \to p) + \int_A f(p'' \to p' \to p) L(p'' \to p') G(p'' \leftrightarrow p') \, \mathrm{d}p'',$$

where A is the union of all the surfaces of in the scene.

The key observation to rewrite the LTE once again is that radiance $L(p_1 \to p_0)$ is contributed to by different paths, all starting from a light source and ending at p_0. For example, a path of length 1 is simply a path starting from the light source and ending at p_0. This corresponds to the radiance coming directly from the light source hitting the point p_0. A path of length 2 is simply a path starting from the light source, hitting a point

p_1 which then reflects radiance towards the point p_2. In fact, the contribution to radiance from paths of length 2 is exactly the radiance due to direct illumination. For convenience, given a path $\bar{p}_n = p_0, \ldots, p_n$ of length n, we define the *throughput* of the path as

$$T(\bar{p}_n) = \prod_{i=1}^{n-1} f(p_{i+1} \to p_i \to p_{i-1}) G(p_{i+1} \leftrightarrow p_i).$$

Then, the contribution of radiance from paths of length n becomes

$$L_n(p_1 \to p_0) = \int_{A^{n-1}} L_e(p_n \to p_{n-1}) T(\bar{p}_n) \, \mathrm{d}p_2 \cdots \mathrm{d}p_n.$$

Then, the total radiance arriving at p_0 from the point p_1 is given by

$$L(p_1 \to p_0) = \sum_{n=1}^{\infty} L_n(p_1 \to p_0).$$

For the moment we will discuss how to estimate $L_n(p_1 \to p_0)$ for a fixed $n > 2$. To estimate $L_n(p_1 \to p_0)$, we could sample $p_2, \ldots, p_{n-1}$ over A uniformly (at least assuming A is compact) and p_n uniformly over the light sources (assuming the union of the light sources is compact). The issue is that the values of the sample will be zero whenever there are a pair of points p_i and p_{i+1} which are not mutually visible, resulting in high variance and long convergence times.

What we will do instead is to build the sample incrementally: given the point p_i, the next point will be sampled with weight according to the BRDF f at that point given that the ray came from the direction p_{i-1}. Once we have a path of length $n-1$, we sample p_n by picking uniformly over the union of the surfaces of the light sources. This recudes variance considerably as we never result in a point which is not visible and moreover we tend to sample directions which contribute more to the radiance. Recall that we can choose whatever distribution to sample the path space from as long as we normalize the contributions by the distribution and the distribution is nonzero whenever the function we are integrating is nonzero. This provides us with a Monte Carlo estimator for $L_n(p_1 \to p_0)$ for a fixed $n > 2$.

Now, we wish to compute the sum of all the contributions of paths for each length. The way we do it is by first splitting the sum

$$L(p_1 \to p_0) = L_1(p_1 \to p_0) + L_2(p_1 \to p_0) + \sum_{n=3}^{\infty} L_n(p_1 \to p_0).$$

Then, we can easily compute the contribution from $n = 1$ (which will only be nonzero if the ray is in the direction of a light source), as well a the contribution from $n = 2$ (which is direct illumination and we already know how to compute).

To compute the rest of the infinite sum, we use a random value with support equal to all positive integers. Then, we will sample a path of that length. Of course, the sample has to be appropriately weighted by the probability mass function for the random variable. This method of computing $L(p_1 \to p_0)$ is called *path tracing* and it is a consistent and unbiased estimator which has low variance for many scenes.

8 Conclusion

We have barely scratched the surface of realistic image synthesis. We have described the camera film as a continuous function measuring light coming from the scene. An image however is a discrete array. So, care needs to be taken to ensure that the method we use to sample the continuous function does not add artefacts like aliasing, blurring or the Gibbs phenomenon. Even after sampling radiance values for each pixel in the image, we need do deal with the fact that the radiance values can be any value in $[0, \infty)$ (even if the total energy is finite) and we need to map these values to the range $[0, 1]$ to be able to display them. The reason for the radiance being unbounded is that given a light source, we can focus the energy on an arbitrarity small disk in the film plane (using mirrors and glass), which increase the radiance on that disk. Moreover, the process of sampling a ray and finding which point intersects this ray, called *ray tracing* can be subject to numerical instability. There are certain scenes where many points are not directly illuminated by the light source, causing a zero visibility term when we sample the light source which causes high variance. There are other algorithms such as bidirectional path tracing which can deal with these cases. In addition, many variance reduction techniques can be used to further reduce variance.

References

[1] Matt Pharr, Wenzel Jakob, and Greg Humphreys. 2016. *Physically Based Rendering: From Theory to Implementation* (3rd. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.