



ugr | Universidad
de **Granada**

TRABAJO FIN DE GRADO
DOBLE GRADO INGENIERÍA INFORMÁTICA Y
MATEMÁTICAS

Métodos de Monte Carlo de baja varianza para simulación numérica de iluminación global

Autor
Emilio José Hoyo Medina

Directores
Carlos Ureña Almagro
María del Carmen Segovia García

Granada, septiembre de 2021

Métodos de Monte Carlo de baja varianza para simulación numérica de iluminación global

Emilio José Hoyo Medina

Palabras clave: ray-tracing, ecuación de renderización, método de Monte Carlo , muestreo de variables aleatorias.

Resumen

Los métodos de Monte Carlo permiten estimar ciertas funciones, siendo en ocasiones la única solución manejable gracias a que su ratio de convergencia es independiente de la dimensionalidad del espacio de trabajo. Este es el motivo de su importancia dentro de los renderizadores fotorrealistas.

Este trabajo estudia la base matemática de los principales métodos de Monte Carlo, permitiéndonos así detallar los algoritmos de simulación numérica de iluminación global más básicos utilizados en la síntesis de imágenes realistas. Posteriormente se analizarán e implementarán una serie de artículos recientes que buscan reducir la varianza de los métodos de Monte Carlos usados en ray-tracing.

Low variance Monte Carlo methods for numerical simulation of global illumination

Emilio José Hoyo Medina

Keywords: ray-tracing, rendering equation, Monte Carlo method, sampling random variables.

Abstract

Photorealistic rendering is a type of rendering that use the principles of physics to model the behaviour of light. Most photorealistic renderers are based on the ray tracing algorithm. Ray tracing is an algorithm in which the color of each pixel in the final image is calculated by tracing rays from the camera into the scene, and calculating the amount of light traveling along it. This algorithm was initially introduced by Whitted ([Whi80]), and has gained relevance since then. It is used in areas such as animation and film production.

In ray tracers, each time the camera generates a ray, in order to calculate the amount of light that travels along this ray, the first task is to find the intersection of the ray with the scene, obtaining a point P . Then, the radiance reflected in the direction of the camera from P is approximated. This reflected radiance is given by the rendering equation, a second-type Fredholm integral equation that describes the global radiance distribution in the scene. This equation in most cases cannot be solved analytically, so it must be approximated numerically.

Monte Carlo methods are a very useful numerical approximation tool in this context. The aims of this methods are to solve one or both of the following problems:

- There are Monte Carlo methods focused on sampling random variables following a given probability distribution that take values in a space E .
- To use these generated samples to approximate the expectation of a function (defined over E) under this distribution. Once the first problem has been solved, the second problem can be solved by using an estimator known as Monte Carlo estimator.

The reason Monte Carlo methods are used instead of other numerical approximation algorithms in rendering is that their convergence ratio is independent of the dimensionality of the space sampled. That characteristic makes the Monte Carlo methods in several situations the only viable solution.

Addressed problem and report description

The addressed problem in this work can be divided into two sub-problems:

- To study the main Monte Carlo methods used in rendering, from a mathematical and formal approach.
- To study different ways to reduce the variance of the Monte Carlo estimator, especially when it is used to approximate the rendering equation.

In order to address the first problem, basic concepts related to stochastic processes have been needed, which have been acquired through the book [Wil91].

The main content of this report has been divided into two blocks, the first one focused on the mathematical theory behind Monte Carlo methods and its applications in ray tracing, and the other on the implementation of software in a ray tracer:

1. The first block aims to study the mathematical basis of the main Monte Carlo methods used in rendering, as well as to describe the two basic algorithms used to approximate the rendering equation: direct lighting and path tracing. These two algorithms are based on calculating an estimate of the integral of the equation through a Monte Carlo estimator.
2. The second block aims to study and implement recent papers that describe random variable sampling methods that reduce the variance of the estimators used to approximate the rendering equation.

The chapter 3 begins by detailing the results and definitions related to stochastic processes that will be used in the rest of the chapter. Next we will detail the theorems and definitions that will end up leading us to a proof of the Strong Law of Large Numbers, a theorem that ensures the convergence of the Monte Carlo estimator when the number of samples taken tends to infinity. After this, the most well-known Monte Carlo methods for random variable sampling used in realistic renderers are detailed, being necessary to detail some aspects of the Markov processes for the explanation of the Metropolis method. Finally, three essential variance reduction methods in rendering are explained: Importance sampling, stratification and multiple importance sampling. Importance sampling is based on a good choice of the distribution that we sample when approximating an integral. Stratification is based on dividing the sampled space into stratum, and sampling each stratum separately. Multiple importance sampling solves the problem of combining different sampling methods in a single estimator.

The chapter 4 serves as a link between the first part and the second, and details from a formal point of view the two main approximation algorithms

for the rendering equation based on Monte Carlo methods. On one hand we have the direct lighting algorithm, which consists of taking samples only from the directions that point towards the light sources in the scene. Indirect lighting due to light reflection through the scene is not taken into account in this approach. On the other hand we have the path tracing algorithm, which generates random paths through the scene, taking into account the characteristics of the matter with which the path intersects. In this chapter we acquire the knowledge about the ray tracing algorithm necessary to understand the importance of sampling random variables defined on the set of directions pointing towards a light source. This process is known as light sampling.

The chapter 5 details alternative area light sampling methods, detailing their advantages and disadvantages. Three recent area light sampling algorithms are discussed. The first is used for rectangular light sources, the second for disk-shaped light sources, and the third for spherical light sources. These algorithms will be implemented in an open source photorealistic renderer. pbrt has been selected for implementation because it has a freely available book, [PJH16], which explains both the mathematical theory behind the rendering systems and its practical implementation. Therefore, this book, in addition to being an important reference in the concepts that this work deals with, serves as documentation of the system used to implement the algorithms. Finally, software tests of the implementation will be carried out to show the results obtained.

It is relevant to note that for the elaboration of this project a previous study of ray-tracing system has been required, since this was a rendering approach practically unknown to the author of this work. In the understanding of the ray tracing algorithm, in addition to pbrt, has been key [Shi20a], [Shi20b] and [Shi20c], three books that give an easy-to-understand introduction to ray tracing, detailing the development of a simple ray tracer.

Conclusions

The objectives of the work have been fulfilled in a satisfactory way, having acquired quite important general knowledge about photorealistic rendering, which goes beyond those detailed in this work, and which has aroused great interest in the author. The mathematical aspects treated based on stochastic processes have also been interesting.

However, there are many more topics in this field to develop. The light sampling algorithms described in this work, while reducing the variance of the Monte Carlo estimator that approximates the rendering equation, they have their drawbacks. Many of them significantly increase the execution time required to generate an image, so possibilities for improvements in terms of efficiency could be analyzed.

Another way of improvement would be to try to find parameterizations of the projected solid angle associated with rectangular or disk-shaped light

sources. It may also be interesting to investigate the parametrization of the solid angle or projected solid angle subtended by other geometric shapes.

Moreover, the use and characteristics of other production renderers can be studied. Other forms of approximation of the rendering equation, such as bidirectional methods or the Metropolis light transport method, are still pending. An also very relevant aspect that we have overlooked is the behavior of light in environments other than vacuum.

Yo, **Emilio José Hoyo Medina**, alumno de la titulación Doble grado en ingeniería informática y matemáticas de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación** y de la **Facultad de Ciencias** de la **Universidad de Granada**, con DNI XXXXXXXXX, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Emilio José Hoyo Medina

Granada a 6 de septiembre de 2021.

D. **Carlos Ureña Almagro**, Profesor del Departamento Lenguajes y Sistemas Informáticos de la Universidad de Granada.

Dña. **María del Carmen Segovia García**, Profesora del Departamento Estadística e Investigación Operativa de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado *Métodos de Monte Carlo de baja varianza para simulación numérica de iluminación global*, ha sido realizado bajo su supervisión por **Emilio José Hoyo Medina**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 6 de septiembre de 2021.

Los directores:

Carlos Ureña Almagro María del Carmen Segovia García

Agradecimientos

Muchas gracias a mis dos tutores, Mari Carmen y Carlos, por la ayuda durante la elaboración de este trabajo.

Gracias a mi familia, amigos, y en especial gracias a Raquel, sin cuyo apoyo no habría podido llegar hasta aquí.

Índice general

1. Introducción	1
1.1. El algoritmo de ray-tracing	1
1.1.1. Cámara	2
1.1.2. Intersección de rayos con objetos de la escena	3
1.1.3. Visibilidad	3
1.1.4. Distribución de la dispersión de la luz en superficies	4
1.1.5. Distribución de la luz	5
1.1.6. Propagación de los rayos	12
1.2. Concepto de método de Monte Carlo y su relación con el ray-tracing	13
1.3. Problema a resolver	15
1.4. Contenido de la memoria y principales fuentes y herramientas empleadas	15
2. Objetivos	17
I Matemáticas	19
3. Métodos de Monte Carlo	21
3.1. Conceptos básicos	21
3.2. Convergencia del estimador de Monte Carlo	28
3.2.1. Tiempos de parada	28
3.2.2. Número de upcrossings y martingalas inversas	30
3.2.3. Ley fuerte de los números grandes	35
3.3. Métodos de Monte Carlo para muestreo de variables aleatorias	40
3.3.1. Método de inversión	41
3.3.2. Método de rechazo	42
3.3.3. Método de Metropolis	45
3.3.4. Transformación entre distribuciones	51
3.4. Métodos para reducir la varianza	53
3.4.1. Muestreo de importancia	53
3.4.2. Estratificación o muestreo estratificado	54

3.4.3. Muestreo de importancia múltiple	55
4. Aproximación de la ecuación de renderización	57
4.1. Iluminación directa	57
4.2. Path-Tracing	58
II Informática	65
5. Métodos de muestreo directo de fuentes de luz	67
5.1. Muestreo uniforme de rectángulos esféricos	70
5.1.1. Construcción de la parametrización M	71
5.1.2. Resultados obtenidos	75
5.2. Muestreo uniforme de elipses esféricas	78
5.2.1. Descripción del sistema de referencia y otros parámetros	78
5.2.2. Construcción de la parametrización M_r	80
5.2.3. Construcción de la parametrización M_p	82
5.2.4. Resultados obtenidos	83
5.3. Muestreo de casquetes esféricos proyectados	86
5.3.1. Descripción del sistema de referencia y otros parámetros	86
5.3.2. Construcción de la parametrización M_s	88
5.3.3. Construcción de la parametrización M_t	89
5.3.4. Resultados obtenidos	91
5.4. Implementación en pbrt	93
6. Conclusiones y trabajo futuro	97
Bibliografía	101

Índice de figuras

1.1. Simulación de la cámara estenopeica	2
1.2. Reflexión de la luz	4
1.3. Ángulo plano	6
1.4. Ángulo sólido	6
1.5. Medida del ángulo sólido de una superficie simple y regular a trozos	8
5.1. Puntos relevantes de la fuente de luz rectangular F	71
5.2. Descripción de los rectángulos esféricos $\pi_P(F)$ y Q_v	73
5.3. Relación entre x_v y φ_v	74
5.4. Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido. Imágenes generadas con una muestra por píxel y con el algoritmo de iluminación directa.	75
5.5. Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido. Imágenes generadas con una muestra por píxel.	76
5.6. Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido.	77
5.7. Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido en medios distintos al vacío.	77
5.8. Comparativa del tiempo de ejecución.	78
5.9. Sistema de referencia y puntos relevantes	79
5.10. Descripción del mapa M_r	81
5.11. Descripción del mapa M_p	82
5.12. Comparativa entre muestreo uniforme respecto al área, mapa radial y paralelo en medios distintos al vacío.	84
5.13. Comparativa entre muestreo uniforme respecto al área, mapa radial y paralelo. Imágenes generadas con una muestra por píxel y con el algoritmo de iluminación directa.	85
5.14. Comparativa del tiempo de ejecución.	86
5.15. Parametrización de la esfera Q y la luna L	87

5.16. Descripción del mapa M_s .	89
5.17. Descripción del mapa M_t .	90
5.18. Comparativa entre muestreo uniforme respecto al ángulo sólido y muestreo uniforme respecto al ángulo sólido proyectado, usando los mapas M_s y M_t . Imágenes generadas con una muestra por píxel e iluminación directa.	92
5.19. Comparativa del tiempo de ejecución.	93

Capítulo 1

Introducción

La renderización es el proceso de generar una imagen a partir de la descripción de una escena 2D o 3D. Dicho proceso puede ser abordado de muchas formas, siendo nuestro objeto de estudio los renderizadores físicamente realistas que utilizan el algoritmo de *ray-tracing*, denominados *ray tracers*. Los renderizadores físicamente realistas o fotorrealistas utilizan los principios de la física para modelar el comportamiento de la luz y su interacción con la materia. En este capítulo se presentará el algoritmo de ray-tracing y su fundamentación física y matemática, así como también se describirá brevemente el concepto de método de Monte Carlo y su uso dentro de los ray-tracers. Posteriormente se establecerán los contenidos de este trabajo, comentando las fuentes utilizadas.

1.1. El algoritmo de ray-tracing

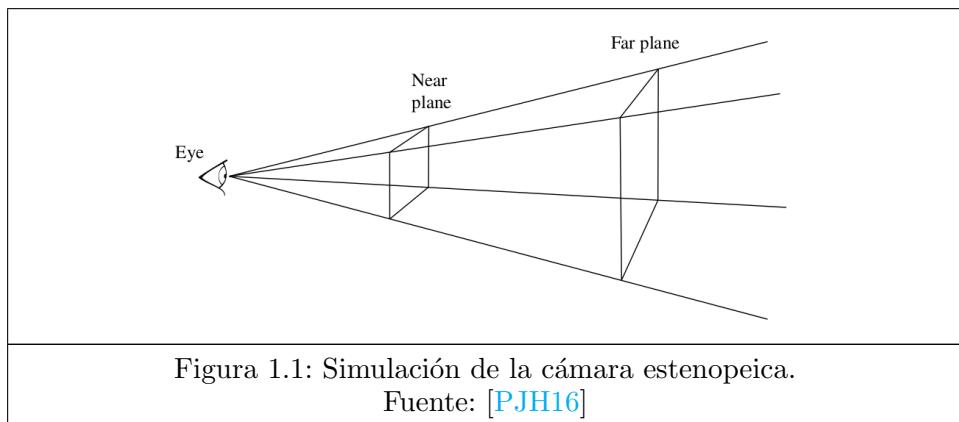
El objetivo de los renderizadores fotorrealistas es crear imágenes de escenas 3D que simulen la realidad de la manera más fiel posible, es decir, buscan generar imágenes que parezcan fotografías reales. La mayor parte de renderizadores fotorrealistas utilizan ray-tracing, un algoritmo que consiste en simular el camino que siguen los rayos de luz a través de la escena para poder así determinar el color de cada píxel de la imagen final. Este enfoque más realista de la renderización empezó a cobrar importancia durante la década de 1980, con el artículo [Whi80], que introdujo la idea de usar ray-tracing para el cálculo de la distribución de la luz en escenas 3D. Desde entonces se han detallado numerosos modelos enfocados a la síntesis realista de imágenes. Destacar también la importancia de [Kaj86], que presentó una solución al problema del cálculo de iluminación global a través de un modelo riguroso basado en la ecuación de renderización, mostrando una forma de aproximar dicha ecuación mediante el uso de métodos de Monte

Carlo, y [Vea98], que desarrolló nuevos algoritmos para ray-tracing basados en métodos de Monte Carlo.

Vamos por tanto a detallar los aspectos más relevantes del algoritmo de ray-tracing, siguiendo las ideas detalladas en el libro [PJH16]. Hay una serie de objetos y fenómenos que todos los ray tracers deben ser capaces de simular de manera más o menos realista: la cámara desde la que se visualiza la escena, la intersección de rayos con objetos de la escena, la visibilidad de las fuentes de luz desde un punto de la escena, la distribución de la cantidad de luz reflejada y transmitida por la superficie de los objetos (*surface scattering*), la distribución de la luz en la escena y la propagación de los rayos a través de medios diferentes al vacío.

1.1.1. Cámara

Si queremos generar una imagen realista necesitamos simular el funcionamiento de una cámara real, y la forma en que modelemos la cámara determinará el modo en que la escena es vista. El modelo de cámara más simple utilizado en ray-tracing está basado en la cámara estenopeica, una cámara sin lente consistente en una caja con un pequeño agujero por el que entra la luz y un trozo de papel fotográfico donde queda grabada la imagen.



La cámara estenopeica puede ser simulada haciendo una abstracción y suponiendo que el papel de película se encuentra a cierta distancia en la dirección en la que la cámara está apuntando. Tal y como podemos ver en la figura 1.1, uniendo la posición dónde se encuentra la cámara con los filos de la película (near plane) se obtiene una región del espacio con forma de pirámide, que será la región del espacio que será captada en la imagen final. Sin embargo falta determinar que color tendrá la imagen en cada uno de sus píxeles. Para ello se traza un número concreto de rayos que pasan por cada

píxel y a continuación se calcula la cantidad de luz que viaja a través de cada rayo trazado, lo cual constituye el principal problema a resolver dentro de los ray tracers. Por último se hace una media de los rayos que pasan por cada píxel para calcular su color.

El modelo de cámara recién presentado, a pesar de ser muy usado dentro de la renderización por ordenador, está bastante alejado del funcionamiento de una cámara actual, por lo que es deseable un modelo más complejo dentro de un renderizador fotorrealista. Esto se consigue simulando la utilización de múltiples lentes que nos permiten conseguir diversos efectos en la imagen final. Por otro lado, es común utilizar un parámetro temporal asociado a los rayos trazados, trazando varios rayos a lo largo del tiempo. La posición de los objetos de la escena puede cambiar en el tiempo, con lo que de esta manera se puede simular que ciertos objetos estén en movimiento al momento de tomar la fotografía.

Aunque puedan utilizarse modelos de cámara más complejos, todos estos modelos coinciden en que su principal tarea es determinar el espacio dentro de la escena que será captado en la imagen y trazar los rayos correspondientes a cada píxel.

1.1.2. Intersección de rayos con objetos de la escena

Una vez los rayos son trazados por la cámara, estos interseccionarán con objetos de la escena. La primera tarea del renderizador es ser capaz de calcular el punto en que cada rayo intersecciona con la escena, así como calcular las propiedades geométricas locales del objeto interseccionado.

1.1.3. Visibilidad

Recordemos que el objetivo es calcular la cantidad de luz que viaja por cada rayo que llega a la película de la cámara, o lo que es lo mismo, si consideramos el punto dónde el rayo intersecciona con la escena, calcular la cantidad de luz que viaja desde dicho punto en dirección a la cámara. Este proceso que se utiliza para calcular el color que se percibe de un punto de la escena se conoce como *sombreado* o *shading*.

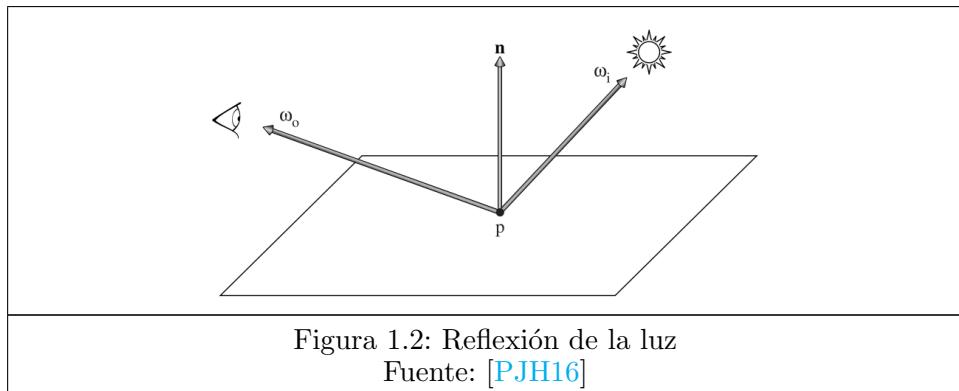
Para calcular la cantidad de luz que es reflejada desde un punto en una dirección, primero es necesario estimar la cantidad de luz que llega a él. Por tanto es esencial que el renderizador sea capaz de determinar si una fuente de luz es visible desde el punto de la escena que está siendo sombreado.

1.1.4. Distribución de la dispersión de la luz en superficies

Llegados a este punto vemos que la forma en que los objetos de la escena reflejan y transmiten la luz que les llega es de gran importancia. Esto depende de las características de la materia que compone el objeto. La función que describe la forma en que los objetos de la escena reflejan la luz que reciben se denomina *función de distribución de reflectancia bidireccional* o *BRDF* (*bidirectional reflectance distribution function*), y la notaremos por:

$$f_r : \mathbb{R}^3 \times \mathbb{S}^2 \times \mathbb{S}^2 \rightarrow \mathbb{R}_0^+$$

Dónde notamos por \mathbb{S}^2 a la esfera unidad de \mathbb{R}^3 . De esta manera, sea p un punto en la superficie de un objeto, ω_i la dirección de un rayo incidente, y ω_o la dirección de un rayo reflejado, entonces la cantidad de energía reflejada por la superficie del objeto en el punto p desde la dirección ω_i hacia la dirección ω_o es $f_r(p, \omega_o, \omega_i)$. Obviamente al tratarse de energía reflejada, si consideramos la esfera cuyo polo norte es la normal en p a la superficie (notada como n en la figura 1.2), si ω_i se encuentra en el hemisferio inferior de dicha esfera entonces $f_r(p, \omega_o, \omega_i)$ será igual a 0.



De manera similar podemos definir la función que describe la cantidad de energía transmitida, es decir, la cantidad de energía que atraviesa el objeto y escapa en una dirección. A dicha función la llamaremos *función de distribución de transmitancia bidireccional* o *BTDF* (*bidirectional transmittance distribution function*) y la notamos por f_t .

Por último consideramos la función que describe de manera conjunta la energía reflejada y transmitida, la *función de distribución de dispersión bidireccional* o *BSDF* (*bidirectional scattering distribution function*) que notamos por f .

En el caso de un objeto que presente reflexión especular perfecta, el BSDF en su superficie se puede describir de manera sencilla, ya que dada

una dirección ω_i , toda la luz procedente de esa dirección será reflejada en una única dirección saliente ω_o , que será aquella que forme el mismo ángulo con la normal a la superficie que ω_i . Este sería el caso de un espejo perfecto. Sin embargo la mayor parte de objetos de la vida real no se comportan de esta manera, por lo que en un ray tracer habrá diversos modelos que simulen diferentes materiales.

1.1.5. Distribución de la luz

La radiometría y la óptica geométrica nos otorgan las herramientas físicas y matemáticas que necesitamos para describir el comportamiento de la luz, y forman la base de la mayor parte de algoritmos utilizados en ray-tracing. La radiometría es la ciencia que trata la medición de radiación electromagnética, ya sea visible o no. La óptica geométrica tiene en cuenta propiedades macroscópicas de la luz, que son suficientes para describir la forma en que la luz interacciona con objetos mucho más grandes que su longitud de onda. Por otra parte, no es raro simular fenómenos propios de la óptica física, aunque en nuestro caso no tendremos en cuenta este tipo de fenómenos. Hay una serie de asunciones que haremos acerca del comportamiento de la luz para simplificar el cálculo de la energía radiante en la escena:

- Linealidad: El efecto combinado de dos entradas en un sistema óptico es siempre igual a la suma del efecto de cada una de las entradas individualmente.
- Conservación de la energía: La energía radiante reflejada o transmitida por una superficie nunca será mayor que la energía incidente sobre dicha superficie.
- Estado de reposo: Se asume que la distribución de la luz en la escena ha alcanzado el equilibrio, por lo que no habrá cambios en la distribución de la radiancia a lo largo del tiempo.

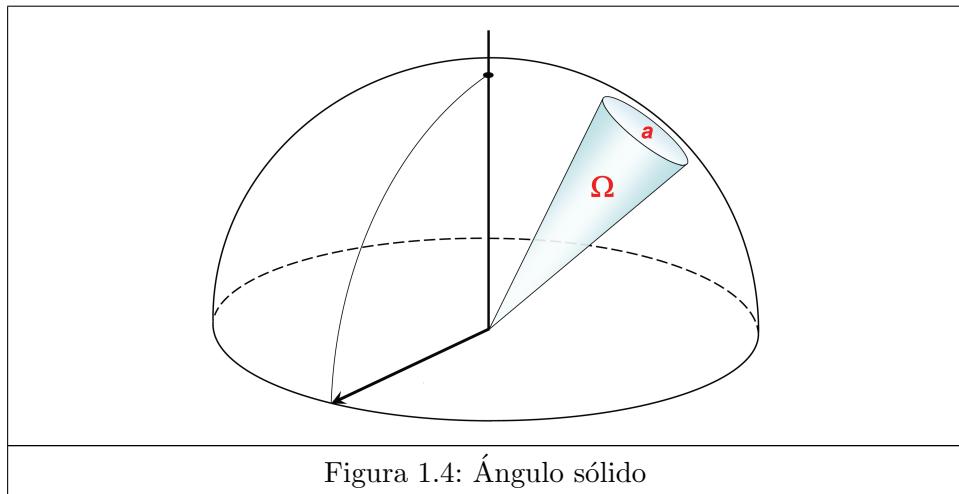
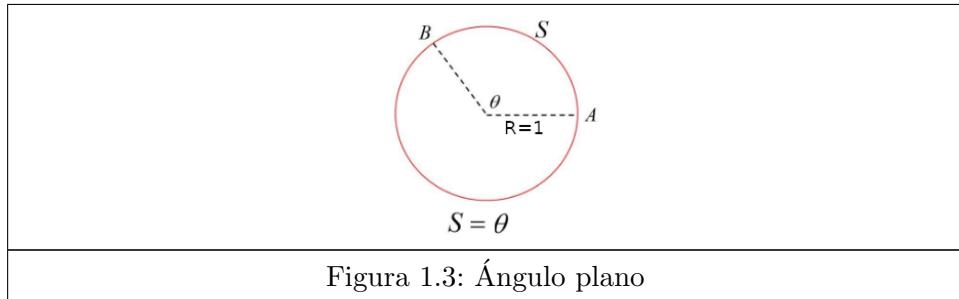
Antes de presentar las magnitudes radiométricas centrales en renderización, vamos a definir el concepto de *ángulo sólido*.

Ángulo sólido

Como ya sabemos, en una circunferencia de radio unidad, el ángulo en radianes entre dos radios es igual a la longitud de circunferencia comprendida entre ellos (figura 1.3). Conceptualmente, el ángulo sólido se trata de una extensión de los ángulos planos a \mathbb{R}^3 , y mide el tamaño con el que se ve

un objeto desde un punto. Podemos entender un objeto como una superficie simple, regular a trozos y orientable.

Definición 1.1. Sea $P \in \mathbb{R}^3$ y consideremos la esfera unidad con centro P , $\mathcal{B}(P, 1)$. Sea $\pi : \mathbb{R}^3/\{P\} \rightarrow \mathcal{B}(P, 1)$ la proyección sobre la esfera $\mathcal{B}(P, 1)$, es decir, $\pi(r) = P + \frac{r-P}{\|r-P\|}$. Sea $A \subseteq \mathbb{R}^3/\{P\}$ una superficie simple, regular a trozos y orientable, y tomamos $I = \int_{\pi(A)} dS$. Entonces diremos que el ángulo sólido subtendido por A desde P mide I estereoradianes.



En la definición anterior hemos usado dS para denotar la integral de superficie. Como vemos el ángulo sólido subtendido por un objeto claramente depende del punto de referencia P . Sin embargo, supongamos que queremos medir el ángulo sólido subtendido por un objeto desde un punto P . En renderización se trabaja con \mathbb{R}^3 como espacio afín, por lo que podemos aplicar una traslación del marco de referencia para situar P en el origen de coordenadas. Por tanto, a efectos prácticos siempre podemos suponer que trabajamos en la esfera unidad \mathbb{S}^2 . Vamos a definir ahora la función que mide el ángulo sólido en la esfera unidad:

Definición 1.2. Sea \mathcal{S} la familia de superficies contenidas en la esfera unidad S^2 , definimos la función $\mu : \mathcal{S} \rightarrow \mathbb{R}_0^+$ como:

$$\mu(C) = \int_C dS, \quad \forall C \in \mathcal{S}$$

La función μ mide el ángulo sólido subtendido por cada superficie simple de la esfera.

Vamos ahora a derivar una serie de resultados básicos relacionados con el ángulo sólido que nos serán útiles durante el resto del trabajo. Consideremos la parametrización de la esfera unidad $\Theta : [0, \pi] \times [0, 2\pi] \rightarrow \mathbb{R}^3$, con:

$$\Theta(\theta, \varphi) = (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta), \quad \forall (\theta, \varphi) \in [0, \pi] \times [0, 2\pi]$$

Derivando parcialmente Θ y calculando el producto vectorial de sus derivadas parciales vemos que:

$$\Theta_\theta(\theta, \varphi) = \frac{\partial \Theta}{\partial \theta}(\theta, \varphi) = (\cos \theta \cos \varphi, \cos \theta \sin \varphi, -\sin \theta)$$

$$\Theta_\varphi(\theta, \varphi) = \frac{\partial \Theta}{\partial \varphi}(\theta, \varphi) = (-\sin \theta \sin \varphi, \sin \theta \cos \varphi, 0)$$

$$\Theta_\theta(\theta, \varphi) \times \Theta_\varphi(\theta, \varphi) = (\sin^2 \theta \cos \varphi, \sin^2 \theta \sin \varphi, \sin \theta \cos \theta)$$

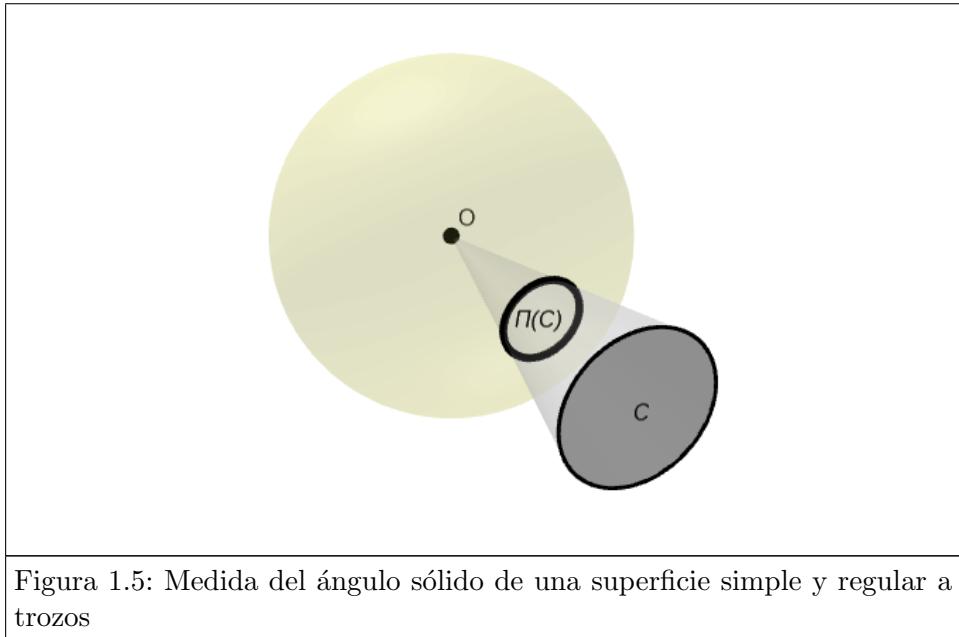
Con lo que concluimos que:

$$\begin{aligned} \mu(C) &= \int_C dS = \iint_{\Theta^{-1}(C)} \|\Theta_\theta(\theta, \varphi) \times \Theta_\varphi(\theta, \varphi)\| d\theta d\varphi \\ &= \iint_{\Theta^{-1}(C)} \sin \theta d\theta d\varphi \quad \forall C \in \mathcal{S} \end{aligned} \tag{1.1}$$

De aquí en adelante, por comodidad, notaremos por π a la proyección sobre la esfera unidad S^2 , es decir, $\pi(p) = \frac{p}{\|p\|}$, $\forall p \in \mathbb{R}^3 / \{(0, 0, 0)\}$. Pasamos ahora a demostrar la siguiente proposición:

Proposición 1.1. Sea μ la medida del ángulo sólido, y sea C una superficie en $\mathbb{R}^3 / \{(0, 0, 0)\}$ simple, regular a trozos y orientable, tal que cada recta que empiece en el origen y acabe en un punto de C sólo interseccione con C una vez. Consideramos la función $g : \mathbb{R}^3 / \{(0, 0, 0)\} \rightarrow \mathbb{R}^3$, con $g(p) = \frac{p}{\|p\|^3}$. Entonces se cumple que:

$$\mu(\pi(C)) = \int_C g \cdot dS \tag{1.2}$$



Notación 1.1. Notaremos por $u \cdot v$ al producto escalar de los vectores u y v .

*Demuestra*ción. Sea C_r la porción del espacio comprendida entre C y $\pi(C)$, y sea ∂C_r su frontera. Sea $C_p = \partial C_r / (C \cup \pi(C))$ la superficie que delimita lateralmente la región C_r . Vamos ahora a calcular la divergencia de g . Sea $(x, y, z) \in \mathbb{R}^3$:

$$\begin{aligned} \operatorname{div} g(x, y, z) &= \frac{\partial g}{\partial x}(x, y, z) + \frac{\partial g}{\partial y}(x, y, z) + \frac{\partial g}{\partial z}(x, y, z) = \\ &= \frac{(x^2 + y^2 + z^2)^{\frac{3}{2}} - 3x^2(x^2 + y^2 + z^2)^{\frac{1}{2}}}{(x^2 + y^2 + z^2)^3} + \\ &\quad + \frac{(x^2 + y^2 + z^2)^{\frac{3}{2}} - 3y^2(x^2 + y^2 + z^2)^{\frac{1}{2}}}{(x^2 + y^2 + z^2)^3} + \\ &\quad + \frac{(x^2 + y^2 + z^2)^{\frac{3}{2}} - 3z^2(x^2 + y^2 + z^2)^{\frac{1}{2}}}{(x^2 + y^2 + z^2)^3} = 0 \end{aligned}$$

Y, por tanto, aplicando el teorema de divergencia tenemos que:

$$\int_{\partial C_r} g \, dS = \iiint_{C_r} \operatorname{div} g \, dx dy dz = 0$$

Sea $C_p = \bigcup_{i=1}^n S_i$ una descomposición finita de C_p como unión de su-

perficies simples y regulares. Fijado $i \in \{1 \dots n\}$, sea $\Psi^i : W^i \rightarrow \mathbb{R}^3$ una parametrización simple y suave de S_i orientada mediante la normal exterior, y sean Ψ_u^i, Ψ_v^i sus derivadas parciales. Es fácil ver que dado un punto $P \in S_i$, se tiene que el vector que va desde el origen hasta P es perpendicular a la normal a S_i en P . Por tanto tenemos que:

$$\int_{S_i} g \, dS = \iint_{W^i} \frac{\Psi^i(u, v) \cdot (\Psi_u^i(u, v) \times \Psi_v^i(u, v))}{\|\Psi^i(u, v)\|^3} \, dudv = 0$$

Dónde se ha usado que $\Psi^i(u, v) \cdot (\Psi_u^i(u, v) \times \Psi_v^i(u, v)) = 0, \forall (u, v) \in W^i$, ya que el producto vectorial de las derivadas parciales es igual a la normal a la superficie y el producto escalar de vectores perpendiculares es nulo. Sabemos que $\partial C_r = C \cup \pi(C) \cup C_p$, y usando las dos igualdades anteriores tenemos que:

$$0 = \int_{\partial C_r} g \, dS = \int_C g \, dS + \int_{\pi(C)} g \, dS$$

Sea $\Phi : W \rightarrow \mathbb{R}^3$, una parametrización simple y suave de $\pi(C)$ orientada con la normal exterior, y sean Φ_u, Φ_v sus derivadas parciales. Como $\pi(C)$ está contenido en la esfera unidad, para todo $(u, v) \in W$ se cumple que $-\Phi(u, v) = \frac{\Phi_u(u, v) \times \Phi_v(u, v)}{\|\Phi_u(u, v) \times \Phi_v(u, v)\|}$ y que $\|\Phi(u, v)\| = 1$, y por ello:

$$\begin{aligned} \int_{\pi(C)} g \, dS &= \iint_W \frac{\Phi(u, v) \cdot (\Phi_u(u, v) \times \Phi_v(u, v))}{\|\Phi(u, v)\|^3} \, dudv = \\ &= - \iint_W \frac{\|\Phi_u(u, v) \times \Phi_v(u, v)\|}{\|\Phi(u, v)\|^2} \, dudv = - \int_{\pi(C)} dS \end{aligned}$$

Y por tanto deducimos que:

$$\begin{aligned} 0 &= \int_C g \, dS + \int_{\pi(C)} g \, dS = \int_C g \, dS - \int_{\pi(C)} dS \Rightarrow \\ \Rightarrow \int_C g \, dS &= \int_{\pi(C)} dS = \mu(\pi(C)) \end{aligned}$$

Tal y como queríamos.

■

Podemos ahora introducir un concepto bastante natural por la forma en que hemos definido la función μ .

Definición 1.3. Sea $C \subseteq \mathbb{S}^2$ una superficie simple y sea $h : C \rightarrow \mathbb{R}$ un campo escalar, $i : C \rightarrow \mathbb{R}^3$ un campo vectorial. Entonces la integral respecto al ángulo sólido de las funciones h e i se define como:

$$\int_C h \, d\mu = \int_C h \, dS$$

$$\int_C i \, d\mu = \int_C i \, dS$$

Y por último podemos generalizar la proposición 1.1 de la siguiente manera:

Proposición 1.2. Sea C una superficie en $\mathbb{R}^3/\{(0,0,0)\}$ simple, regular a trozos y orientable, tal que cada recta que empiece en el origen y acabe en un punto de C sólo interseccione con C una vez. Sea $h : \mathbb{S}^2 \rightarrow \mathbb{R}_0^+$ un campo escalar no negativo. Consideramos la función $g : \mathbb{R}^3/\{(0,0,0)\} \rightarrow \mathbb{R}^3$, con $g(p) = \frac{p}{\|p\|^3}$. Entonces se cumple que:

$$\int_{\pi(C)} h \, d\mu = \int_C (h \circ \pi) g \, dS \quad (1.3)$$

Medida de la luz

Una vez nos hemos familiarizado con el concepto de ángulo sólido, vamos a presentar una serie de magnitudes radiométricas básicas. Todas estas magnitudes dependen de la longitud de onda, y aunque obviaremos esto, es importante tenerlo en cuenta.

- **Energía radiante:** Las fuentes de luz emiten fotones, cada uno de los cuales porta una cantidad de energía que depende de la longitud de onda. La energía se mide en julios (J), y notaremos por Q a la función que mide la energía en una región durante un periodo de tiempo.
- **Flujo radiante:** La energía está medida sobre algún periodo de tiempo, y por tanto, bajo la asunción de estado de reposo, es más conveniente trabajar con magnitudes que se midan en un instante de tiempo. El flujo radiante se define como la cantidad total de energía por unidad de tiempo que atraviesa una región. Se mide en vatios (W), y notaremos por Φ a la función que mide el flujo radiante en una región.
- **Irradiancia y emitancia radiante:** Las dos magnitudes miden el flujo radiante por unidad de superficie, en el caso de la irradiancia se mide el flujo llegando a la superficie, y en el caso de la emitancia

radiante se mide el flujo saliendo de la superficie. Se mide en vatios por metro cuadrado (W/m^{-2}) y notaremos por E a la función que mide la irradiancia.

- **Radiancia incidente y saliente:** Se trata de la magnitud que más utilizaremos. La irradiancia nos da una medida del flujo radiante por unidad de superficie, pero no tiene en cuenta que el flujo tiene una distribución direccional, es decir, la cantidad de flujo que llega a una superficie varía según la dirección del flujo incidente. La radiancia mide el flujo por unidad de superficie proyectada y por unidad de ángulo sólido. Cuando hablamos de superficie proyectada nos referimos a que calculamos el flujo en una superficie hipotética que es perpendicular a la dirección del flujo incidente/saliente. Se mide en vatios por estereorradián por metro cuadrado ($\text{W sr}^{-1} \text{ m}^{-2}$) y notaremos por $L : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}_0^+$ a la función que mide la radiancia.

Notemos por $\mathcal{H}^2(n)$ al hemisferio superior de la esfera unidad cuyo polo norte es el vector n . Si consideramos una superficie con normal n en un punto p , la radiancia generalmente no es continua en p . Por tanto tiene sentido distinguir entre la función radiancia encima de la superficie y debajo de ella:

$$\begin{aligned} L^+(p, \omega) &= \lim_{t \rightarrow 0^+} L(p + tn, \omega) \\ L^-(p, \omega) &= \lim_{t \rightarrow 0^-} L(p - tn, \omega) \end{aligned}$$

Podemos definir por tanto las funciones de radiancia incidente y radiancia saliente como sigue (siempre consideramos que la dirección ω apunta hacia fuera de la superficie, aunque sea radiancia incidente):

$$L_i(p, \omega) = \begin{cases} L^+(p, -\omega), & \omega \cdot n > 0 \\ L^-(p, -\omega), & \omega \cdot n < 0 \end{cases}$$

$$L_s(p, \omega) = \begin{cases} L^+(p, \omega), & \omega \cdot n > 0 \\ L^-(p, \omega), & \omega \cdot n < 0 \end{cases}$$

Por como están definidas las magnitudes anteriores tenemos las siguientes relaciones entre ellas:

- La irradiancia en un punto $p \in \mathbb{R}^3$ perteneciente a una superficie con normal n en p cumple que:

$$E(p) = \int_{\mathbb{S}^2} L_i(p, \omega) |n \cdot \omega| d\mu(\omega)$$

Dónde se multiplica por $|n \cdot \omega|$ debido al hecho de que la radiancia se mide respecto a la superficie proyectada. Notar que $|n \cdot \omega|$ es igual al valor absoluto del coseno del ángulo que forman n y ω .

- El flujo en una superficie $A \subseteq \mathbb{R}^3$ cumple que:

$$\Phi(A) = \int_A E \, dS$$

Por último, la radiancia saliente de una superficie en un punto $p \in \mathbb{R}^3$ y en una dirección $\omega_o \in \mathbb{S}^2$ a causa de la iluminación incidente en p se puede calcular como sigue, con n la normal a la superficie en p :

$$L_s(p, \omega_o) = \int_{\mathbb{S}^2} f(p, \omega_o, \omega_i) L_i(p, \omega_i) |n \cdot \omega_i| \, d\mu(\omega_i) \quad (1.4)$$

donde f es la *BDSF* descrita en el apartado anterior. Esta ecuación se conoce como ecuación de dispersión. Si además la superficie es emisiva en el punto p , tenemos que la radiancia saliente total en el punto $p \in \mathbb{R}^3$ y en una dirección $\omega_o \in \mathbb{S}^2$ se calcula como:

$$L_s(p, \omega_o) = L_e(p, \omega_o) + \int_{\mathbb{S}^2} f(p, \omega_o, \omega_i) L_i(p, \omega_i) |n \cdot \omega_i| \, d\mu(\omega_i) \quad (1.5)$$

dónde la función $L_e(p, \omega_o)$ describe la radiancia emitida en el punto p y en la dirección ω_o . Esta ecuación se denomina *ecuación de renderización*, o *rendering equation*, y es fundamental en ray-tracing, ya que es la ecuación utilizada para estimar el color de los puntos con los que interseccionan los rayos trazados por la cámara, estimando el valor de radiancia saliente en dirección a la cámara en el punto de intersección. La complejidad de la integral dependerá de la escena que estemos renderizando, y como podemos intuir no tiene una solución analítica. Hay diferentes formas de abordar el problema de aproximar la ecuación de renderización, como veremos en el capítulo 4.

Hay que destacar que el algoritmo de ray-tracing tiene una naturaleza recursiva, ya que con el objetivo de estimar el color de un punto, se trazarán rayos desde ese punto en direcciones dónde el valor de radiancia incidente o de BDSF sea grande. Dicho rayo puede volver a interseccionar con otro componente de la escena, del cuál tendremos que estimar a su vez la radiancia saliente en dirección al punto inicial.

1.1.6. Propagación de los rayos

En todo lo descrito con anterioridad hemos supuesto que los rayos se propagan por el vacío, por lo que la energía radiante que viaja por ellos no

se ve atenuada. Sin embargo en la vida real ciertos componentes ambientales como la niebla o el humo invalidan dicha suposición. Es por ello que es habitual que los ray tracers simulen este tipo de efectos ambientales, dónde la energía radiante se ve atenuada al viajar por el espacio en presencia de partículas en suspensión. No obstante, en lo referente a los objetivos de este trabajo podemos obviar este tipo de efectos.

1.2. Concepto de método de Monte Carlo y su relación con el ray-tracing

Los métodos de Monte Carlo son un conjunto de algoritmos basados en el muestreo aleatorio. Sus objetivos son:

- Sea (Γ, \mathcal{A}, P) un espacio probabilístico, sea $X : \Gamma \rightarrow \mathbb{R}^n$ una variable o un vector aleatorio y sea P_X su distribución de probabilidad, generar un conjunto de muestras $\{x^{(r)}\}_{r=1}^R$ independientes que sigan dicha distribución.
- Estimar la esperanza de determinadas funciones bajo dicha distribución, es decir, sea g una función medible definida sobre $(\mathbb{R}^n, \mathcal{B}^n)$, estimar el valor de:

$$E[g(X)] = \int g(s) dP(s)$$

Vemos que una vez resolvamos el primer problema, podemos usar el siguiente estimador para resolver el segundo:

$$\hat{g} = \frac{1}{R} \sum_{r=1}^R g(x^{(r)}) \quad (1.6)$$

Ya que es claro que si las muestras $\{x^{(r)}\}_{r=1}^R$ son independientes y siguen la distribución P_X , entonces la esperanza del estimador cumple que:

$$E[\hat{g}] = E\left[\frac{1}{R} \sum_{r=1}^R g(x^{(r)})\right] = \frac{1}{R} \sum_{r=1}^R E[g(x^{(r)})] = \frac{1}{R} R E[g(X)] = E[g(X)]$$

Al estimador definido en 1.6 lo llamaremos estimador Monte Carlo de g . Habitualmente trabajaremos con vectores aleatorios continuos, así que salvo que se especifique lo contrario, así lo supondremos. Supongamos ahora que queremos estimar la integral de una función $h : K \rightarrow \mathbb{R}$ en lugar de su esperanza. Sea $X : \Gamma \rightarrow K$ un vector aleatorio tal que su función de densidad f_X cumple que $f_X(s) \neq 0$ para todo s tal que $|h(s)| > 0$, y sea

$\{x^{(r)}\}_{r=1}^R$ un conjunto de muestras independientes siguiendo la distribución de probabilidad de X . Entonces para aproximar la integral de h podemos utilizar el estimador Monte Carlo de la función $g = \frac{h}{f_X}$:

$$\hat{g} = \frac{1}{R} \sum_{r=1}^R \frac{h(x^{(r)})}{f_X(x^{(r)})}$$

En efecto la esperanza de \hat{g} es igual a la integral de h en K :

$$\begin{aligned} E[\hat{g}] &= E\left[\frac{1}{R} \sum_{r=1}^R \frac{h(x^{(r)})}{f_X(x^{(r)})}\right] = \frac{1}{R} \sum_{r=1}^R E\left[\frac{h(x^{(r)})}{f_X(x^{(r)})}\right] = \\ &= \frac{R}{R} \int_K \frac{h(x)}{f_X(x)} f_X(x) dx = \int_K h(x) dx \end{aligned}$$

La elección de la distribución de probabilidad respecto a la que tomamos las muestras es clave, ya que seleccionar una distribución de probabilidad que asigne mayores probabilidades a los puntos donde la función g es mayor supone una importante técnica para reducir la varianza del estimador.

Los métodos de Monte Carlo son una importante herramienta dentro del ray-tracing. El motivo de que estos métodos sean más efectivos en renderización que otras técnicas de integración numérica es que la varianza del estimador es independiente de la dimensionalidad del espacio muestreado. En efecto, sean $\{x^{(r)}\}_{r=1}^R$ un conjunto de muestras independientes siguiendo la distribución de probabilidad de X , y supongamos que $E[g(X)^2] < +\infty$. Entonces tenemos que:

$$\begin{aligned} \text{Var}(\hat{g}) &= \text{Var}\left(\frac{1}{R} \sum_{r=1}^R g(x^{(r)})\right) = \frac{1}{R^2} \text{Var}\left(\sum_{r=1}^R g(x^{(r)})\right) \\ &= \frac{1}{R^2} \sum_{r=1}^R \text{Var}(g(x^{(r)})) = \frac{1}{R} \text{Var}(g(X)) \end{aligned}$$

Esto muestra que la varianza del estimador decrece con un ratio de $\frac{1}{R}$, donde R es el número de muestras.

Estos métodos son ampliamente utilizados dentro de los renderizadores fotorrealistas, y su aplicación más importante en este ámbito es permitirnos estimar el valor de la ecuación de renderización (1.5).

1.3. Problema a resolver

En un ray-tracer, la cantidad de operaciones necesarias para calcular una estimación del color de un píxel es muy alta, ya que conlleva trazar numerosos rayos y calcular sus intersecciones con la escena. Dado que habrá miles de píxeles en la imagen final, es esencial obtener buenos resultados sin necesidad de tener que incrementar en gran medida el número de muestras tomadas por cada píxel, lo cual se consigue mediante una reducción de la varianza de los estimadores usados. Por ello, el principal problema a resolver es el estudio de métodos que reduzcan la varianza de los estimadores Monte Carlo usados en ray-tracing para aproximar la ecuación de renderización.

Para poder abordar este problema se requiere resolver una cuestión previa, realizar un estudio de los métodos de Monte Carlo más utilizados en ray-tracing, lo cuál se abordará desde un punto de vista matemático y formal.

1.4. Contenido de la memoria y principales fuentes y herramientas empleadas

El trabajo realizado está dividido en dos partes, si bien ambas están estrechamente relacionadas. La primera parte, enfocada a los aspectos más matemáticos, asume ciertos conocimientos en probabilidad y teoría de la medida. Esta parte comprende dos capítulos:

- En el capítulo 3 estudiaremos los principales métodos de Monte Carlo. Iniciaremos el capítulo detallando ciertos conceptos básicos relacionados con procesos estocásticos. A continuación presentaremos y demostraremos una serie de resultados que nos llevarán a la ley fuerte de los números grandes, el teorema que nos asegura la convergencia del estimador Monte Carlo. Detallaremos una serie de métodos de Monte Carlo para generar muestras y finalizaremos el capítulo presentando tres métodos de reducción de la varianza en estimadores de Monte Carlo. Este apartado sigue algunas de las ideas detalladas en [GT13], aunque también se han consultado demostraciones alternativas en [Wil91]. También se ha consultado [Vea98] para los métodos de reducción de varianza y [RR04] para ciertos aspectos relacionados con los procesos de Markov y el método de Metrópolis.
- En el capítulo 4 se detallan las dos formas más elementales de aproximar la ecuación de renderización, la iluminación directa y el algoritmo de *path tracing*. Estas aproximaciones se basan en el uso de estimadores de Monte Carlo. En este capítulo la fuente principal es [PJH16], y

1.4. Contenido de la memoria y principales fuentes y herramientas empleadas

para la fundamentación matemática del algoritmo de path tracing se ha consultado el artículo [\[DGP05\]](#).

En la parte relacionada con informática, presentaremos y estudiaremos tres algoritmos que nos permiten tomar muestras respecto a distribuciones de probabilidad sobre el ángulo sólido subtendido por una fuente de luz de área, distintas a las distribuciones utilizadas en los métodos de muestreo clásicos. Estos algoritmos están enfocados a reducir la varianza de los estimadores de la ecuación de renderización. Los algoritmos están recogidos en los artículos [\[UfK13\]](#) [\[nG18\]](#) [\[GUk+17\]](#), y serán implementados en la versión 3 del renderizador de código abierto *pbrt*, cuyo código es accesible a través de [\[PJH\]](#). La versión de pbrt con las implementaciones descritas en este trabajo es accesible a través de [este repositorio de github](#). Se realizarán pruebas para mostrar los resultados obtenidos con cada implementación.

Por último, destacar las fuentes [\[Shi20a\]](#), [\[Shi20b\]](#) y [\[Shi20c\]](#), que fueron esenciales como introducción en el ámbito de los ray tracers. [\[PJH16\]](#) también ha sido fundamental, tanto por ser una guía de uso del renderizador pbrt, como por detallar muchos de los conceptos y algoritmos más relevantes en ray-tracing.

Capítulo 2

Objetivos

Los objetivos inicialmente previstos fueron los siguientes:

- Realizar una revisión bibliográfica de los métodos de Monte Carlo.
- Investigar técnicas de reducción de la varianza en métodos de Monte Carlo.
- Realizar un análisis, implementación y pruebas de software de algoritmos descritos en la literatura basados en métodos de Monte Carlo para la síntesis de imágenes realistas.
- Analizar posibilidades de mejora de los algoritmos en eficiencia o reducción de varianza.

Los primeros dos objetivos han sido alcanzados en el capítulo 3. Por otro lado, aunque inicialmente no era un objetivo del trabajo, en el capítulo 4 se han analizado desde un punto de vista matemático las dos formas más elementales de aproximar la ecuación de renderización, ya que de esta manera se pone de manifiesto la utilidad de los métodos de Monte Carlo en la síntesis de imágenes realistas. El tercer objetivo se ha alcanzado en el capítulo 5, mientras que no ha sido posible abordar el último objetivo.

Parte I

Matemáticas

Capítulo 3

Métodos de Monte Carlo

Al final del capítulo anterior se mencionaron los problemas a los que pretenden poner solución los métodos de Monte Carlo. En este capítulo se demostrará la convergencia del estimador de Monte Carlo y se describirán diversos métodos de Monte Carlo con aplicación en renderización. Por último veremos tres métodos de reducción de varianza en el estimador Monte Carlo.

3.1. Conceptos básicos

Vamos a presentar una serie de conceptos que nos serán útiles posteriormente. Empezaremos por describir el concepto de proceso estocástico.

Notación 3.1. Notaremos por \mathcal{B}^n a la σ -álgebra de Borel en \mathbb{R}^n .

Definición 3.1 (Proceso estocástico). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Un proceso estocástico es una familia de variables o vectores aleatorios $\{X_t\}_{t \in T}$ definidos sobre (Γ, \mathcal{A}, P) con valores en un espacio de Borel (E, \mathcal{B}_E) , con $E \subseteq \mathbb{R}^n$. El conjunto T se denomina espacio paramétrico y es un conjunto ordenado arbitrario, y el espacio (E, \mathcal{B}_E) se denomina espacio de estados.*

Definición 3.2. *Sea $\{X_t\}_{t \in T}$ un proceso estocástico. Diremos que $\{X_t\}_{t \in T}$ es un proceso estocástico en tiempo discreto si T es un conjunto discreto. Diremos que $\{X_t\}_{t \in T}$ es un proceso estocástico en tiempo continuo si T es un conjunto no numerable. Diremos que $\{X_t\}_{t \in T}$ es un proceso estocástico real si el espacio de estados (E, \mathcal{B}_E) cumple que $E \subseteq \mathbb{R}$, es decir, si $\{X_t\}_{t \in T}$ es una sucesión de variables aleatorias.*

Pasamos ahora a definir el concepto de esperanza y probabilidad condicionada.

Definición 3.3 (Probabilidad restringida). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico y $\mathcal{D} \subseteq \mathcal{A}$ una σ -álgebra. La restricción de P a la σ -álgebra \mathcal{D} , denotada por $P_{\mathcal{D}}$, se define por*

$$P_{\mathcal{D}}(S) = P(S), \quad \forall S \in \mathcal{D}$$

y es una medida de probabilidad sobre (Γ, \mathcal{D}) .

Definición 3.4 (Propiedad casi segura). *Sea Γ un espacio muestral y sea \mathcal{A} una σ -álgebra sobre Γ . Sea P una medida de probabilidad sobre (Γ, \mathcal{A}) . Se dice que una propiedad se cumple casi seguramente respecto a P , y lo notamos por c.s.- P , si se cumple para todo conjunto $S \in \mathcal{A}$ tal que $P(S) > 0$.*

Definición 3.5 (P-equivalencias). *Sea Γ un espacio muestral y sea \mathcal{A} una σ -álgebra sobre Γ . Sea P una medida de probabilidad sobre (Γ, \mathcal{A}) . Se dice que dos funciones medibles $X : (\Gamma, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}^n)$, $Y : (\Gamma, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}^n)$ son P -equivalentes si $X = Y$ c.s.- P .*

Definición 3.6 (Esperanza de una variable aleatoria condicionada a una σ -álgebra). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. La esperanza condicionada de una variable aleatoria integrable $X : (\Gamma, \mathcal{A}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ a una σ -álgebra $\mathcal{D} \subseteq \mathcal{A}$ se nota por $E[X/\mathcal{D}]$ y es la única función \mathcal{D} -medible, salvo $P_{\mathcal{D}}$ -equivalencias, que verifica:*

$$\int_S E[X/\mathcal{D}] \, dP_{\mathcal{D}} = \int_S X \, dP, \quad \forall S \in \mathcal{D} \quad (3.1)$$

Definición 3.7 (Probabilidad condicionada a una σ -álgebra). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. La probabilidad condicionada de un suceso $S \in \mathcal{A}$ a una σ -álgebra $\mathcal{D} \subseteq \mathcal{A}$ se nota por $P(S/\mathcal{D})$ y se define como:*

$$P(S/\mathcal{D}) = E[\mathbb{1}_S/\mathcal{D}], \quad \text{c.s.-}P_{\mathcal{D}}$$

donde $\mathbb{1}_S$ es la función indicadora del conjunto S .

Definición 3.8 (σ -álgebra generada por un vector aleatorio). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico y sea $X : (\Gamma, \mathcal{A}, P) \rightarrow (\mathbb{R}^n, \mathcal{B}^n)$ un vector aleatorio. La σ -álgebra generada por $X = (X_1, \dots, X_n)$ se nota por $\sigma(X) = \sigma(X_1, \dots, X_n)$ y se define por $\sigma(X) = \{X^{-1}(B)/B \in \mathcal{B}^n\}$. $\sigma(X)$ es la mínima σ -álgebra que hace a X medible.*

Definición 3.9 (Esperanza de una variable aleatoria condicionada a un vector aleatorio). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea X una variable aleatoria integrable, y sea $Y = (Y_1, \dots, Y_n)$ un vector aleatorio, ambos definidos sobre (Γ, \mathcal{A}, P) . La esperanza condicionada de X al vector aleatorio Y se nota por $E[X/Y_1, \dots, Y_n]$ y se define por:*

$$E[X/Y_1, \dots, Y_n] = E[X/\sigma(Y_1, \dots, Y_n)], \quad \text{c.s.-}P_{\sigma(Y_1, \dots, Y_n)}$$

Proposición 3.1. $E[X/Y_1, \dots, Y_n]$ es una función de Y_1, \dots, Y_n .

Definición 3.10 (Probabilidad condicionada a un vector aleatorio). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico y sea $Y = (Y_1, \dots, Y_n)$ un vector aleatorio definido sobre (Γ, \mathcal{A}, P) . La probabilidad condicionada de un suceso $S \in \mathcal{A}$ al vector aleatorio Y se nota por $P(S/Y_1, \dots, Y_n)$ y se define por:*

$$P(S/Y_1, \dots, Y_n) = E[\mathbb{1}_S/Y_1, \dots, Y_n], \text{ c.s.-}P_{\sigma(Y_1, \dots, Y_n)}$$

donde $\mathbb{1}_S$ es la función indicadora del conjunto S .

Pasamos a ver una serie de propiedades de la esperanza condicionada que utilizaremos en demostraciones posteriores. Antes de enunciar dichas propiedades necesitamos definir el concepto de independencia de σ -álgebras.

Definición 3.11 (Independencia de σ -álgebras). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sean $S_1, S_2 \subseteq \mathcal{A}$ dos σ -álgebras y sea $Y = (Y_1, \dots, Y_n)$ un vector aleatorio definido sobre (Γ, \mathcal{A}, P) . Diremos que S_1 y S_2 son independientes si cumplen que:*

$$\forall s_1 \in S_1, \forall s_2 \in S_2, P(s_1 \cap s_2) = P(s_1)P(s_2)$$

De igual forma diremos que Y es independiente de S_1 si $\sigma(Y_1, \dots, Y_n)$ es independiente de S_1 .

Proposición 3.2. Sean X, Y variables aleatorias integrables definidas sobre el espacio probabilístico (Γ, \mathcal{A}, P) , sea $\mathcal{D} \subseteq \mathcal{A}$ una σ -álgebra. Entonces:

- (a) Sean $a, b \in \mathbb{R}$, entonces $E[aX+bY/\mathcal{D}] = aE[X/\mathcal{D}]+bE[Y/\mathcal{D}]$, c.s.- $P_{\mathcal{D}}$.
- (b) Si $X \geq 0$ c.s.- $P \Rightarrow E[X/\mathcal{D}] \geq 0$, c.s.- $P_{\mathcal{D}}$.
- (c) Si $X \geq Y$ c.s.- $P \Rightarrow E[X/\mathcal{D}] \geq E[Y/\mathcal{D}]$, c.s.- $P_{\mathcal{D}}$.
- (d) $E[E[X/\mathcal{D}]] = E[X]$.
- (e) Si X es \mathcal{D} -medible $\Rightarrow E[X/\mathcal{D}] = X$, c.s.- $P_{\mathcal{D}}$.
- (f) Sea $\{Z_n\}_{n \in \mathbb{N}}$ una sucesión creciente de variables aleatorias positivas integrables sobre (Γ, \mathcal{A}, P) , que convergen puntualmente a una variable aleatoria integrable Z , entonces:

$$\lim_{n \rightarrow +\infty} E[Z_n/\mathcal{D}] = E[Z/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

- (g) Si XY es integrable y X es \mathcal{D} -medible, $\Rightarrow E[XY/\mathcal{D}] = XE[Y/\mathcal{D}]$, c.s.- $P_{\mathcal{D}}$.

(h) Si X es independiente de $\mathcal{D} \Rightarrow E[X/\mathcal{D}] = E[X]$, c.s.- $P_{\mathcal{D}}$.

(i) Sean $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \mathcal{A}$, entonces:

$$E[E[X/\mathcal{D}_2]/\mathcal{D}_1] = E[E[X/\mathcal{D}_1]/\mathcal{D}_2] = E[X/\mathcal{D}_1], \text{ c.s.-}P_{\mathcal{D}_1}$$

(j) Sea $g : \mathbb{R} \rightarrow \mathbb{R}$ una función continua y convexa, tal que $E[g(X)] < \infty$, entonces:

$$g(E[X/\mathcal{D}]) \leq E[g(X)/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

(k) $|E[X/\mathcal{D}]| \leq E[|X|/\mathcal{D}]$, c.s.- $P_{\mathcal{D}}$

Demostración. Las propiedades (a), (b), (d), (e), (h) e (i) se demuestran fácilmente usando la definición de esperanza condicionada (definición 3.6). Por otro lado, la propiedad (c) se deduce fácilmente de las propiedades (a) y (b), así como la propiedad (k) se deduce de la propiedad (c). Nos centramos por tanto en las propiedades (f), (g) y (j).

Demostración de la propiedad (f):

Tomamos $Y_n = Z - Z_n$. Dado que $Y_n \geq 0$ c.s.- P , por la propiedad (b), $E[Y_n/\mathcal{D}] \geq 0$ c.s.- $P_{\mathcal{D}}$ para todo $n \in \mathbb{N}$. Además, por ser $\{Y_n\}_{n \in \mathbb{N}}$ decreciente, por la propiedad (c), se cumple que $\{E[Y_n/\mathcal{D}]\}_{n \in \mathbb{N}}$ también es decreciente. Al ser decreciente y positiva, la sucesión $\{E[Y_n/\mathcal{D}]\}_{n \in \mathbb{N}}$ converge puntualmente. Sea α el límite puntual de $\{E[Y_n/\mathcal{D}]\}_{n \in \mathbb{N}}$, es decir, $\alpha(\gamma) := \lim_{n \rightarrow +\infty} E[Y_n/\mathcal{D}](\gamma)$, $\forall \gamma \in \Gamma$. Queremos demostrar por tanto que $\alpha = 0$ c.s.- $P_{\mathcal{D}}$.

Es fácil ver que por el teorema de la convergencia monótona se tiene que $\lim_{n \rightarrow +\infty} E[Z_n] = E[Z]$. Deducimos que, usando la propiedad (d):

$$\lim_{n \rightarrow +\infty} E[E[Y_n/\mathcal{D}]] = \lim_{n \rightarrow +\infty} E[Y_n] = E[Z] - \lim_{n \rightarrow +\infty} E[Z_n] = 0$$

Y por tanto, usando el teorema de la convergencia dominada, se tiene que $E[\alpha] = \lim_{n \rightarrow +\infty} E[E[Y_n/\mathcal{D}]] = 0$. Dado que $\alpha \geq 0$ c.s.- $P_{\mathcal{D}}$ por ser límite de funciones positivas, y teniendo que $E[\alpha] = 0$, deducimos que $\alpha = 0$ c.s.- $P_{\mathcal{D}}$. Por último:

$$0 = \alpha = \lim_{n \rightarrow +\infty} E[Y_n/\mathcal{D}] = E[Z/\mathcal{D}] - \lim_{n \rightarrow +\infty} E[Z_n/\mathcal{D}] \text{ c.s.-}P_{\mathcal{D}}$$

⇓

$$\lim_{n \rightarrow +\infty} E[Z_n/\mathcal{D}] = E[Z/\mathcal{D}] \text{ c.s.-}P_{\mathcal{D}}$$

Como queríamos.

Demostración de la propiedad (g):

En primer lugar lo demostraremos para funciones indicadoras. Sea $A \in \mathcal{D}$, y sea $\mathbb{1}_A$ la función indicadora de A . Por definición, la esperanza condicionada es la única función \mathcal{D} -medible salvo $P_{\mathcal{D}}$ equivalencias que verifica 3.1. Por tanto:

$$\int_S E[Y\mathbb{1}_A/\mathcal{D}] dP_{\mathcal{D}} = \int_S Y\mathbb{1}_A dP, \forall S \in \mathcal{D}$$

Sea $S \in \mathcal{D}$, es claro que $S \cap A \in \mathcal{D}$, y vemos que:

$$\int_S Y\mathbb{1}_A dP = \int_{S \cap A} Y dP = \int_{S \cap A} E[Y/\mathcal{D}] dP_{\mathcal{D}} = \int_S \mathbb{1}_A E[Y/\mathcal{D}] dP_{\mathcal{D}}, \forall S \in \mathcal{D}$$

Dado que $\mathbb{1}_A E[Y/\mathcal{D}]$ es \mathcal{D} -medible por ser producto de funciones \mathcal{D} -medibles, por la unicidad de la definición de esperanza condicionada deducimos que $E[Y\mathbb{1}_A/\mathcal{D}] = \mathbb{1}_A E[Y/\mathcal{D}]$, c.s.- $P_{\mathcal{D}}$, $\forall A \in \mathcal{D}$.

Vamos ahora a demostrarlo para funciones simples, es decir, para funciones de la forma $X = \sum_{i=0}^m a_i \mathbb{1}_{B_i}$, con $B_i \in \mathcal{D}$ para todo i . Vemos que, usando la linealidad (propiedad (a)) y el resultado para funciones indicadoras:

$$E[XY/\mathcal{D}] = \sum_{i=0}^m a_i E[\mathbb{1}_{B_i} Y/\mathcal{D}] = \sum_{i=0}^m a_i \mathbb{1}_{B_i} E[Y/\mathcal{D}] = X E[Y/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

El siguiente paso es demostrar el resultado para funciones medibles positivas. Sea X una función medible positiva integrable definida sobre (Γ, \mathcal{A}, P) . Por el teorema de aproximación de Lebesgue sabemos que existe una sucesión creciente $\{X_n\}_{n \in \mathbb{N}}$ de funciones simples positivas que converge puntualmente hacia X . Vemos que:

$$XE[Y/\mathcal{D}] = \lim_{n \rightarrow +\infty} X_n E[Y/\mathcal{D}] \stackrel{(1)}{=} \lim_{n \rightarrow +\infty} E[X_n Y/\mathcal{D}] \stackrel{(2)}{=} E[XY/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

donde en (1) se ha usado el resultado recién demostrado para funciones simples y en (2) se ha usado la propiedad (f). Por último falta demostrar el resultado para variables aleatorias integrables. Sea X una variable aleatoria integrable. X se puede descomponer como suma de su parte positiva y su parte negativa, notadas por $X^+ = \max\{X, 0\}$, $X^- = -\min\{X, 0\}$, con $X = X^+ - X^-$. Por tanto se tiene que:

$$\begin{aligned} XE[Y/\mathcal{D}] &= X^+ E[Y/\mathcal{D}] - X^- E[Y/\mathcal{D}] \stackrel{(1)}{=} E[X^+ Y/\mathcal{D}] - E[X^- Y/\mathcal{D}] \\ &\stackrel{(2)}{=} E[(X^+ - X^-)Y/\mathcal{D}] = E[XY/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}} \end{aligned}$$

donde en (1) se ha usado el resultado para funciones positivas y en (2) se ha usado la linealidad de la esperanza condicionada.

Demostración de la propiedad (j):

Si $g : \mathbb{R} \rightarrow \mathbb{R}$ es una función continua y convexa, existe una función $h : \mathbb{R} \rightarrow \mathbb{R}$ no decreciente tal que $g(x) - g(y) \geq h(y)(x - y)$, $\forall x, y \in \mathbb{R}$. Aplicando dicha desigualdad a X y $E[X/\mathcal{D}]$:

$$g(X) - g(E[X/\mathcal{D}]) \geq h(E[X/\mathcal{D}])(X - E[X/\mathcal{D}]), \text{ c.s.-}P_{\mathcal{D}}$$

Usando la propiedad (c) y la linealidad (propiedad (a)) tenemos que:

$$E[g(X)/\mathcal{D}] - E[g(E[X/\mathcal{D}])/\mathcal{D}] \geq E[h(E[X/\mathcal{D}])(X - E[X/\mathcal{D}])/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

$g(E[X/\mathcal{D}])$ y $h(E[X/\mathcal{D}])$ son \mathcal{D} -medibles por ser g continua y h no decreciente, por lo que podemos aplicar las propiedades (e) y (g) para obtener:

$$E[g(X)/\mathcal{D}] - g(E[X/\mathcal{D}]) \geq h(E[X/\mathcal{D}])E[(X - E[X/\mathcal{D}])/\mathcal{D}], \text{ c.s.-}P_{\mathcal{D}}$$

Basta ver que:

$$E[(X - E[X/\mathcal{D}])/\mathcal{D}] = E[X/\mathcal{D}] - E[E[X/\mathcal{D}]/\mathcal{D}] = 0 \text{ c.s.-}P_{\mathcal{D}}$$

donde se ha usado la linealidad y la propiedad (e). ■

Las anteriores definiciones de esperanza de una variable aleatoria se pueden extender a vectores aleatorios. En lo siguiente, diremos que un vector aleatorio es integrable si lo son sus componentes.

Definición 3.12 (Esperanza de un vector aleatorio condicionado a una σ -álgebra). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico y sea $X = (X_1, \dots, X_n)$ un vector aleatorio integrable definido sobre (Γ, \mathcal{A}, P) . La esperanza condicionada de X a una σ -álgebra $\mathcal{D} \subseteq \mathcal{A}$ se nota por $E[X/\mathcal{D}]$ y se define por:*

$$E[X/\mathcal{D}] = (E[X_1/\mathcal{D}], \dots, E[X_n/\mathcal{D}]), \text{ c.s.-}P_{\mathcal{D}}$$

Definición 3.13 (Esperanza de un vector aleatorio condicionado a un vector aleatorio). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico y sean $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_n)$ dos vectores aleatorios definidos sobre (Γ, \mathcal{A}, P) . Supongamos además que X es integrable. La esperanza condicionada de X al vector aleatorio Y se nota por $E[X/Y_1, \dots, Y_n]$ y se define por*

$$E[X/Y_1, \dots, Y_n] = E[X/\sigma(Y_1, \dots, Y_n)], \text{ c.s.-}P_{\sigma(Y_1, \dots, Y_n)}$$

Por último veremos dos tipos de procesos estocásticos de gran importancia.

Definición 3.14 (Filtración de σ -álgebras). *Sea (Γ, \mathcal{A}) un espacio medible. Una filtración de σ -álgebras $\{\mathcal{G}_t\}_{t \in T}$ es una sucesión de σ -álgebras cumpliendo que:*

$$\forall s, t \in T, s \leq t, \mathcal{G}_s \subseteq \mathcal{G}_t \subseteq \mathcal{A}$$

Definición 3.15 (Adaptabilidad a una filtración). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_t\}_{t \in T}$ un proceso estocástico definido sobre (Γ, \mathcal{A}, P) y sea $\{\mathcal{G}_t\}_{t \in T}$ una filtración de σ -álgebras. Se dice que $\{X_t\}_{t \in T}$ está adaptado a la filtración $\{\mathcal{G}_t\}_{t \in T}$ si X_t es \mathcal{G}_t -medible, $\forall t \in T$.*

Notación 3.2. $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$

Definición 3.16 (Martingala, submartingala y supermartingala). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico real (en tiempo discreto) definido sobre (Γ, \mathcal{A}, P) , con X_n integrable para todo $n \in \mathbb{N}_0$. Sea $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una filtración de σ -álgebras tal que $\{X_n\}_{n \in \mathbb{N}_0}$ está adaptada a $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$. Entonces diremos que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es una martingala si cumple que:*

$$E[X_{n+1}/\mathcal{G}_n] = X_n, \text{ c.s.-}P_{\mathcal{G}_n}, \forall n \in \mathbb{N}_0$$

Diremos que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es una submartingala si cumple que:

$$E[X_{n+1}/\mathcal{G}_n] \geq X_n, \text{ c.s.-}P_{\mathcal{G}_n}, \forall n \in \mathbb{N}_0 \quad (3.2)$$

Por último, diremos que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es una supermartingala si cumple que:

$$E[X_{n+1}/\mathcal{G}_n] \leq X_n, \text{ c.s.-}P_{\mathcal{G}_n}, \forall n \in \mathbb{N}_0$$

Definición 3.17 (Proceso de Markov). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_t\}_{t \in T}$ un proceso estocástico definido sobre (Γ, \mathcal{A}, P) con espacio de estados (E, \mathcal{B}_E) . Sea $\{\mathcal{G}_t\}_{t \in T}$ una filtración de σ -álgebras tal que $\{X_t\}_{t \in T}$ está adaptada a $\{\mathcal{G}_t\}_{t \in T}$. Entonces diremos que $\{X_t, \mathcal{G}_t\}_{t \in T}$ es un proceso de Markov si cumple que:*

$$\forall s, t \in T, s < t, \forall B \in \mathcal{B}_E, P(X_t \in B / \mathcal{G}_s) = P(X_t \in B / X_s)$$

En el caso de procesos de Markov en tiempo discreto se da la siguiente caracterización.

Proposición 3.3. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con espacio de estados (E, \mathcal{B}_E) . Sea $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una filtración de σ -álgebras tal que $\{X_n\}_{n \in \mathbb{N}_0}$ está adaptada a $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$. Entonces $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es un proceso de Markov si y solo si cumple que:*

$$\forall n \in \mathbb{N}, \forall B \in \mathcal{B}_E, P(X_n \in B / \mathcal{G}_{n-1}) = P(X_n \in B / X_{n-1})$$

3.2. Convergencia del estimador de Monte Carlo

Cuando presentamos el estimador de Monte Carlo, vimos que su esperanza es igual al valor que queremos aproximar, lo cual implica que si la varianza del estimador no es muy grande por lo general obtendremos valores cercanos al buscado. En esta sección, siguiendo las ideas en [GT13], enunciaremos y demostraremos el teorema que nos asegura la convergencia del estimador de Monte Carlo al valor buscado cuando el número de muestras tomadas tiende a infinito. Antes de poder abarcar dicha demostración necesitamos probar una serie de resultados previos.

3.2.1. Tiempos de parada

El concepto de *tiempo de parada* proviene de los juegos de apuestas, y puede ser entendido como una regla que dice cuando el jugador debería dejar de apostar. Formalmente, un tiempo de parada es un tipo de variable aleatoria que toma valores en el espacio paramétrico de un proceso estocástico y que tiene unas características muy útiles.

Definición 3.18 (Tiempo de parada de una filtración). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una filtración de σ -álgebras, y sea R una variable aleatoria definida sobre (Γ, \mathcal{A}, P) y tomando valores en $\mathbb{N}_0 \cup \{+\infty\}$. Diremos que R es un tiempo de parada para la filtración $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ si se cumple que $R^{-1}(n) = \{R = n\} \in \mathcal{G}_n, \forall n \in \mathbb{N}_0$. Es fácil ver que, por ser $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una filtración, las dos siguientes condiciones son equivalentes:*

$$\{R = n\} \in \mathcal{G}_n, \forall n \in \mathbb{N}_0 \Leftrightarrow \{R \leq n\} \in \mathcal{G}_n, \forall n \in \mathbb{N}_0 \quad (3.3)$$

Definición 3.19. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una martingala (submartingala, supermartingala) sobre (Γ, \mathcal{A}, P) . Sea R un tiempo de parada para la filtración $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$. Entonces definimos la variable aleatoria $X_R : (\Gamma, \mathcal{A}, P) \rightarrow (E, \mathcal{B}_E)$ como:*

$$X_R(\gamma) = X_{R(\gamma)}(\gamma), \forall \gamma \in \Gamma$$

Una vez presentado este concepto, podemos demostrar el siguiente teorema.

Teorema 3.1 ([GT13]). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una submartingala sobre (Γ, \mathcal{A}, P) . Sean R, L dos tiempos de parada para la filtración $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ satisfaciendo que $R \leq L \leq K$, c.s.-P para algún $K \in \mathbb{N}_0$. Entonces se cumple que:*

$$E[X_R] \leq E[X_L]$$

Demostración. En primer lugar es fácil ver que:

$$X_R = \sum_{j=0}^K X_j \mathbb{1}_{\{L \geq j\}} \mathbb{1}_{\{R=j\}}, \text{ c.s.-}P$$

donde $\mathbb{1}_A$ representa la función indicadora del suceso A . Vemos también que, dado $j \in \{0, \dots, K\}$, se cumple la siguiente igualdad:

$$X_j \mathbb{1}_{\{L \geq j\}} = X_j \mathbb{1}_{\{L \geq j\}} - X_j \mathbb{1}_{\{L \geq K+1\}} = \sum_{i=j}^K (X_i \mathbb{1}_{\{L \geq i\}} - X_{i+1} \mathbb{1}_{\{L \geq i+1\}}), \text{ c.s.-}P$$

donde se ha usado que $\mathbb{1}_{\{L \geq K+1\}} = 0$, *c.s.-*P. Por tanto:

$$\begin{aligned} X_R &= \sum_{j=0}^K \sum_{i=j}^K (X_i \mathbb{1}_{\{L \geq i\}} - X_{i+1} \mathbb{1}_{\{L \geq i+1\}}) \mathbb{1}_{\{R=j\}} \\ &= \sum_{j=0}^K \sum_{i=j}^K X_i \mathbb{1}_{\{L=i\}} \mathbb{1}_{\{R=j\}} + \sum_{j=0}^K \sum_{i=j}^K (X_i - X_{i+1}) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}, \text{ c.s.-}P \end{aligned}$$

donde se ha usado que $\mathbb{1}_{\{L \geq i\}} = \mathbb{1}_{\{L=i\}} + \mathbb{1}_{\{L \geq i+1\}}$, $\forall i \in \{0, \dots, K\}$. También vemos que:

$$\sum_{j=0}^K \sum_{i=j}^K X_i \mathbb{1}_{\{L=i\}} \mathbb{1}_{\{R=j\}} = X_L \sum_{j=0}^K \sum_{i=j}^K \mathbb{1}_{\{L=i\}} \mathbb{1}_{\{R=j\}} = X_L, \text{ c.s.-}P$$

Y por tanto:

$$X_R - X_L = \sum_{j=0}^K \sum_{i=j}^K (X_i - X_{i+1}) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}, \text{ c.s.-}P \quad (*)$$

Fijamos $j, i \in \mathbb{N}_0$, con $0 \leq j \leq i \leq K$. Entonces se tiene que el suceso $\{R=j\} \in \mathcal{G}_j \subseteq \mathcal{G}_i$ por ser R un tiempo de parada. De igual forma, en virtud de la equivalencia 3.3, se tiene que el suceso $\{L \geq i+1\} = \{L \leq i\}^c \in \mathcal{G}_i$. Por tanto $\mathbb{1}_{\{L \geq i+1\}}$, $\mathbb{1}_{\{R=j\}}$ son \mathcal{G}_i -medibles. Vemos por tanto que:

$$\begin{aligned} E[(X_i - X_{i+1}) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}] &\stackrel{(1)}{=} E[E[(X_i - X_{i+1}) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}} / \mathcal{G}_i]] \\ &\stackrel{(2)}{=} E[E[(X_i - X_{i+1}) / \mathcal{G}_i] \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}] \\ &\stackrel{(3)}{=} E[(X_i - E[X_{i+1} / \mathcal{G}_i]) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}] \stackrel{(4)}{\leq} 0 \end{aligned}$$

donde en (1) se ha usado la propiedad 2.3.(d), en (2) se ha usado la propiedad 2.3.(g), en (3) se ha usado la linealidad de la esperanza condicionada

(propiedad 2.3.(a)) y la propiedad 2.3.(e) y por último en (4) se ha usado la definición de submartingala (3.2) y que la esperanza de una variable aleatoria no positiva es no positiva.

Por tanto podemos tomar esperanzas a ambos lados de la igualdad $(*_1)$ obteniendo:

$$E[X_R] - E[X_L] = \sum_{j=0}^K \sum_{i=j}^K E[(X_i - X_{i+1}) \mathbb{1}_{\{L \geq i+1\}} \mathbb{1}_{\{R=j\}}] \leq 0$$

Y por tanto $E[X_R] \leq E[X_L]$, como queríamos. ■

3.2.2. Número de upcrossings y martingalas inversas

Pasamos a definir el concepto de número de *upcrossings* de una sucesión de números reales. Intuitivamente, fijado un intervalo $[a, b]$, el número de upcrossings de una sucesión es el número de veces que, partiendo de un valor menor que a , acaba en un valor mayor que b , es decir, el número de veces que 'atraviesa' el intervalo 'de abajo hacia arriba'.

Definición 3.20 (Número de upcrossings). *Sean $a, b \in \mathbb{R}$, $a < b$, y sea $\{X_n\}_{n \in \mathbb{N}_0}$ una sucesión de números reales. Definimos dos sucesiones $\{\tau_j\}_{j \in \mathbb{N}_0}$, $\{\xi_j\}_{j \in \mathbb{N}_0}$, con $\xi_0 = \tau_0 = 0$. Definimos las familias de forma recursiva:*

$$\tau_{j+1} = \inf\{k \geq \xi_j : X_k \leq a\}, \quad \xi_{j+1} = \inf\{k \geq \tau_{j+1} : X_k \geq b\}$$

Bajo la convención de que $\inf(\emptyset) = +\infty$, tenemos que las sucesiones tomarán valores en $\mathbb{N}_0 \cup \{+\infty\}$. Para $m \geq 1$, el número de upcrossings de $[a, b]$ entre los tiempos 0 y m por la sucesión $\{X_n\}_{n \in \mathbb{N}_0}$ se define como:

$$U_m[a, b](\{X_n\}_{n \in \mathbb{N}_0}) := \max\{j \geq 0 : \xi_j \leq m\}$$

Notaremos el número de upcrossings como U_m para simplificar la notación cuando no lleve a confusión.

Podemos extender el concepto de número de upcrossings a procesos estocásticos como veremos a continuación.

Definición 3.21 (Número de upcrossings de un proceso estocástico). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico real en tiempo discreto sobre (Γ, \mathcal{A}, P) . Sean $a, b \in \mathbb{R}$, $a < b$, $m \in \mathbb{N}_0$. Definimos los procesos estocásticos $\{\tau_j\}_{j \in \mathbb{N}_0}$, $\{\xi_j\}_{j \in \mathbb{N}_0}$ de la siguiente manera:*

$$\xi_0 = \tau_0 = 0, \text{ c.s.-}P$$

$$\tau_{j+1}(\gamma) = \inf\{k \geq \xi_j(\gamma) : X_k(\gamma) \leq a\}, \forall \gamma \in \Gamma$$

$$\xi_{j+1}(\gamma) = \inf\{k \geq \tau_{j+1}(\gamma) : X_k(\gamma) \geq b\}, \forall \gamma \in \Gamma$$

Entonces el número de upcrossings de $[a, b]$ entre los tiempos 0 y m por el proceso $\{X_n\}_{n \in \mathbb{N}_0}$, que notaremos por $U_m[a, b](\{X_n\}_{n \in \mathbb{N}_0})$ (o por U_m para simplificar la notación), se define como:

$$U_m(\gamma) := \max\{j \geq 0 : \xi_j(\gamma) \leq m\}, \forall \gamma \in \Gamma$$

Con el objetivo de demostrar una desigualdad relacionada con el número de upcrossings de una submartingala, enunciaremos y demostraremos dos proposiciones previas que necesitaremos.

Proposición 3.4. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico real (en tiempo discreto) definido sobre (Γ, \mathcal{A}, P) . Sea $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una filtración de σ -álgebras tal que $\{X_n\}_{n \in \mathbb{N}_0}$ está adaptado a $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$. Sean $a, b \in \mathbb{R}, a < b$, $m \in \mathbb{N}_0$. Entonces, en el contexto de la definición 3.21, τ_j, ξ_j son tiempos de parada respecto a $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}, \forall j \in \mathbb{N}_0$.*

Demuestra. Empezaremos viendo que τ_j es un tiempo de parada $\forall j \in \mathbb{N}_0$. La demostración para ξ_j se puede abarcar de manera similar y no la incluiremos aquí. Claramente $\{\tau_0 \leq n\} = \Gamma \in \mathcal{G}_n, \forall n \in \mathbb{N}_0$, y por tanto τ_0 es un tiempo de parada.

De aquí en adelante notar que $\bigcap_{i=n}^{n-1} A_i = \Gamma$ para cualesquiera conjuntos A_i y $\forall n \in \mathbb{N}_0$, es decir, la intersección vacía es igual al total. Fijado $N \in \mathbb{N}_0$, consideramos ahora el conjunto $\{\tau_1 = N\}$. Es claro que $\{\tau_1 = N\} = (\bigcap_{i=0}^{N-1} \{X_i > a\}) \cap \{X_N \leq a\}$. Dado que X_i es \mathcal{G}_N -medible $\forall i \leq N$, tenemos que $\{X_i > a\} = \{X_i \leq a\}^c \in \mathcal{G}_N, \forall i \leq N$, donde usamos $(\cdot)^c$ para notar el complementario de un conjunto. De igual forma, $\{X_N \leq a\} \in \mathcal{G}_N$, por lo que deducimos que $\{\tau_1 = N\} \in \mathcal{G}_N, \forall N \in \mathbb{N}_0$, y τ_1 es un tiempo de parada.

Por último, fijamos $j \in \mathbb{N}, j \geq 2$. Fijamos también $N \in \mathbb{N}_0$. Distinguimos dos casos:

- Si $N < 2(j - 1)$: Es fácil ver que $\{\tau_j = N\} = \emptyset \in \mathcal{G}_N$.
- Si $N \geq 2(j - 1)$: Definimos los conjuntos $I(k) := \{0, \dots, k\}, \forall k \in \mathbb{N}_0$. Consideramos el siguiente conjunto de funciones:

$$F = \{f : I(2j-2) \rightarrow I(N) / f(2j-2) = N; f(k) < f(k+1), \forall k \in I(2j-2)\}$$

Intuitivamente el anterior conjunto de funciones recoge las posibles formas de elegir de forma ordenada $2j - 2$ índices de entre los N

primeros índices de la sucesión (ya que el último está fijo). Dado $f \in F$, definimos los siguientes conjuntos:

$$A_f = \bigcap_{k=0}^{j-2} (\{X_{f(2k)} \leq a\} \cap (\bigcap_{l=f(2k)+1}^{f(2k+1)-1} \{X_l < b\}))$$

$$B_f = \bigcap_{k=0}^{j-2} (\{X_{f(2k+1)} \geq b\} \cap (\bigcap_{l=f(2k+1)+1}^{f(2k+2)-1} \{X_l > a\}))$$

Y vemos que:

$$\{\tau_j = N\} = \bigcup_{f \in F} ((\bigcap_{l=0}^{f(0)-1} \{X_l > a\}) \cap A_f \cap B_f \cap \{X_N \leq a\})$$

Como hicimos en el caso $j = 1$, es fácil ver que estamos uniendo e interseccionando conjuntos que pertenecen a \mathcal{G}_N , y por tanto $\{\tau_j = N\} \in \mathcal{G}_N$.

Como fijamos un N arbitrario, deducimos que τ_j es un tiempo de parada. De igual forma el j fijado es arbitrario, y concluimos que τ_j es un tiempo de parada $\forall j \in \mathbb{N}_0$. ■

Proposición 3.5. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una submartingala sobre (Γ, \mathcal{A}, P) . Sea Φ una función creciente, convexa y continua, tal que $E[\Phi(X_n)] < +\infty$, $\forall n \in \mathbb{N}_0$. Entonces $\{\Phi(X_n), \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es una submartingala.*

Demostración. Esta proposición es una consecuencia directa de la propiedad 2.3.(j). En efecto:

$$E[\Phi(X_{n+1})]/\mathcal{G}_n \geq \Phi(E[X_{n+1}]/\mathcal{G}_n) \stackrel{(1)}{\geq} \Phi(X_n), \text{ c.s.-}P_{\mathcal{G}_n}, \forall n \in \mathbb{N}_0$$

donde en (1) se ha usado que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ es una submartingala y que Φ es creciente. ■

Ya tenemos las herramientas necesarias para demostrar la siguiente desigualdad. En esta demostración es clave el uso del teorema 3.1.

Teorema 3.2 ([GT13]). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}_0}$ una submartingala sobre (Γ, \mathcal{A}, P) . Sean $a, b \in \mathbb{R}$, $a < b$, $m \in \mathbb{N}_0$. Entonces:*

$$E[U_m] \leq \frac{1}{b-a} E((X_m - a)^+)$$

donde usamos $(\cdot)^+$ para notar la función parte positiva.

Notación 3.3. $s \wedge z = \min\{s, z\}$

Demuestra. Tomamos $Y_n := (X_n - a)^+$, $\forall n \in \mathbb{N}_0$. La función $\Phi(s) = (s - a)^+$ claramente es continua, convexa y no decreciente. Por tanto, aplicando la proposición 3.5, tenemos que $\{Y_n\}_{n \in \mathbb{N}_0}$ es una submartingala. Fijamos ahora $m \in \mathbb{N}_0$. Es claro que por la definición de $\{\tau_j\}_{j \in \mathbb{N}_0}$, se cumple que $\tau_{m+1} > m$, por lo que $Y_{\tau_{m+1} \wedge m} = Y_m$. Por tanto tenemos que:

$$\begin{aligned} Y_m &= Y_{\tau_{m+1} \wedge m} = Y_{\tau_1 \wedge m} + \sum_{i=1}^m (Y_{\tau_{i+1} \wedge m} - Y_{\tau_i \wedge m}) \\ &= Y_{\tau_1 \wedge m} + \sum_{i=1}^m (Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m} + Y_{\xi_i \wedge m} - Y_{\tau_i \wedge m}) \\ &= Y_{\tau_1 \wedge m} + \sum_{i=1}^m (Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m}) + \sum_{i=1}^m (Y_{\xi_i \wedge m} - Y_{\tau_i \wedge m}) \end{aligned}$$

Es claro que $Y_n \geq 0$, $\forall n \in \mathbb{N}_0$ por como la hemos definido, y por tanto:

$$Y_m \geq \sum_{i=1}^m (Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m}) + \sum_{i=1}^m (Y_{\xi_i \wedge m} - Y_{\tau_i \wedge m})$$

Por otro lado es fácil ver que:

$$\sum_{i=1}^m (Y_{\xi_i \wedge m} - Y_{\tau_i \wedge m}) \geq \sum_{i=1}^{U_m} Y_{\xi_i \wedge m} \geq U_m(b - a)$$

Y por tanto:

$$Y_m \geq \sum_{i=1}^m (Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m}) + U_m(b - a)$$

Como hemos mencionado anteriormente, $\{Y_n\}_{n \in \mathbb{N}_0}$ es una submartingala. Por otro lado, en vista de la proposición 3.4, es sencillo comprobar que $\tau_{i+1} \wedge m$ y $\xi_i \wedge m$ son tiempos de parada para $\{\mathcal{G}_n\}_{n \in \mathbb{N}_0}$, $\forall i \in \{1, \dots, m\}$.

También es claro que $\xi_i \wedge m \leq \tau_{i+1} \wedge m \leq m$, *c.s.-P*, y por tanto podemos aplicar el teorema 3.1 para ver que $E[Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m}] \geq 0$. Aplicando esperanzas a ambos lados de la desigualdad anterior:

$$E[Y_m] \geq \sum_{i=1}^m E[Y_{\tau_{i+1} \wedge m} - Y_{\xi_i \wedge m}] + E[U_m](b-a) \geq (b-a)E[U_m]$$

Como queríamos. ■

Presentamos ahora el concepto de *martingala inversa* (*backward martingale*).

Definición 3.22 (Martingala inversa). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ una sucesión de σ -álgebras cumpliendo que:*

$$\forall n \in \mathbb{N}, \mathcal{G}_{n+1} \subseteq \mathcal{G}_n \subseteq \mathcal{A} \quad (3.4)$$

Sea $\{X_n\}_{n \in \mathbb{N}}$ un proceso estocástico real (en tiempo discreto) definido sobre (Γ, \mathcal{A}, P) , con X_n \mathcal{G}_n -medible e integrable para todo $n \in \mathbb{N}$. Entonces diremos que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}}$ es una martingala inversa si cumple que:

$$E[X_n | \mathcal{G}_{n+1}] = X_{n+1}, \text{ c.s.-}P_{\mathcal{G}_{n+1}}, \forall n \in \mathbb{N}$$

Introducimos la siguiente notación que nos permite adaptar el concepto de upcrossings a martingalas inversas.

Notación 3.4. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}}$ un proceso real en tiempo discreto definido sobre (Γ, \mathcal{A}, P) . Sean $a, b \in \mathbb{R}, a < b$, $m \in \mathbb{N}$. Consideramos el proceso estocástico $\{X'_n\}_{n \in \mathbb{N}_0}$ definido como sigue:*

$$\begin{aligned} X'_n &= X_{m-n}, \forall n \in \{0, \dots, m-1\} \\ X'_n &= X_{n-m+1}, \forall n \geq m \end{aligned} \quad (3.5)$$

Notaremos por $U_{-m}[a, b](\{X_n\}_{n \in \mathbb{N}})$ al número de upcrossings del intervalo $[a, b]$ entre los tiempos 0 y $m-1$ por el proceso $\{X'_n\}_{n \in \mathbb{N}_0}$, es decir, $U_{-m}[a, b](\{X_n\}_{n \in \mathbb{N}}) := U_{m-1}[a, b](\{X'_n\}_{n \in \mathbb{N}_0})$. Por comodidad lo notaremos por U_{-m} cuando no de lugar a ambigüedad.

Y vemos que el teorema 3.2 se puede adaptar de la siguiente manera.

Teorema 3.3. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}}$ una martingala inversa definida sobre (Γ, \mathcal{A}, P) . Sean $a, b \in \mathbb{R}, a < b$, $m \in \mathbb{N}$. Entonces:*

$$E[U_{-m}] \leq \frac{1}{b-a} E((X_1 - a)^+)$$

Demostración. Fijado $m \in \mathbb{N}_0$, consideramos el proceso estocástico $\{X'_n\}_{n \in \mathbb{N}_0}$ definido como en (3.5). Consideramos también la sucesión de σ -álgebras $\{\mathcal{G}'_n\}_{n \in \mathbb{N}_0}$ definida como sigue:

$$\mathcal{G}'_n = \mathcal{G}_{m-n}, \forall n \in \{0, \dots, m-1\}$$

$$\mathcal{G}'_n = \mathcal{G}_{n-m+1}, \forall n \geq m$$

Vemos por tanto que:

$$E[X'_{n+1} / \mathcal{G}'_n] = E[X_{m-n-1} / \mathcal{G}_{m-n}] \stackrel{(1)}{=} X_{m-n} = X'_n, \text{ c.s.-}P_{\mathcal{G}'_n}, \forall n \in \{0, \dots, m-2\}$$

donde en (1) se ha usado que $\{X_n, \mathcal{G}_n\}_{n \in \mathbb{N}}$ es una martingala inversa.
Además:

$$\mathcal{G}'_n = \mathcal{G}_{m-n} \subseteq \mathcal{G}_{m-n-1} = \mathcal{G}'_{n+1}, \forall n \in \{0, \dots, m-2\}$$

Deducimos que entre los términos 0 y $m-1$, $\{X'_n, \mathcal{G}'_n\}_{n \in \mathbb{N}_0}$ se comporta como una martingala y, en particular, como una submartingala. Es fácil concluir que, como $U_{m-1}[a, b](\{X'_n\}_{n \in \mathbb{N}_0})$ sólo tiene en cuenta dichos términos, aplicando el teorema 3.2 se cumple que:

$$E[U_{-m}] = E[U_{m-1}[a, b](\{X'_n\}_{n \in \mathbb{N}_0})] \leq \frac{E[(X'_{m-1} - a)^+]}{b - a} = \frac{E[(X_1 - a)^+]}{b - a}$$

■

Por último antes de enunciar el teorema principal, enunciamos un teorema que será clave en su demostración.

Teorema 3.4 (Ley 0-1 de Kolmogorov, ver [Wil91]). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes, definidas sobre (Γ, \mathcal{A}, P) . Consideramos la σ -álgebra $\mathcal{T} = \bigcap_{k=1}^{+\infty} \sigma(X_k, X_{k+1}, \dots)$. Entonces cualquier variable aleatoria \mathcal{T} -medible es constante casi seguramente.*

3.2.3. Ley fuerte de los números grandes

Ya podemos enunciar y demostrar el resultado principal de esta sección, que nos permite asegurar la convergencia del estimador de Monte Carlo.

Teorema 3.5 (Ley fuerte de los números grandes, [GT13]). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, definidas sobre (Γ, \mathcal{A}, P) . Supongamos que*

$$E[|X_1|] < +\infty$$

Para $N \in \mathbb{N}$, consideramos el siguiente estimador de (X_1, \dots, X_N) :

$$S_N := \frac{1}{N} \sum_{i=1}^N X_i$$

Entonces se cumple que:

$$\lim_{N \rightarrow +\infty} S_N = E[X_1], \text{ c.s.-}P$$

Demostración. Dividiremos la demostración en 4 pasos.

Paso 1

Tomamos $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ una sucesión de σ -álgebras definida como:

$$\mathcal{G}_n := \sigma(S_n, S_{n+1}, \dots), \forall n \in \mathbb{N}$$

Claramente $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$ satisface (3.4). Tomamos ahora una sucesión de variables aleatorias $\{Y_n\}_{n \in \mathbb{N}}$ definida como sigue:

$$Y_n := E[X_1 / \mathcal{G}_n], \forall n \in \mathbb{N} \quad (*_2)$$

Claramente, para todo $n \in \mathbb{N}$, Y_n es \mathcal{G}_n -medible e integrable (por definición de esperanza condicionada). Vemos además que:

$$E[Y_n / \mathcal{G}_{n+1}] = E[E[X_1 / \mathcal{G}_n] / \mathcal{G}_{n+1}] \stackrel{(1)}{=} E[X_1 / \mathcal{G}_{n+1}] = Y_{n+1}, \text{ c.s.-}P_{\mathcal{G}_{n+1}}, \forall n \in \mathbb{N}$$

donde en (1) se ha utilizado la propiedad 2.3.(i). Deducimos que $\{Y_n, \mathcal{G}_n\}_{n \in \mathbb{N}}$ es una martingala inversa.

Por último veamos que $Y_n = S_n, \forall n \in \mathbb{N}$. Es claro que S_n es \mathcal{G}_n -medible para todo $n \in \mathbb{N}$. Por tanto:

$$S_n \stackrel{(1)}{=} E[S_n / \mathcal{G}_n] = \frac{1}{n} \sum_{i=1}^n E[X_i / \mathcal{G}_n] \stackrel{(2)}{=} \frac{1}{n} \sum_{i=1}^n E[X_1 / \mathcal{G}_n] = Y_n, \text{ c.s.-}P_{\mathcal{G}_n}, \forall n \in \mathbb{N}$$

donde en (1) se usa la propiedad 2.3.(e), y en (2) se usa que todas las variables X_i están idénticamente distribuidas. Notar que, en particular, $Y_1 = X_1$.

Paso 2

En este paso demostraremos que $\{Y_n\}_{n \in \mathbb{N}}$ converge puntualmente casi seguramente a una variable aleatoria integrable Y . Fijamos $a, b \in \mathbb{R}$. Teniendo en cuenta la notación introducida en 3.4, vamos a considerar la sucesión

$\{U_{-n}[a, b](\{Y_n\}_{n \in \mathbb{N}})\}_{n \in \mathbb{N}} = \{U_{-n}\}_{n \in \mathbb{N}}$. $\{U_{-n}\}_{n \in \mathbb{N}}$ es no decreciente por como está definida y, por tanto, es claro que $\forall \gamma \in \Gamma$, $\{U_{-n}(\gamma)\}_{n \in \mathbb{N}}$ tiene límite en $\mathbb{R} \cup \{+\infty\}$. Definimos la función $U : (\Gamma, \mathcal{A}, P) \rightarrow \mathbb{R} \cup \{+\infty\}$ como el límite puntual de $\{U_{-n}\}_{n \in \mathbb{N}}$, es decir, $U(\gamma) = \lim_{n \rightarrow +\infty} U_{-n}(\gamma)$, $\forall \gamma \in \Gamma$. Vemos que, usando el teorema de la convergencia monótona, el teorema 3.3 y la hipótesis de que $E[|X_1|] < +\infty$:

$$\begin{aligned} E[U] &= \lim_{n \rightarrow +\infty} E[U_{-n}] \leq \frac{1}{b-a} E[(Y_1 - a)^+] = \frac{1}{b-a} E[(X_1 - a)^+] \\ &\leq \frac{1}{b-a} E[|X_1 - a|] \leq \frac{1}{b-a} (E[|X_1|] + E[|a|]) < +\infty \end{aligned}$$

Por tanto la variable aleatoria U es finita *c.s.-P*, y la sucesión $\{Y_n\}_{n \in \mathbb{N}}$ cruza el intervalo $[a, b]$ un número finito de veces casi seguramente. De esto deducimos que existirá un conjunto $O \subset \Gamma$ de probabilidad nula tal que, fijado $\gamma \in \Gamma \setminus O$, se cumplirá una de las dos siguientes posibilidades:

- $\exists N \in \mathbb{N}$ tal que $\forall n \geq N$, $Y_n(\gamma) > a$, lo cual implica que:

$$\liminf_{n \rightarrow +\infty} Y_n(\gamma) \geq a$$

- $\exists N \in \mathbb{N}$ tal que $\forall n \geq N$, $Y_n(\gamma) < b$, lo cual implica que:

$$\limsup_{n \rightarrow +\infty} Y_n(\gamma) \leq b$$

Por tanto:

$$P(\{\liminf_{n \rightarrow +\infty} Y_n < a, \limsup_{n \rightarrow +\infty} Y_n > b\}) = 0$$

Dado que esto se cumple $\forall a, b \in \mathbb{R}$, con $a < b$, deducimos que:

$$P(\{\liminf_{n \rightarrow +\infty} Y_n < \limsup_{n \rightarrow +\infty} Y_n\}) = 0$$

Y podemos afirmar que $\liminf_{n \rightarrow +\infty} Y_n = \limsup_{n \rightarrow +\infty} Y_n$, *c.s.-P*, con lo que $\{Y_n\}_{n \in \mathbb{N}}$ tiene límite en $\mathbb{R} \cup \{-\infty, +\infty\}$, *c.s.-P*. Definimos la variable aleatoria Y como el límite inferior punto a punto de $\{Y_n\}_{n \in \mathbb{N}}$, es decir, $Y(\gamma) := \liminf_{n \rightarrow +\infty} Y_n(\gamma)$, $\forall \gamma \in \Gamma$, y tenemos que $Y = \lim_{n \rightarrow +\infty} Y_n$, *c.s.-P*. Como podemos observar:

$$E[|Y_n|] \stackrel{(1)}{=} E[|E[X_1/\mathcal{G}_n]|] \stackrel{(2)}{\leq} E[E[|X_1|/\mathcal{G}_n]] \stackrel{(3)}{=} E[|X_1|], \forall n \in \mathbb{N}$$

En la igualdad (1) hemos usado la definición de $\{Y_n\}_{n \in \mathbb{N}}$ en $(*_2)$, en (2) hemos usado la propiedad 2.3.(k) y en (3) hemos usado la propiedad 2.3.(d).

Dado que el valor absoluto es una función continua, se cumple que $|Y| = \lim_{n \rightarrow +\infty} |Y_n|$, *c.s.-P.* Por último, aplicando el lema de Fatou y la desigualdad anterior obtenemos:

$$E[|Y|] \leq \liminf_{n \rightarrow +\infty} E[|Y_n|] \leq E[|X_1|] < +\infty$$

Con lo que Y es absolutamente integrable, y por tanto, integrable.

Paso 3

Dado que no podemos asegurar que $\{Y_n\}_{n \in \mathbb{N}}$ sea monótona ni acotada, no podemos aplicar el teorema de la convergencia monótona ni el de la convergencia dominada para afirmar que converge en el espacio $L_1(\Gamma, \mathcal{A}, P)$ de las funciones integrables definidas sobre (Γ, \mathcal{A}, P) . Por tanto vamos a demostrarlo utilizando otros métodos. Fijamos $C \in \mathbb{R}^+$. Empezamos viendo que:

$$\begin{aligned} E[|Y - Y_n|] &= E[|Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}] + E[|Y - Y_n| \mathbf{1}_{\{|Y_n| > C\}}] \\ &\leq E[|Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}] + E[(|Y| + |Y_n|) \mathbf{1}_{\{|Y_n| > C\}}], \forall n \in \mathbb{N} \end{aligned}$$

Por otro lado, usando la definición de $\{Y_n\}_{n \in \mathbb{N}}$ ([\(*2\)](#)), la propiedad [2.3.\(g\)](#), y la propiedad [2.3.\(k\)](#):

$$\begin{aligned} |Y_n| \mathbf{1}_{\{|Y_n| > C\}} &= |E[X_1 / \mathcal{G}_n] \mathbf{1}_{\{|Y_n| > C\}}| \\ &= |E[X_1 \mathbf{1}_{\{|Y_n| > C\}} / \mathcal{G}_n]| \\ &\leq E[|X_1| \mathbf{1}_{\{|Y_n| > C\}} / \mathcal{G}_n], \text{ *c.s.-P.* } \forall n \in \mathbb{N} \end{aligned}$$

Por tanto:

$$E[|Y_n| \mathbf{1}_{\{|Y_n| > C\}}] \leq E[E[|X_1| \mathbf{1}_{\{|Y_n| > C\}} / \mathcal{G}_n]] = E[|X_1| \mathbf{1}_{\{|Y_n| > C\}}], \forall n \in \mathbb{N}$$

donde se ha usado la propiedad [2.3.\(d\)](#). Utilizando esta desigualdad, tomamos $Z := |Y| + |X_1|$, y vemos que:

$$E[|Y - Y_n|] \leq E[|Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}] + E[Z \mathbf{1}_{\{|Y_n| > C\}}], \forall n \in \mathbb{N} \quad (*_3)$$

Claramente, por ser Y integrable, $|Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}$ está acotada *c.s.-P.*, por lo que podemos usar el teorema de la convergencia dominada para asegurar que:

$$\lim_{n \rightarrow +\infty} E[|Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}] = E[\lim_{n \rightarrow +\infty} |Y - Y_n| \mathbf{1}_{\{|Y_n| \leq C\}}] = E[0] = 0 \quad (*_4)$$

Fijamos ahora $B \in \mathbb{R}^+$, y vemos que:

$$Z \mathbf{1}_{\{|Y_n| > C\}} \leq Z \mathbf{1}_{\{Z > B\}} + B \mathbf{1}_{\{|Y_n| > C\}}, \text{ *c.s.-P.* } \forall n \in \mathbb{N}$$

Y por tanto:

$$E[Z\mathbb{1}_{\{|Y_n|>C\}}] \leq E[Z\mathbb{1}_{\{Z>B\}}] + BP(|Y_n|>C), \forall n \in \mathbb{N} \quad (*_5)$$

Dado que B y C son reales positivos arbitrarios, y dado que Y y X_1 son integrables (y por tanto absolutamente integrables), para todo $\epsilon > 0$ podemos elegir B_ϵ lo suficientemente grande tal que:

$$E[Z\mathbb{1}_{\{Z>B_\epsilon\}}] < \epsilon \quad (*_6)$$

Y de igual forma podemos elegir C_ϵ lo suficientemente grande tal que:

$$B_\epsilon P(|Y|>C_\epsilon) < \epsilon \quad (*_7)$$

Utilizando que $\{\mathbb{1}_{|Y_n|>C}\}_{n \in \mathbb{N}}$ converge puntualmente a $\mathbb{1}_{|Y|>C}$, $\forall C \in \mathbb{R}^+$ casi seguramente, y el teorema de la convergencia dominada, vemos que:

$$\lim_{n \rightarrow +\infty} B_\epsilon P(|Y_n|>C_\epsilon) = B_\epsilon P(|Y|>C_\epsilon) < \epsilon$$

Por último, fijamos $\epsilon > 0$, y tomando límites en la desigualdad $(*_3)$, y aplicando los resultados obtenidos en $(*_4)$, $(*_5)$, $(*_6)$ y $(*_7)$, vemos que:

$$\begin{aligned} \lim_{n \rightarrow +\infty} E[|Y - Y_n|] &\leq \lim_{n \rightarrow +\infty} E[|Y - Y_n|\mathbb{1}_{\{|Y_n| \leq C_\epsilon\}}] + \lim_{n \rightarrow +\infty} E[Z\mathbb{1}_{\{|Y_n|>C_\epsilon\}}] \\ &\leq E[Z\mathbb{1}_{\{Z>B_\epsilon\}}] + \lim_{n \rightarrow +\infty} B_\epsilon P(|Y_n|>C_\epsilon) \\ &< \epsilon + B_\epsilon P(|Y|>C_\epsilon) < 2\epsilon \end{aligned}$$

Como esto se cumple para todo $\epsilon > 0$, concluimos que:

$$\lim_{n \rightarrow +\infty} E[|Y - Y_n|] = 0$$

Con lo que acabamos de demostrar que $\{Y_n\}_{n \in \mathbb{N}}$ converge en L_1 a Y . Esto implica que $E[Y] = \lim_{n \rightarrow +\infty} E[Y_n]$. Aplicando la definición de $\{Y_n\}_{n \in \mathbb{N}}$ en $(*_2)$, y la propiedad 2.3.(d):

$$E[Y] = \lim_{n \rightarrow +\infty} E[Y_n] = \lim_{n \rightarrow +\infty} E[E[X_1]/\mathcal{G}_n] = E[X_1]$$

Paso 4

Por último queremos demostrar que Y es constante casi seguramente. Recordemos que en el paso 1 comprobamos que $Y_n = S_n$, $\forall n \in \mathbb{N}$. Consideramos la σ -álgebra $\mathcal{T} = \bigcap_{k=1}^{+\infty} \sigma(X_k, X_{k+1}, \dots)$. Fijado $k \in \mathbb{N}$, vemos que:

$$Y = \lim_{N \rightarrow +\infty} S_N = \lim_{N \rightarrow +\infty} \frac{X_1 + \dots + X_N}{N} = \lim_{N \rightarrow +\infty} \frac{X_k + \dots + X_{N+k}}{N}$$

Por tanto Y es $\sigma(X_k, X_{k+1}, \dots)$ -medible por ser límite de funciones $\sigma(X_k, X_{k+1}, \dots)$ -medibles. Dado que k es arbitrario, se obtiene que Y es $\sigma(X_k, X_{k+1}, \dots)$ -medible, $\forall k \in \mathbb{N}$, con lo que Y es \mathcal{T} -medible. Deducimos usando el teorema 3.4 que Y es constante, y dado que al final del paso 3 comprobamos que $E[Y] = E[X_1]$, podemos concluir el resultado buscado:

$$\lim_{N \rightarrow +\infty} S_N = Y = E[X_1], \text{ c.s.-}P$$

■

Este teorema nos asegura la convergencia del estimador de Monte Carlo definido en (1.6) cuando el número de muestras tiende a infinito. En efecto, sea (Γ, \mathcal{A}, P) un espacio probabilístico, sea $X : (\Gamma, \mathcal{A}, P) \rightarrow (E, \mathcal{B}_E)$ un vector aleatorio y sea $g : (E, \mathcal{B}_E) \rightarrow (\mathbb{R}, \mathcal{B})$ una función medible. Considerando $\{x^{(r)}\}_{r \in \mathbb{N}}$ una sucesión de vectores aleatorios idénticamente distribuidos según la distribución de X e independientes, entonces $\{g(x^{(r)})\}_{r \in \mathbb{N}}$ es una sucesión de variables aleatorias idénticamente distribuidas según la distribución de $g(X)$ e independientes. La ley fuerte de los números grandes nos asegura por tanto que los estimadores:

$$S_N = \frac{1}{N} \sum_{i=0}^N g(x^{(i)})$$

convergen a $E[g(x^{(1)})] = E[g(X)]$ cuando N tiende a infinito casi seguramente.

3.3. Métodos de Monte Carlo para muestreo de variables aleatorias

Hemos visto un estimador que nos permite aproximar la esperanza de una función bajo una distribución a partir de muestras independientes siguiendo dicha distribución, que es uno de los objetivos de los métodos de Monte Carlo. En esta sección estudiaremos los métodos más utilizados en renderización que nos permiten abarcar el otro objetivo, tomar muestras que sigan una distribución de probabilidad. Notar que hay ciertas distribuciones que son fácilmente simuladas computacionalmente, como por ejemplo la distribución uniforme unidimensional, cuyas muestras se obtienen a partir de un generador de números pseudoaleatorios.

3.3.1. Método de inversión

Este método se utiliza para muestrear variables aleatorias y se basa en hallar una inversa continua a la izquierda de la función de distribución de la variable que queremos muestrear. Empezamos definiendo dicha función inversa.

Definición 3.23. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico, sea X una variable aleatoria y sea F_X su función de distribución. Definimos la inversa (continua a la izquierda) de F_X , notada por F_X^- , como:*

$$\begin{aligned} F_X^- : [0, 1] &\rightarrow \mathbb{R} \\ u \rightarrow F_X^-(u) &= \inf\{y \in \mathbb{R} / F(y) \geq u\} \end{aligned}$$

El método de inversión se basa en el siguiente teorema.

Teorema 3.6 (Método de inversión). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico, sea X una variable aleatoria definida sobre (Γ, \mathcal{A}, P) , sea F_X su función de distribución. Sea $V \sim U(0, 1)$ una variable aleatoria definida sobre (Γ, \mathcal{A}, P) , donde notamos por $U(a, b)$ a la distribución uniforme en el intervalo $[a, b]$. Entonces las variables aleatorias $Y := F_X^-(V)$, X están idénticamente distribuidas, o lo que es lo mismo, $F_X(t) = F_Y(t)$, $\forall t \in \mathbb{R}$.*

Demuestra. Como vemos $F_V(s) = P(V \leq s) = s$, $\forall s \in [0, 1]$. Deducimos que, dado $t \in \mathbb{R}$:

$$F_Y(t) = P(Y \leq t) = P(F_X^-(V) \leq t) \stackrel{(1)}{=} P(V \leq F_X(t)) = F_X(t)$$

En la igualdad (1) se ha usado que $\{F_X^-(V) \leq t\} = \{V \leq F_X(t)\}$. En efecto, dado $\gamma \in \{F_X^-(V) \leq t\}$, tenemos que, por ser F_X no decreciente y continua por la derecha, $F_X(F_X^-(V(\gamma))) \geq V(\gamma)$, y por tanto $F_X(t) \geq V(\gamma)$, con lo que $\gamma \in \{V \leq F_X(t)\}$. Por otro lado, dado $\gamma \in \{V \leq F_X(t)\}$, es claro que por definición de F_X^- , $F_X^-(V(\gamma)) \leq t$, con lo que $\gamma \in \{F_X^-(V) \leq t\}$ y tenemos lo que queríamos.

Por tanto concluimos que X e Y están idénticamente distribuidas. ■

Surgen dos problemas a la hora de utilizar este método. Por un lado tenemos la limitación de que sólo es aplicable para variables aleatorias, y por tanto no se puede aplicar en múltiples dimensiones, y por otro lado no siempre podremos hallar de manera sencilla la inversa de la función de distribución de manera analítica. Además, dada una función $f : \mathbb{R} \rightarrow \mathbb{R}_0^+$ continua, habrá ocasiones en que queramos muestrear una variable aleatoria

cuya función de densidad sea $\frac{f}{\int_{\mathbb{R}} f(x)dx}$, y no siempre podremos calcular la integral de la función analíticamente.

Ejemplo 3.1 (Variables aleatorias discretas). *Un ejemplo claro de distribuciones de probabilidad respecto a las que podemos generar muestras usando el método de inversión son aquellas asociadas a variables aleatorias discretas. Supongamos que la función masa de probabilidad de una variable aleatoria X viene dada por:*

$$P(X = x_i) = p_i, \quad x_i \in \mathbb{R}, \quad p_i \in [0, 1], \quad \forall i \in I = \{1, 2, \dots\} \subseteq \mathbb{N}$$

con $\sum_{i \in I} p_i = 1$. Entonces dada una muestra u generada según la distribución uniforme en $[0, 1]$, se puede generar una muestra x siguiendo la distribución de X tomando $x = x_j$, con j satisfaciendo:

$$\sum_{i=1}^{j-1} p_i < u \leq \sum_{i=1}^j p_i$$

3.3.2. Método de rechazo

Este método nos permite muestrear una distribución a partir de otra distribución conocida. Sea (Γ, \mathcal{A}, P) un espacio probabilístico, y sean X, Y dos vectores aleatorios continuos definidos sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) , $E \subseteq \mathbb{R}^n$. Consideremos f_X, f_Y las funciones de densidad de X e Y , y supongamos que existe $c \geq 1$ tal que $f_X(p) \leq cf_Y(p)$, $\forall p \in \mathbb{R}^n$. Supongamos también que sabemos tomar muestras independientes que sigan la distribución de la variable Y . Entonces el método de rechazo consiste en tomar una muestra y siguiendo la distribución de Y , y una muestra v siguiendo la distribución uniforme en $[0, 1]$. Entonces si $v \leq \frac{f_X(y)}{cf_Y(y)}$, y es una muestra siguiendo la distribución de X . El nombre del método corresponde al hecho de que si las muestras y, v no satisfacen la desigualdad anterior, son rechazadas y se toman otras muestras.

El siguiente teorema nos asegura que tomando sucesivas muestras, llegará un punto en que una muestra sea aceptada, es decir, no rechazaremos infinitas muestras. También nos asegura que las muestras aceptadas siguen la distribución deseada.

Teorema 3.7 ([GT13]). *Sea $\{Y_n\}_{n \in \mathbb{N}}$ una sucesión de vectores aleatorios independientes e idénticamente distribuidos con función de densidad g , definida sobre en \mathbb{R}^n . Sea $\{V_n\}_{n \in \mathbb{N}}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas siguiendo una distribución uniforme en $[0, 1]$. Supongamos que todos los términos de $\{Y_n\}_{n \in \mathbb{N}}$ son a su vez independientes de todos los términos de $\{V_n\}_{n \in \mathbb{N}}$. Sea X un vector aleatorio con*

función de densidad f , definida sobre \mathbb{R}^n . Supongamos que se cumple que $f(p) \leq cg(p)$, $\forall p \in \mathbb{R}^n$ para cierto $c \geq 1$. Consideramos la variable aleatoria M y el vector aleatorio X' , definidos como sigue:

$$M := \inf\{k \in \mathbb{N} / U_k \leq \frac{f(Y_k)}{cg(Y_k)}\}$$

$$X' := Y_M$$

Entonces M es finito casi seguramente (en particular $E[M] = c$), y X' sigue la misma distribución que X . Además X' y M son independientes y $P(M = k) = \frac{1}{c}(1 - \frac{1}{c})^{k-1}$, $\forall k \in \mathbb{N}$.

*Demuestra*cción. Por ser U_k independiente de Y_k , $\forall k \in \mathbb{N}$, tenemos que la función de densidad conjunta de U_k e Y_k cumple que $f_{(U_k, Y_k)} = f_{U_k} f_{Y_k} = g$, ya que la función de densidad de U_k es constantemente igual a 1. Por tanto, dado $A \in \mathcal{B}^n$:

$$\begin{aligned} P(U_k \leq \frac{f(Y_k)}{cg(Y_k)}, Y_k \in A) &= \int_A \int_0^{\frac{f(p)}{cg(p)}} f_{(U_k, Y_k)}(u, p) \, du \, dp \\ &= \int_A \frac{f(p)}{cg(p)} g(p) \, dp = \frac{1}{c} \int_A f(p) \, dp, \quad \forall k \in \mathbb{N} \end{aligned}$$

Tomando $A = \mathbb{R}^n$ y usando que f es una función de densidad (y por tanto su integral en \mathbb{R}^n vale 1), deducimos que $P(U_k \leq \frac{f(Y_k)}{cg(Y_k)}) = \frac{1}{c}$, $\forall k \in \mathbb{N}$. Por tanto, usando que todas las variables y vectores son independientes, deducimos que $\forall k \in \mathbb{N}$:

$$\begin{aligned} P(M = k) &= \prod_{i=1}^{k-1} P(U_i > \frac{f(Y_i)}{cg(Y_i)}) P(U_k \leq \frac{f(Y_k)}{cg(Y_k)}) \\ &= \prod_{i=1}^{k-1} (1 - P(U_i \leq \frac{f(Y_i)}{cg(Y_i)})) \frac{1}{c} \\ &= (1 - \frac{1}{c})^{k-1} \frac{1}{c} \end{aligned}$$

Con lo que M sigue una distribución geométrica de parámetro $\frac{1}{c}$, por lo que su esperanza es igual a $E[M] = c < +\infty$, y deducimos que M es finito

casi seguramente. También vemos que $\forall k \in \mathbb{N}, \forall A \in \mathcal{B}^n$:

$$\begin{aligned}
 P(M = k, Y_M \in A) &= P(M = k, Y_k \in A) \\
 &= \prod_{i=1}^{k-1} P(U_i > \frac{f(Y_i)}{cg(Y_i)}) P(U_k \leq \frac{f(Y_k)}{cg(Y_k)}, Y_k \in A) \\
 &= \prod_{i=1}^{k-1} (1 - P(U_i \leq \frac{f(Y_i)}{cg(Y_i)})) \frac{1}{c} \int_A f(p) dp \\
 &= (1 - \frac{1}{c})^{k-1} \frac{1}{c} \int_A f(p) dp = P(M = k) \int_A f(p) dp
 \end{aligned}$$

De lo que deducimos que M y $X' = Y_M$ son independientes, que la distribución de probabilidad de X' viene dada por $P(X' \in A) = \int_A f(p) dp$, y que por tanto la función de densidad de X' es f , concluyendo que X' sigue la misma distribución que X .

■

Supongamos que sabemos evaluar dos funciones continuas $f^* : \mathbb{R}^n \rightarrow \mathbb{R}_0^+$ y $g^* : \mathbb{R}^n \rightarrow \mathbb{R}_0^+$, que $f^*(p) \leq c'g^*(p)$, $\forall p \in \mathbb{R}^n$ para cierto $c' \geq 1$, que se pueden tomar muestras tales que su función de densidad es $g = \frac{g^*}{\int_{\mathbb{R}^n} g^*(s)ds}$ y que queremos generar muestras cuya función de densidad sea $f = \frac{f^*}{\int_{\mathbb{R}^n} f^*(s)ds}$. Una ventaja de este método es que se puede aplicar a las funciones g^* y f^* sin necesidad de calcular su integral, considerando la constante c' . En efecto, basta tomar f , g , y $c = \frac{c' \int_{\mathbb{R}^n} g^*(s)ds}{\int_{\mathbb{R}^n} f^*(s)ds}$ para ver que el método sigue siendo válido.

Por otro lado, uno de los problemas de este algoritmo que nos encontramos al aplicarlo en múltiples dimensiones es hallar la constante c . Además al aumentar la dimensionalidad usualmente la constante será grande, con lo que el ratio de aceptación $\frac{1}{c}$ será bajo, y rechazaremos muchas muestras antes de encontrar una válida. Otro problema de este método es que necesitamos una función de densidad g que, multiplicada por una constante, sea mayor o igual que la función objetivo, y esto no siempre será posible.

El método de rechazo se usa en renderización para depurar algoritmos que hacen uso de otros métodos de Monte Carlo. El ejemplo más clásico de aplicación del método de rechazo es generar muestras uniformemente distribuidas en un círculo de la siguiente manera.

Ejemplo 3.2 (Muestreando un círculo). *Supongamos que queremos tomar muestras uniformes en el círculo de radio unidad centrado en el origen. Este procedimiento se puede generalizar fácilmente a radio y centro arbitrarios.*

Deducimos que la función de densidad que queremos muestrear es $f(x, y) = \frac{4}{\pi}$, $\forall (x, y) \in \mathbb{R}^2$ tal que $x^2 + y^2 \leq 1$.

Por otro lado, como ya se mencionó, en muchos lenguajes de programación es usual disponer de un generador de números pseudoaleatorios, con lo que podemos generar fácilmente muestras siguiendo la distribución uniforme unidimensional en $[0, 1]$. Sean v_1, v_2 dos muestras siguiendo la distribución uniforme en $[0, 1]$, entonces (v_1, v_2) sigue la distribución uniforme en $[0, 1]^2$. Por tanto la otra función de densidad que utilizaremos para aplicar el método de rechazo es la distribución $g(x, y) = 1$, $\forall (x, y) \in [0, 1]^2$. Claramente $f(x, y) \leq \frac{4}{\pi}g(x, y)$, $\forall (x, y) \in \mathbb{R}^2$

Como vemos, dado $(x, y) \in [0, 1]^2$, $\frac{f(x,y)}{\frac{4}{\pi}g(x,y)} = 1$ si $x^2 + y^2 \leq 1$, $\frac{f(x,y)}{\frac{4}{\pi}g(x,y)} = 0$ en otro caso. Por tanto no hay que generar muestras uniformes adicionales, si la muestra generada está dentro del círculo unidad la aceptaremos y la rechazaremos en caso contrario.

3.3.3. Método de Metropolis

En primer lugar vamos a describir el método. Despues introduciremos una serie de definiciones y teoremas que nos llevarán a la demostración de que el método produce los resultados esperados.

Sea (Γ, \mathcal{A}, P) un espacio probabilístico, y sea X un vector aleatorio continuo definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) , $E \subseteq \mathbb{R}^n$. Consideremos f_X la función de densidad de X , y consideremos una función $T : E \times E \rightarrow \mathbb{R}_0^+$ tal que $T_x := T(x, \cdot)$ es una función de densidad en E para todo $x \in E$. Supongamos que $T_x(s) = 0$ para todo s tal que $f_X(s) = 0$. Supongamos también que el vector aleatorio con función de densidad T_x es fácil de simular para todo $x \in E$. Consideramos la función $a : E \times E \rightarrow \mathbb{R}_0^+$ definida de la siguiente manera:

$$a(x, x') = \min\left\{1, \frac{f_X(x')T(x', x)}{f_X(x)T(x, x')}\right\}$$

Entonces el algoritmo básico del método de Metropolis para generar N muestras es el descrito a continuación.

1. Partimos de un punto x_i , con $i = 0$.
2. Generamos una muestra x' con función de densidad $T(x_i, \cdot)$, y una muestra v con distribución uniforme en $[0, 1]$.
3. Calculamos $a = \min\left\{1, \frac{f_X(x')T(x', x_i)}{f_X(x_i)T(x_i, x')}\right\}$

4. Si $v < a$, entonces $x_{i+1} = x'$. En caso contrario, $x_{i+1} = x_i$.
5. Tomamos $i = i + 1$, y si $i \leq N$, repetimos desde 2.

Este método nos devolverá un conjunto de muestras $\{x_n\}_{n \in \{1, \dots, N\}}$. Vamos ahora a enunciar la notación, las definiciones y los resultados necesarios para demostrar la convergencia del método hacia la distribución objetivo.

Definición 3.24 (Filtración natural). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico en tiempo discreto definido sobre (Γ, \mathcal{A}, P) . Entonces se denomina filtración natural de $\{X_n\}_{n \in \mathbb{N}_0}$ a la sucesión de σ -álgebras $\{\sigma(X_0, \dots, X_n)\}_{n \in \mathbb{N}_0}$.*

Definición 3.25 (Proceso de Markov respecto a la filtración natural). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso estocástico en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Entonces se dirá que $\{X_n\}_{n \in \mathbb{N}_0}$ es un proceso de Markov si lo es respecto a su filtración natural, es decir, si cumple que:*

$$\forall n \in \mathbb{N}, \forall B \in \mathcal{B}_E, P(X_n \in B / X_0, \dots, X_{n-1}) = P(X_n \in B / X_{n-1})$$

Notación 3.5. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $X = (X_1, \dots, X_n)$ un vector aleatorio definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) , $E \subseteq \mathbb{R}^n$. Entonces para $y = (y_1, \dots, y_n) \in \mathbb{R}^n$, notaremos por $\{X \leq y\}$ al conjunto $\{X_1 \leq y_1, \dots, X_n \leq y_n\}$, y notaremos por I_y al conjunto $(-\infty, y_1] \times \dots \times (-\infty, y_n]$.*

Definición 3.26 (Proceso de Markov homogéneo). *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Entonces se dirá que $\{X_n\}_{n \in \mathbb{N}_0}$ es un proceso de Markov homogéneo si cumple que:*

$$\forall n \in \mathbb{N}, \forall B \in \mathcal{B}_E, \forall x \in E, P(X_n \in B / X_{n-1} = x) = P(X_1 \in B / X_0 = x)$$

En cuyo caso definimos la función de transición en un paso, $P(x, B)$, como:

$$\forall B \in \mathcal{B}_E, \forall x \in E, P(x, B) = P(X_1 \in B / X_0 = x)$$

De igual forma, definimos la función de distribución de transición en un paso como:

$$\forall y \in \mathbb{R}^n, \forall x \in E, F(y/x) = P(X_1 \leq y / X_0 = x)$$

Y, por último, la función de densidad de transición en un paso, $f(y/x)$, se define como la función que cumple que:

$$\forall y \in \mathbb{R}^n, \forall x \in E, \int_{I_y} f(s/x) ds = F(y/x)$$

Proposición 3.6. Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Entonces las probabilidades de transición en n pasos son independientes del instante de tiempo en el que se realiza la transición, es decir:

$$\forall n, m \in \mathbb{N}, \forall B \in \mathcal{B}_E, \forall x \in E, P(X_{m+n} \in B | X_m = x) = P(X_n \in B | X_0 = x)$$

Definición 3.27 (Distribución de transición en k pasos). Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Definimos la función de transición en k pasos, $P_k(x, B)$, como:

$$\forall n \in \mathbb{N}, \forall B \in \mathcal{B}_E, \forall x \in E, P_k(x, B) = P(X_k \in B | X_0 = x)$$

Definición 3.28 (Distribución estacionaria). Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Una distribución η sobre (E, \mathcal{B}_E) se dice que es estacionaria frente a $P(x, B)$ si cumple que:

$$\forall B \in \mathcal{B}_E, \eta(B) = \int_E P(x, B) \eta(dx)$$

Una función de distribución G sobre \mathbb{R}^n es estacionaria frente a la función de distribución de transición $F(y/x)$ si cumple que:

$$\forall y \in \mathbb{R}^n, G(y) = \int_E F(y/x) dG(x)$$

Una función de densidad g sobre \mathbb{R}^n es estacionaria frente a la función de densidad de transición $f(y/x)$ si cumple que:

$$\forall y \in \mathbb{R}^n, g(y) = \int_E f(y/x) g(x) dx$$

Definición 3.29 (Proceso de Markov irreducible). Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Se dice que $\{X_n\}_{n \in \mathbb{N}_0}$ es σ -irreducible si existe una medida no nula σ -finita Φ definida sobre (E, \mathcal{B}_E) tal que para todo $B \in \mathcal{B}_E$, con $\Phi(B) > 0$, y para todo $x \in E$, existe un entero positivo $k = k(x, B)$ tal que $P_k(x, B) > 0$.

Definición 3.30 (Proceso de Markov aperiódico). Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Supongamos que η es una distribución estacionaria. Se dice que $\{X_n\}_{n \in \mathbb{N}_0}$ es aperiódico si no existe $d \geq 2$ y conjuntos disjuntos $S_1, \dots, S_d \in \mathcal{B}_E$ cumpliendo que:

$$P(x, S_{i+1}) = 1, \forall x \in S_i, \forall i \in \{1, \dots, d-1\}$$

$$P(x, S_1) = 1, \forall x \in S_d$$

$$\eta(S_1) > 0$$

Definición 3.31. Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Se dice que $\{X_n\}_{n \in \mathbb{N}_0}$ es reversible con respecto a una función de densidad g definida sobre \mathbb{R}^n si cumple que:

$$g(x)f(y/x) = g(y)f(x/y), \forall x, y \in E \quad (3.6)$$

En cuyo caso diremos que g cumple la propiedad de balance detallado.

Definición 3.32 (Función delta de Dirac). La función delta de Dirac, notada por $\delta : \mathbb{R} \rightarrow \mathbb{R}$, es una función que está definida de tal manera que cumple que:

$$\delta(x) = 0, \forall x \neq 0 \quad \int_{\mathbb{R}} \delta(x) dx = 1$$

Una importante propiedad es la función delta de Dirac es que dado $x_0 \in \mathbb{R}$, $f : \mathbb{R} \rightarrow \mathbb{R}$:

$$\int_B \delta(x - x_0) f(x) dx = \mathbf{1}_B(x_0) f(x_0), \forall B \in \mathcal{B}$$

Esta función se puede generalizar a \mathbb{R}^n , definiéndola de la siguiente manera:

$$\delta_n(x_1, \dots, x_n) = \delta(x_1) \dots \delta(x_n), \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$

Manteniéndose la propiedad de que, dado $x_0 \in \mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$\int_B \delta_n(x - x_0) f(x) dx = \mathbf{1}_B(x_0) f(x_0), \forall B \in \mathcal{B}$$

Una vez vistas estas definiciones, demostraremos la siguiente proposición.

Proposición 3.7. Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Si una función de densidad g definida sobre \mathbb{R}^n cumple la propiedad de balance detallado, entonces g es estacionaria.

Demostración. Basta integrar a un lado de la igualdad (3.6) y aplicarla:

$$\int_E g(x)f(y/x)dx = \int_E g(y)f(x/y)dx = g(y) \int_E f(x/y)dx = g(y), \forall y \in \mathbb{R}^n$$

donde se ha usado que, dado $y \in \mathbb{R}^n$, $f(\cdot/y)$ es una función de densidad, y por tanto tiene integral 1.



Y llegamos al teorema clave que nos permitirá demostrar que el método de Metropolis converge (ver demostración en [RR04]).

Teorema 3.8. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico. Sea $\{X_n\}_{n \in \mathbb{N}_0}$ un proceso de Markov homogéneo en tiempo discreto definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{D}) . Si \mathcal{D} es generada por un conjunto numerable y $\{X_n\}_{n \in \mathbb{N}_0}$ es Φ -irreducible, aperiódico y tiene una distribución estacionaria η . Entonces:*

$$\forall B \in \mathcal{B}_E, \lim_{n \rightarrow +\infty} P(X_n \in B) = \eta(B)$$

Ya tenemos todo lo necesario para demostrar la convergencia del método de Metropolis.

Teorema 3.9. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico, y sea X un vector aleatorio continuo definido sobre (Γ, \mathcal{A}, P) con valores en (E, \mathcal{B}_E) . Sea f_X la función de densidad de X , y sea P_X la distribución de probabilidad de X . Consideremos una función $T : E \times E \rightarrow \mathbb{R}_0^+$ cumpliendo que $T(y, \cdot)$ es una función de densidad en E para todo $y \in E$. Supongamos que, dado $y \in E$, $T(y, s) = 0$ para todo s tal que $f_X(s) = 0$. Consideramos la función $a : E \times E \rightarrow \mathbb{R}_0^+$ definida de la siguiente manera:*

$$a(x, x') = \min\left\{1, \frac{f_X(x')T(x', x)}{f_X(x)T(x, x')}\right\}$$

Sea $x^{(0)} \in E$, con $f_X(x^{(0)}) > 0$. Consideramos la sucesión $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ generada por el algoritmo (1) iterando infinitas veces. Entonces $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ es un proceso de Markov homogéneo y P_X es una distribución estacionaria. Supongamos que $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ es P_X -irreducible y aperiódico. Entonces se verifica que:

$$\forall B \in \mathcal{B}_E, \lim_{n \rightarrow +\infty} P_n(\cdot, B) = P_X(B), \text{ c.s.-}P_X$$

En concreto,

$$\forall B \in \mathcal{B}_E, \lim_{n \rightarrow +\infty} P_n(x^{(0)}, B) = P_X(B)$$

Demostración. Claramente $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ es un proceso de Markov, ya que cada elemento de la sucesión es generado a partir del elemento anterior. También es fácil ver que es homogéneo, pues dado $x^{(i)}$, $i \in \mathbb{N}_0$, el elemento $x^{(i+1)}$ se genera según la función de densidad $T(x^{(i)}, \cdot)$, que no tiene ninguna dependencia temporal. Por tanto vamos a calcular la función de densidad de transición en un paso. Dada una muestra $x \in E$, generada según la función de densidad $T(x^{(i)}, \cdot)$, entonces $a(x^{(i)}, x)$ representa la probabilidad de que x sea aceptada, y, por tanto, $x^{(i+1)} = x$. Teniendo esto en cuenta vemos que:

$$\forall B \in \mathcal{B}_E, x \in E/B, P(x^{(i+1)} \in B / x^{(i)} = x) = \int_B T(x, y) a(x, y) dy$$

$$\begin{aligned} \forall B \in \mathcal{B}_E, x \in B, P(x^{(i+1)} \in B/x^{(i)} = x) &= \int_B T(x, y)a(x, y)dy + \\ &\quad + (1 - \int_E T(x, s)a(x, s)ds) \end{aligned}$$

donde $h(x) := (1 - \int_E T(x, s)a(x, s)ds)$ representa la probabilidad de rechazar la muestra generada según $T(x, \cdot)$. Vemos por tanto que podemos expresar la función de densidad de transición utilizando la función delta de Dirac definida en (3.32):

$$f(y/x) = T(x, y)a(x, y) + \delta_n(y - x)h(y), \forall x, y \in E$$

Vamos a demostrar ahora que f_X satisface la propiedad de balance detallado. Sean $x, y \in E, x \neq y$:

$$\begin{aligned} f_X(x)f(y/x) &= f_X(x)T(x, y)a(x, y) = f_X(x)T(x, y) \min\{1, \frac{f_X(y)T(y, x)}{f_X(x)T(x, y)}\} \\ &= \min\{f_X(x)T(x, y), f_X(x)T(x, y)\frac{f_X(y)T(y, x)}{f_X(x)T(x, y)}\} \\ &= \min\{f_X(x)T(x, y), f_X(y)T(y, x)\} = f_X(y)f(x/y) \end{aligned}$$

Por tanto f_X es una función de densidad estacionaria, o lo que es lo mismo, P_X es una distribución estacionaria. Concluimos el resultado usando la hipótesis de que $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ es P_X -irreducible y aperiódico, aplicando el teorema (3.8). ■

El teorema anterior nos asegura que, para un número de muestras suficientemente alto, las muestras generadas mediante el método de metrópolis seguirán la distribución que se quiere muestrear. Como vemos el punto inicial puede ser generado siguiendo cualquier distribución de probabilidad, la única restricción es que debe cumplir que $f_X(x^{(0)}) > 0$. Intuitivamente, la hipótesis de que $\{x^{(n)}\}_{n \in \mathbb{N}_0}$ sea P_X -irreducible se obtiene eligiendo una función T que nos permita llegar a cualquier conjunto $B \subseteq E$, con $P_X(B) > 0$, en un número finito de iteraciones del algoritmo, lo cual es una característica deseable ya que nos permite muestrear el espacio de estados completo. La aperiodicidad, por otro lado, es una propiedad que casi siempre se cumple y que es sencilla de comprobar. El método de Metropolis corresponde a una clase de algoritmos llamados *Markov chain Monte Carlo*, o *MCMC*.

Un problema asociado a este método es que a diferencia de los métodos anteriores, las muestras generadas no son independientes, ya que el algoritmo genera la muestra $x^{(i+1)}$ a partir de la muestra $x^{(i)}$. Es por ello que es posible que tengamos que generar muchas más muestras de las que vamos a utilizar

con el objetivo de que las muestras sean independientes entre sí. El número de muestras que desecharemos antes de aceptar una se elegirá en función de la correlación que haya entre muestras adyacentes.

Otro problema derivado de este método es que aunque el proceso de Markov acaba convergiendo a la distribución buscada, las primeras muestras pueden no ser representativas de dicha distribución, y esto se produce debido a que la muestra inicial está tomada respecto de una distribución diferente. Es por esto que suele ser necesario un periodo inicial en el que todas las muestras se desechan.

Una ventaja de este método es que su ratio de convergencia y su tiempo de ejecución no se ven demasiado perjudicados al aumentar la dimensionalidad del espacio muestrado. Esto hace que a veces este método, así como otros algoritmos de tipo MCMC, sean la única solución viable para aproximar una distribución en varias dimensiones. Dentro de este método es muy importante la elección de la función T , ya que de ello dependerá la velocidad de convergencia del proceso de Markov hacia la distribución deseada.

En cuanto a su aplicación en renderización, el método de Metropolis puede usarse para mejorar la eficiencia de los algoritmos bidireccionales, que son algoritmos que aprovechan el carácter reversible de la óptica geométrica para trazar caminos entre dos puntos, simulando el camino de la luz partiendo de ambos puntos y juntando los dos caminos generados en un punto intermedio de la escena.

3.3.4. Transformación entre distribuciones

Habrá ocasiones en que nos resulte costoso generar muestras de un vector aleatorio Y , pero exista una función g biyectiva y diferenciable tal que $Y = g(X)$, con X fácilmente simulable. Entonces podremos generar muestras que sigan la distribución de Y aplicando un cambio de variable, de la manera en que se especifica en la siguiente proposición.

Proposición 3.8. *Sea (Γ, \mathcal{A}, P) un espacio probabilístico, y sean X, Y dos vectores aleatorios continuos definido sobre (Γ, \mathcal{A}, P) , con valores en E_X, E_Y respectivamente. Sea f_Y la función de densidad de X , y sea g una función biyectiva y diferenciable tal que $Y = g(X)$. Entonces la función de densidad de X cumple que:*

$$f_X(x) = f_Y(g(x))|det(J_g(x))|, \forall x \in E_X$$

donde $|det(J_g(x))|$ denota al valor absoluto del determinante del Jacobiano de g en x .

Detallaremos ahora dos transformaciones que tienen gran importancia en renderización. Recordemos que notamos por π a la proyección sobre la esfera unidad \mathbb{S}^2 . En el contexto de la proposición 1.2, consideremos un vector aleatorio ω con valores en la porción de la esfera $\pi(C)$. Claramente, en dicho contexto, la función π restringida a C , $\pi|_C$, es una biyección entre la superficie C y la porción de la esfera $\pi(C)$. Supongamos que conocemos la función de densidad de ω , f_ω , y que queremos tomar muestras en C siguiendo la distribución de $X := (\pi|_C)^{-1}(\omega)$. Entonces tenemos que, dado $A \subseteq C$:

$$\begin{aligned} P(\omega \in \pi(A)) &= \int_{\pi(A)} f_\omega(s) d\mu(s) = \int_A f_\omega(\pi(x)) \frac{x}{\|x\|^3} dS \\ &= \int_A f_\omega(\pi(x)) \frac{x \cdot n(x)}{\|x\|^3} dS = \int_A f_\omega(\pi(x)) \frac{\cos(x, n(x))}{\|x\|^2} dS \\ &= P(X \in A) \end{aligned}$$

donde $n(x)$ es la normal a la superficie en el punto x , y $\cos(x, n(x))$ es el coseno del ángulo formado por la normal a la superficie en x y el vector x . Por tanto concluimos que:

$$f_X(x) = f_\omega(\pi(x)) \frac{\cos(x, n(x))}{\|x\|^2} \quad (3.7)$$

Por otro lado, consideremos la proyección sobre el plano XY del hemisferio superior de la esfera unidad centrada en el origen, $\rho : \mathcal{H}^2 \rightarrow \mathbb{R}^3$, con:

$$\rho(x, y, z) = (x, y, 0), \forall (x, y, z) \in \mathcal{H}^2 \quad (3.8)$$

Notar que esta proyección depende del sistema de referencia usado. Sea \mathbb{D} el disco unidad con centro el origen contenido en el plano XY . Sea $A \subseteq \mathcal{H}^2$ un conjunto medible, entonces es fácil demostrar que se cumple que dada una función $g : \mathbb{D} \rightarrow \mathbb{R}_0^+$ medible:

$$\int_A g(\rho(\omega)) \cos(\omega, (0, 0, 1)) d\mu(\omega) = \int_{\rho(A)} g dS$$

Consideremos ahora un vector aleatorio X con valores en \mathbb{D} . Claramente la función ρ es una biyección entre el hemisferio \mathcal{H}^2 y el disco \mathbb{D} . Supongamos que conocemos la función de densidad de X , f_X , y que queremos conocer la distribución de $\omega := (\rho^{-1}(X))$. Entonces tenemos que, por la igualdad anterior:

$$f_\omega(\omega) = f_X(\rho(\omega)) \cos(\omega, (0, 0, 1)) \quad (3.9)$$

Notar que dado que trabajamos en un espacio afín, la definición de ρ y las igualdades anteriores dependen del sistema de referencia usado.

3.4. Métodos para reducir la varianza

En esta sección presentaremos tres métodos que nos permiten reducir la varianza del estimador de Monte Carlo, los cuales están recogidos en [Vea98]. Estos tres métodos tienen una gran importancia en renderización.

3.4.1. Muestreo de importancia

El muestreo de importancia se basa en una idea bastante sencilla. Sea X un vector aleatorio, supongamos que queremos aproximar $E[g(X)]$ para una función medible g con valores en \mathbb{R} . Supongamos que X es continuo y tiene función de densidad f_X . Sea Y otro vector aleatorio con función de densidad f_Y estrictamente positiva. Entonces vemos que:

$$E[g(X)] = \int g(x)f_X(x)dx = \int \frac{g(x)}{f_Y(x)}f_Y(x)f_X(x)dx = E\left[\frac{g(Y)}{f_Y(Y)}f_X(Y)\right]$$

Consideramos la variable aleatoria $Z := \frac{g(Y)}{f_Y(Y)}f_X(Y)$. Entonces:

$$\text{Var}(Z) = \int \frac{g(x)^2}{f_Y(x)}f_X(x)^2dx - (E[g(X)])^2$$

El muestreo de importancia consiste en buscar una función de densidad f_Y muestreable tal que la varianza de Z sea considerablemente más pequeña que la de $g(X)$. Teniendo en cuenta la igualdad anterior, la varianza de Z será menor que la de $g(X)$ cuando $E[Z^2]$ sea menor que $E[g(X)^2]$. En ese caso, será preferible muestrear la variable aleatoria Y y aproximar $E[Z]$ con el estimador Monte Carlo, ya que, como vimos, la varianza del estimador Monte Carlo es proporcional a la varianza de la función cuya esperanza approxima. Al reducir la varianza del estimador, se requerirán menos simulaciones para acercarse al valor buscado.

Como ya se mencionó al final del capítulo anterior, habrá situaciones en las que en vez de la esperanza de una función queramos calcular su integral. Supongamos que queremos calcular la integral de una función h , y sea X una variable aleatoria tal que $f_X(x) > 0$ para todo x tal que $|h(x)| > 0$. Tomando $g = \frac{h}{f_X}$, podemos aproximar la integral de h aplicando el estimador Monte Carlo a la función g . El muestreo de importancia en este caso se reduce a elegir la variable X de tal manera que la esperanza de g al cuadrado, $E[g(X)^2] = \int \frac{h(x)^2}{f_X(x)}dx$, sea baja, y esto se consigue mediante funciones de densidad que asignen mayores probabilidades a regiones donde h tome valores altos.

3.4.2. Estratificación o muestreo estratificado

El muestreo estratificado consiste en separar el dominio de integración en regiones disjuntas, llamadas estratos, y tomar un número fijo de muestras en cada estrato. Supongamos que queremos aproximar la esperanza de una función f en un dominio de integración Λ . Consideremos N regiones disjuntas, $\Lambda_1, \dots, \Lambda_N \subset \Lambda$, cumpliendo que:

$$\bigcup_{i=1}^N \Lambda_i = \Lambda$$

Consideramos $n_1, \dots, n_N \in \mathbb{N}_0$ el número de muestras que se tomarán en cada estrato, y p_1, \dots, p_N las funciones de densidad definidas sobre cada estrato. Esto lleva a un estimador de la forma:

$$F = \sum_{i=1}^N v_i \frac{1}{n_i} \sum_{j=1}^{n_i} f(X_{i,j})$$

donde v_i es el volumen fraccional del estrato i ($\sum v_i = 1$, $v_i \in (0, 1]$), y $X_{i,j}$ es la j -ésima muestra independiente tomada siguiendo la función de densidad p_i . Mediante una serie de derivaciones sencillas, suponiendo que $n_i = v_i M$ donde M es el número total de muestras, puede verse que:

$$\text{Var}(F) = \frac{1}{N} \sum v_i \sigma_i^2$$

donde σ_i^2 es la varianza de f en Λ_i . La varianza del estimador tomando M muestras sin estratificación cumple que (ver demostración en [Vea98]):

$$\text{Var}(F_M) = \frac{1}{M} [\sum v_i \sigma_i^2 + \sum v_i (\mu_i - I)^2]$$

donde μ_i es la esperanza de f en Λ_i , e I es la esperanza de f en el dominio completo. Dado que la suma de la derecha es no negativa, deducimos que la estratificación nunca incrementa la varianza. De hecho la suma de la derecha solo valdrá 0 cuando la función f tenga la misma media en todos los estratos. La reducción en la varianza que produce este método viene del hecho de que impide que todas las muestras se acumulen en una región y permite muestrear el espacio de estados completo.

En los ray tracers la mayoría de distribuciones se simulan partiendo de una o varias muestras siguiendo la distribución uniforme en $[0, 1]^2$, por lo que suele usarse estratificación para generar dichas muestras uniformes.

3.4.3. Muestreo de importancia múltiple

El muestreo de importancia múltiple se puede ver como una extensión del muestreo estratificado en la que no es necesario que los estratos sean disjuntos. Este método consiste en tomar muestras de diferentes distribuciones y combinarlos de manera que nos quede un estimador insesgado. Consideraremos un conjunto p_1, \dots, p_N de funciones de densidad, un conjunto w_1, \dots, w_N de pesos y un conjunto n_1, \dots, n_N de enteros. El estimador multi-muestra de la integral de f viene dado por:

$$F = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} w_i(X_{i,j}) \frac{f(X_{i,j})}{p_i(X_{i,j})} \quad (3.10)$$

donde $X_{i,j}$ es la j -ésima muestra independiente tomada siguiendo la función de densidad p_i . Enunciamos una proposición que nos indica las características que deben cumplir los pesos para que el estimador sea insesgado.

Proposición 3.9. *Sea F el estimador definido en 3.10. Supongamos que las funciones w_i cumplen las siguientes condiciones:*

- $\sum_{i=1}^n w_i(x) = 1$ para todo x tal que $f(x) \neq 0$
- $w_i(x) = 0$ para todo x tal que $p_i(x) = 0$.

Entonces se cumple que:

$$E[F] = \int f(x)dx$$

Demostración. En efecto:

$$E[F] = \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \int w_i(x) \frac{f(x)}{p_i(x)} p_i(x) dx = \int \sum_{i=1}^N w_i(x) f(x) dx = \int f(x) dx$$

■

Este método es especialmente útil cuando el integrando que se quiere aproximar está compuesto por el producto de dos funciones. En ese caso se pueden utilizar dos métodos de muestreo diferentes, haciendo que cada uno se centre en muestrear cada una de las funciones. Es clara la importancia de este método en renderización, ya que en la ecuación de renderización habrá direcciones en las que la radiancia sea alta y la función BSDF no y, por el contrario, habrá direcciones en las que la función BSDF sea alta y la radiancia no. El muestreo de importancia múltiple nos permite combinar el muestreo de la radiancia y de la función de dispersión para así reducir la varianza.

Capítulo 4

Aproximación de la ecuación de renderización

En este capítulo describiremos dos posibles formas de aproximar la ecuación de renderización descrita en 1.5.

4.1. Iluminación directa

Se trata del método de aproximación más sencillo de la ecuación de renderización, ya que no tiene en cuenta la iluminación indirecta. Una vez se calcula la intersección del rayo de cámara con la escena, desde ese punto se muestran todas las fuentes de luz de la escena. Se denomina muestrear una fuente de luz al proceso de generar muestras sobre el conjunto de direcciones que apuntan hacia la fuente de luz, o lo que es lo mismo, sobre el ángulo sólido subtendido por esta. Por tanto se estima la radiancia en un punto $P \in \mathbb{R}^3$ y en una dirección $\omega_o \in \mathbb{S}^2$ como:

$$L(P, \omega_o) = \sum_{F \in \mathcal{F}} \int_{\pi_P(F)} f(P, \omega_o, \omega_i) L_e(P, \omega_i) |n_P \cdot \omega_i| d\mu(\omega_i),$$

donde \mathcal{F} es el conjunto de todas las fuentes de luz de la escena, π_P es la proyección sobre la esfera unidad con centro P y n_P es la normal a la superficie en P . La mayoría de fuentes de luz serán fuentes de luz de área, ya que son las fuentes de luz que más se aproximan a las fuentes de luz reales. El método más sencillo para muestrear una fuente de luz de área es generar una muestra que siga una distribución uniforme en toda su superficie y transformar la función de densidad, que será constante, para obtener la función de densidad respecto al ángulo sólido, en virtud de la ecuación 3.7. Sin embargo esta forma de muestreo puede llegar a implicar alta varianza

dependiendo de la escena, especialmente en puntos que están muy cerca de la fuente de luz muestreada. En el siguiente capítulo veremos diferentes formas de muestrear fuentes de luz en el caso de ciertas figuras geométricas.

Si solo se muestrean direcciones correspondientes a fuentes de luz, aunque la radiancia emitida en esas direcciones sea alta, puede que la función BSDF tome valores bajos en esas direcciones y por tanto nos lleve a una mala aproximación del color en el punto. Es por esto que se suele aplicar muestreo de importancia múltiple en este tipo de algoritmo, muestreando también la función BSDF. Un ray tracer siempre deberá ser capaz de generar muestras siguiendo la distribución de los diferentes tipos de BSDF que simula. Además debe tenerse en cuenta si el punto que está siendo sombreado pertenece a una superficie puramente especular, en cuyo caso se trazan rayos recursivamente en las correspondientes direcciones, fijando un número máximo de rayos trazados.

4.2. Path-Tracing

Antes de estudiar el enfoque tomado en este algoritmo y su justificación, presentaremos una serie de definiciones y resultados previos, recogidos en [Fre].

Definición 4.1 (Operador lineal). *Sean S, M dos espacios normados sobre el mismo cuerpo \mathbb{K} , un operador lineal de S en M es una aplicación $T : S \rightarrow M$ cumpliendo que:*

$$T(\lambda u + v) = \lambda T(u) + T(v) \quad \forall u, v \in S, \forall \lambda \in \mathbb{K}$$

Teorema 4.1 (Teorema del punto fijo). *Sea S un espacio de Banach no vacío, y sea $T : S \rightarrow S$ un operador lipschitziano con constante de lipschitz $l < 1$. Consideremos $f_0 \in S$ arbitrario, y sea $\{f_n\}_{n \in \mathbb{N}_0}$ una sucesión de elementos de S cumpliendo que $f_n = T(f_{n-1})$, $\forall n \in \mathbb{N}$. Entonces $\{f_n\}_{n \in \mathbb{N}_0}$ converge al único punto fijo de T , $f \in S$ tal que $T(f) = f$.*

Definición 4.2 (Operador integral de Fredholm). *Sea $A \subseteq \mathbb{R}^n$, sea (A, \mathcal{A}, ρ_1) un espacio de medida, y sea $k : A \times A \rightarrow \mathbb{R}$ una función absolutamente integrable ($k \in L_1(A \times A)$) y acotada. Entonces se define el operador integral de Fredholm como:*

$$K_k : L_1(A) \rightarrow L_1(A)$$

$$f \rightarrow K_k(f) = \int_A k(\cdot, y) f(y) dy$$

Definición 4.3 (Ecuación integral de Fredholm de segundo tipo). *Sea $A \subseteq \mathbb{R}^n$, sea (A, \mathcal{A}, ρ_1) un espacio de medida, y sea $k : A \times A \rightarrow \mathbb{R}$ una función*

absolutamente integrable ($k \in L_1(A \times A)$) y acotada. Sea K_k el operador integral de Fredholm, y sea $f : A \rightarrow \mathbb{R}$ una función. Supongamos que queremos hallar una función $u : A \rightarrow \mathbb{R}$ satisfaciendo:

$$u = f + K(u)$$

A esta ecuación la llamamos ecuación integral de Fredholm de segundo tipo.

Teorema 4.2. Consideramos la ecuación integral de Fredholm de segundo tipo que acabamos de definir:

$$u = f + K(u)$$

Entonces si K es lipschitziano con constante $l < 1$, la ecuación tiene solución. Dicha solución viene dada por la suma de la serie de Neumann:

$$u = \sum_{i=0}^{\infty} K^i(f)$$

donde $K^i := K \circ \dots \circ K$.

*Demuestra*ción. Supongamos que K es lipschitziano con constante $l < 1$. Consideramos el operador $T(s) := K(s) + f$. Como podemos ver:

$$\|Tu_1 - Tu_2\| = \|Ku_1 - Ku_2\| \leq l\|u_1 - u_2\|, \forall u_1, u_2 \in A$$

Por tanto T también es un operador lipschitziano con constante $l < 1$, por lo que podemos aplicar el teorema del punto fijo de Banach. Tomamos $u_0 \in L_1(A)$ arbitrario, y definimos la sucesión $\{u_n\}_{n \in \mathbb{N}_0}$ como $u_n := T(u_{n-1}) = K(u_{n-1}) + f, \forall n \in \mathbb{N}$. Es fácil ver que:

$$u_n = \sum_{i=0}^{n-1} K^i(f) + K^n(u_0)$$

Entonces $\{u_n\}_{n \in \mathbb{N}_0}$ convergerá al único punto fijo de T , $u \in L_1(A)$. Vemos por otro lado que $\lim_{n \rightarrow +\infty} K^n(u_0) = 0$ por ser K contractiva. Concluimos que $\lim_{n \rightarrow +\infty} u_n = \lim_{n \rightarrow +\infty} \sum_{i=0}^{n-1} K^i(f) = u$ es la única solución de la ecuación de Fredholm, como queríamos.

■

Ya podemos desarrollar el enfoque tomado en este algoritmo, siguiendo las ideas descritas en [DGP05]. Consideramos la función $t : \mathbb{R}^3 \times \mathbb{S}^2 \rightarrow \mathbb{R}^3$

que asigna a cada rayo (definido por su punto inicial y su dirección) su intersección más cercana con la escena. Es fácil ver que entonces $L_i(p, \omega) = L_s(t(p, \omega), -\omega)$, $\forall p \in \mathbb{R}^3$, $\forall \omega \in \mathbb{S}^2$. Por tanto, notando $L = L_s$ por comodidad, y fijando $p \in \mathbb{R}^3$, podemos reescribir la ecuación de la siguiente manera:

$$L(p, \omega_o) = L_e(p, \omega_o) + \int_{\mathbb{S}^2} f(p, \omega_o, \omega_i) L(t(p, \omega_i), -\omega_i) |n \cdot \omega_i| d\mu(\omega_i) \quad (4.1)$$

Y podemos entender la ecuación anterior como una ecuación de Fredholm en la que el operador integral es:

$$K(g) := \int_{\mathbb{S}^2} f(p, \cdot, \omega_i) g(t(p, \omega_i), -\omega_i) |n \cdot \omega_i| d\mu(\omega_i)$$

Como vemos, el operador K es lineal. Vemos que en cualquier sistema físicamente realista se cumple que K es lipschitziano con constante $l < 1$, respecto a una norma adecuada, a causa de la absorción de energía por parte de las superficies. Por tanto podemos aproximar la función L a partir de la serie de Neumann:

$$L_k = \sum_{i=0}^k K^i(L_e), \quad \forall k \in \mathbb{N}_0 \quad (4.2)$$

La radiancia media incidente sobre un píxel P de la imagen final se puede calcular de la siguiente manera:

$$\bar{L}_P = \frac{1}{|P|} \int_{S(P)} L(t(x_{camara}, \omega), -\omega) d\mu \quad (4.3)$$

donde $S(P)$ representa el conjunto de direcciones que atraviesan el píxel P , x_{camara} es la posición de la cámara, y $|P|$ es el área del píxel P . Consideramos ahora el problema de evaluar el siguiente operador:

$$J_g(L) = \int_{\mathbb{S}^2} g(\omega) L(t(x_{camara}, \omega), -\omega) d\mu \quad (4.4)$$

donde la función g es no negativa y pertenece a $L_\infty(\mathbb{S}^2)$. Es de especial interés el caso $g(\omega) = \frac{\mathbf{1}_{S(P)}(\omega)}{|P|}$, $\forall \omega \in \mathbb{S}^2$, que nos lleva a la radiancia media sobre el píxel P . Aplicaremos un estimador de Monte Carlo para aproximar el operador 4.4, aprovechando el hecho de que la serie definida en 4.2 converge.

Supongamos que generamos una dirección ω_0 inicial siguiendo una función de densidad p_0 definida sobre \mathbb{S}^2 respecto al ángulo sólido. Vamos a introducir la notación que usaremos. En primer lugar, tomamos $x_0 := x_{camara}$. Fijamos $N \in \mathbb{N}$. Consideramos $\omega_j \in \mathbb{S}^2$ una dirección para todo

$j \in \{1, \dots, N\}$. Definimos $x_j := t(x_{j-1}, \omega_{j-1})$, n_j la normal a la superficie en x_j , para todo $j \in \{1, \dots, N+1\}$. Por último, para x_i, ω_i , definimos $R(x_i, \omega_i) = f(x_i, -\omega_{i-1}, \omega_i)|n_i \cdot \omega_i|$ para todo $i \in \{1, \dots, N\}$. Vemos que utilizando esta notación se tiene que, fijado $i \in \{1, \dots, N\}$:

$$K^i(L_e)(x_1, -\omega_0) = \int_{(\mathbb{S}^2)^i} R(x_1, \omega_1) \dots R(x_i, \omega_i) L_e(x_{i+1}, -\omega_i) d\mu(\omega_1) \dots d\mu(\omega_i) \quad (4.5)$$

Y, como ya sabemos, el N -ésimo término de la serie de Neumann cumple que:

$$L_N(x_1, -\omega_0) = L_e(x_1, -\omega_0) + \sum_{i=1}^N K^i(L_e)(x_1, -\omega_0) \quad (4.6)$$

Consideramos $\{p_j\}_{j \in \{1, \dots, N\}}$, donde p_j es una función de densidad sobre \mathbb{S}^2 definida respecto al ángulo sólido para todo $j \in \{1, \dots, N\}$. Supongamos que la dirección ω_j es una variable aleatoria generada según la función de densidad p_j para todo $j \in \{0, \dots, N\}$, y que todas las direcciones ω_j son independientes. Definamos los siguientes pesos:

$$W_0 = 1, W_j = W_{j-1} \frac{R(x_j, \omega_j)}{p_j(\omega_j)}, \forall j \in \{1, \dots, N\}$$

Dado que los ω_j son independientes, su función de densidad conjunta es el producto de las funciones de densidad p_j . Por tanto, teniendo en cuenta la igualdad en 4.5, es fácil ver que, bajo la distribución conjunta de $(\omega_1, \dots, \omega_N)$, el estimador $W_j L_e(x_{j+1}, -\omega_j)$, $j \in \{0, \dots, N\}$, cumple que:

$$E[W_j L_e(x_{j+1}, -\omega_j)] = K^j(L_e)(x_1, -\omega_0), j \in \{0, \dots, N\} \quad (4.7)$$

Aplicando el operador 4.4 al término N -ésimo de la serie de Neumann L_N , tenemos que:

$$\begin{aligned} J_g(L_N) &= \int_{\mathbb{S}^2} g(\omega_0)(L_e(x_1, -\omega_0) + \sum_{i=1}^N K^i(L_e)(x_1, -\omega_0)) d\mu(\omega_0) \\ &= \sum_{i=0}^N \int_{\mathbb{S}^2} g(\omega_0) K^i(L_e)(x_1, -\omega_0) d\mu(\omega_0) = \sum_{i=0}^N J_g(K^i(L_e)) \end{aligned}$$

Por lo que definimos el siguiente estimador:

$$P_N[g](\omega_0, \dots, \omega_N) = \frac{g(\omega_0)}{p_0(\omega_0)} \sum_{i=0}^N W_i L_e(x_{i+1}, -\omega_i) \quad (4.8)$$

El cual, bajo la distribución conjunta de $(\omega_0, \dots, \omega_N)$, cumple que:

$$\begin{aligned} E[P_N[g]] &= E\left[\frac{g(\omega_0)}{p_0(\omega_0)} \sum_{j=0}^N W_j L^e(x_{j+1}, -\omega_j)\right] \\ &= \sum_{j=0}^N E\left[\frac{g(\omega_0)}{p_0(\omega_0)} W_j L^e(x_{j+1}, -\omega_j)\right] = \sum_{j=0}^N J_g(K^j(L^e)) = J_g(L_N). \end{aligned}$$

Y acabamos de demostrar el siguiente teorema:

Teorema 4.3. *Dado $N \in \mathbb{N}_0$, bajo la distribución conjunta de $(\omega_0, \dots, \omega_N)$ se cumple que:*

$$E[P_N[g]] = J_g(L_N)$$

Por tanto, podríamos generar M muestras del vector aleatorio $(\omega_0, \dots, \omega_N)$, $\{(\omega_0, \dots, \omega_N)^{(i)}\}_{i \in \{1, \dots, M\}}$, y por la ley fuerte de los números grandes el estimador de Monte Carlo $\frac{1}{M} \sum_{i=1}^M P_N[g](\omega_0, \dots, \omega_N)^{(i)})$ converge al valor $J_g(L_N)$. A su vez, $J_g(L_N)$ tiende a $J_g(L)$ cuando N tiende a infinito debido a que la serie de Neumann converge, por lo que cuanto más grande sea el vector de direcciones que tomamos mejor será la aproximación de $J_g(L)$.

Intuitivamente, este enfoque consiste en trazar un camino de N rayos por la escena. Iniciamos trazando un rayo desde la cámara cuya dirección es elegida de manera aleatoria. Después calculamos el punto donde este rayo interseca la escena, y a partir de ese punto procedemos recursivamente a trazar un rayo con dirección aleatoria. Mientras vamos trazando el camino vamos sumando las componentes $W_j L^e(x_{j+1}, \omega_j)$ del estimador. Cuando hayamos trazado los N rayos, almacenamos el valor obtenido y procedemos de nuevo hasta tener M estimaciones con las que calculamos el valor del estimador de Monte Carlo.

Por último veamos un teorema que nos indica cómo debemos seleccionar las funciones de densidad p_i , con $i \in \{0, \dots, N-1\}$.

Teorema 4.4. *Elegimos la función de densidad inicial y la función de densidad de transición de la siguiente manera:*

$$p_0(\omega_0) = \frac{g(\omega_0)}{\int_{\mathbb{S}^2} g(\omega) d\mu(\omega)} \quad p_j(\omega_j) = \frac{R(x_j, \omega_j)}{\int_{\mathbb{S}^2} R(x_j, \omega) d\mu(\omega)} \quad \forall j \in \mathbb{N}_0 \quad (4.9)$$

Entonces la varianza del estimador (4.8) está acotada.

Demostración. Por comodidad notaremos:

$$\alpha_j = \frac{g(\omega_0)}{p_0(\omega_0)} W_j L^e(x_{j+1}, \omega_j), \quad \forall j \in \mathbb{N}_0$$

Tengamos en cuenta la siguiente desigualdad (aunque no será comprobada):

$$\left(\sum_{j=0}^{\infty} \alpha_j\right)^2 \leq \sum_{j=0}^{\infty} \frac{t^{-j}}{1-t} \alpha_j^2, \quad 0 < t < 1,$$

Entonces tenemos que:

$$\text{Var}(P_N[g]) \leq E[P_N[g]^2] = E\left[\left(\sum_{j=0}^N \alpha_j\right)^2\right] \leq E\left[\left(\sum_{j=0}^{\infty} \alpha_j\right)^2\right] \leq \sum_{j=0}^{\infty} \frac{t^{-j}}{1-t} E[\alpha_j^2]$$

Teniendo en cuenta como hemos definido las funciones de densidad:

$$\begin{aligned} E[\alpha_j^2] &= \int_{(\mathbb{S}^2)^j} \frac{g(\omega_0)^2}{p_0(\omega_0)} \frac{R(x_1, \omega_1)^2}{p_1(\omega_1)} \dots \frac{R(x_j, \omega_j)^2}{p_j(\omega_j)} L_e(x_{j+1}, -\omega_j)^2 d\mu(\omega_1) \dots d\mu(\omega_j) \\ &= S_j \int_{(\mathbb{S}^2)^j} g(\omega_0) R(x_1, \omega_1) \dots R(x_j, \omega_j) L_e(x_{j+1}, -\omega_j)^2 d\mu(\omega_1) \dots d\mu(\omega_j) \\ &\leq S_j \|g\|_\infty \|L_e\|^2(l^j) \leq 4\pi \|g\|_\infty \|L_e\|^2(l^j) \end{aligned}$$

Con l la constante de lipschitz de K , y S_j cumpliendo que:

$$S_j = \int_{\mathbb{S}^2} g(\omega) d\mu(\omega) \prod_{i=1}^j \int_{\mathbb{S}^2} R(x_i, \omega) d\mu(\omega) < \int_{\mathbb{S}^2} g(\omega) d\mu(\omega) \leq 4\pi \|g\|_\infty$$

Donde se ha usado que $\int_{\mathbb{S}^2} R(x_j, \omega) d\mu(\omega) < 1$ para todo j , lo cuál se cumple por la asunción de conservación de la energía. Por tanto:

$$\text{Var}(P_N[g]) \leq \sum_{j=0}^{\infty} \frac{t^{-j}}{1-t} 4\pi \|g\|_\infty \|L_e\|^2(l^j)$$

Y tomando $1 > t > l$:

$$\text{Var}(P_N[g]) \leq \frac{l}{(1-t)(t-l)} 4\pi \|g\|_\infty \|L_e\|^2$$

■

Como vemos, para que el estimador tenga varianza acotada, las direcciones deben ser muestreadas según la función BSDF multiplicada por el coseno del ángulo que forma la dirección con la normal a la superficie. Además si tomamos $g(\omega) = \frac{\mathbb{1}_{S(P)}(\omega)}{|P|}$, $\forall \omega \in \mathbb{S}^2$, el estimador (4.8) aproxima la radiancia incidente media en el píxel P , y la dirección inicial debe ser generada según la distribución uniforme en el conjunto de direcciones que pasan por el pixel. En cada punto x_j , $j \in \{1, \dots, N+1\}$, en lugar de únicamente evaluar su

emisividad en la dirección $-\omega_{j-1}$, es habitual que si el punto no es emisivo se tome como radiancia emitida la radiancia saliente en la dirección $-\omega_{j-1}$ desde el punto x_j , $L(x_j, -\omega_{j-1})$, con lo que tenemos en cuenta la iluminación indirecta en el camino trazado. Por tanto en cada punto del camino tendremos que estimar la radiancia $L(x_j, -\omega_{j-1})$, lo cuál se hace muestreando una fuente de luz aleatoria. Esto se hace para evitar que se tracen excesivos caminos cuya estimación sea nula, ya que será muy habitual trazar caminos donde no se interseque ninguna superficie emisiva. De esta manera reduciremos la varianza del estimador. Este algoritmo para estimar la ecuación de renderización recibe el nombre de *path-tracing*.

Parte II

Informática

Capítulo 5

Métodos de muestreo directo de fuentes de luz

Una característica común a todos los algoritmos de aproximación de la ecuación de renderización es que eventualmente tendrán que muestrear las direcciones que apuntan hacia las fuentes de luz de la escena con el objetivo de utilizar estas muestras para evaluar el correspondiente estimador de Monte Carlo. En efecto, como se ha visto en el capítulo anterior, tanto en el algoritmo de iluminación directa, que muestrea todas las fuentes de luz de la escena en cada punto sombreado, como en el algoritmo de path-tracing, que muestrea una fuente de luz en cada punto de los caminos trazados, se muestrean fuentes de luz.

Supongamos que queremos aproximar la integral de la función $g_P : \mathbb{S}^2 \rightarrow \mathbb{R}_0^+$ sobre el ángulo sólido subtendido por una fuente de luz desde un punto P de la escena. Usualmente tendremos:

$$g_P(\omega) = L_e(t(P, \omega), -\omega) f(P, \omega_o, \omega) \cos(n(P), \omega),$$

para alguna dirección ω_o , donde t es la función que asigna a cada pareja (x, ω) el primer punto visible desde x en dirección ω y $n(P)$ es la normal a la superficie en P . Para estimar dicha integral utilizando un estimador Monte Carlo necesitamos ser capaces de tomar muestras respecto de alguna distribución de probabilidad definida sobre el ángulo sólido subtendido por la fuente de luz.

Como ya se ha mencionado anteriormente, la forma más sencilla de muestrear una fuente de luz de área es generar una muestra con distribución uniforme en su superficie y obtener la función de densidad de la muestra respecto al ángulo sólido, teniendo en cuenta la transformación 3.7. Sin embargo, cuando la fuente de luz está muy cerca del punto P , este método de

muestreo puede llevar a una varianza muy alta. En efecto, vemos que:

$$E\left[\frac{g_P(\omega)^2}{f_\omega(\omega)^2}\right] = \int \frac{g_P(\omega)^2}{f_\omega(\omega)} d\mu(\omega) = \int \frac{g_P(\omega)^2 A_F \cos(n(t(P, \omega)), \omega)}{\|t(P, \omega) - P\|^2} d\mu(\omega),$$

donde A_F es el área de la fuente de luz y f_ω es la función de densidad asociada al método de muestreo uniforme respecto al área. Recordando la discusión en 3.4.1, cuando $\|t(P, \omega) - P\|$ sea cercano a 0, o lo que es lo mismo, cuando la fuente de luz esté muy cerca de P , $E\left[\frac{g_P(\omega)^2}{f_\omega(\omega)^2}\right]$ alcanzará valores altos, y por tanto la varianza del estimador Monte Carlo de la función $\frac{g_P}{f_\omega}$ será muy alta. Que la varianza sea alta implicará que necesitaremos más muestras por píxel para acercarnos al valor real, y esto a su vez incrementará el tiempo de ejecución necesario para generar la imagen final.

Durante los últimos años se han investigado nuevos métodos de muestreo de fuentes de luz en los que las muestras generadas siguen distribuciones de probabilidad con funciones de densidad diferentes a la que se obtiene con el muestreo uniforme respecto al área. Por un lado, se han descrito diversos métodos tales que las muestras ω generadas presentan una función de densidad uniforme respecto al ángulo sólido subtendido por la fuente de luz muestreada, es decir:

$$f_\omega(\omega) = \frac{1}{\mu(\pi_P(F))}, \forall \omega \in \pi_P(F)$$

con F la fuente de luz, π_P la proyección sobre la esfera unidad centrada en P y μ la medida del ángulo sólido. Supongamos que hemos generado R muestras $\{\omega^{(r)}\}_{r=1}^R$, con función de densidad uniforme respecto al ángulo sólido. El estimador Monte Carlo que estima la radiancia saliente en una dirección ω_s y en el punto P a causa de la radiancia incidente procedente de la fuente de luz F viene dado por:

$$\begin{aligned} & \frac{1}{R} \sum_{i=1}^R \frac{L_e(t(P, \omega^{(i)}), -\omega^{(i)}) f(P, \omega_s, \omega^{(i)}) \cos(n(P), \omega^{(i)})}{f_\omega(\omega^{(i)})} = \\ & \frac{1}{R} \sum_{i=1}^R L_e(t(P, \omega^{(i)}), -\omega^{(i)}) f(P, \omega_s, \omega^{(i)}) \cos(n(P), \omega^{(i)}) \mu(\pi_P(F)) \end{aligned}$$

Una ventaja de las muestras generadas de esta forma es que el estimador Monte Carlo asociado, por lo general, en puntos desde los que la fuente de luz sea completamente visible, presenta menos varianza que en el caso del muestreo uniforme respecto al área. Sin embargo, este método tiene la desventaja de que permanece el término coseno multiplicando, por lo que las muestras que estén cerca de formar un ángulo recto con la normal $n(P)$ apenas tendrán aportación a la estimación final. Por tanto, si hay un excesivo número de muestras cerca del horizonte, la estimación puede no ser acertada.

Aunque no hemos tratado en este trabajo el comportamiento de la luz en medios distintos del vacío, describiremos brevemente la ecuación que modela dicho comportamiento para poner de manifiesto otra ventaja de este método de muestreo. Cuando el punto P pertenece a un medio distinto del vacío, se cumple que la radiancia saliente en una dirección ω_s y en el punto P a causa de la radiancia incidente procedente de la fuente de luz F viene dado por:

$$L_s(P, \omega_s) = \int_{\pi_P(F)} L_e((t(P, \omega), -\omega) f_p(P, \omega_s, \omega) T(P, t(P, \omega))) d\mu(\omega)$$

donde f_p es la función de fase y T es la transmitancia. Como vemos el término coseno desaparece en esta integral, por lo que la desventaja de este método en el vacío desaparece en medios isotrópicos que no son el vacío.

Otros métodos de muestreo de gran importancia son aquellos que generan muestras con distribución uniforme en el ángulo sólido proyectado. Tomamos la proyección ρ definida en 3.8, y consideraremos ρ^* que es la proyección ρ respecto a un sistema de referencia cuyo vector z es la normal a la superficie en P , $n(P)$. En vista de 3.9, concluimos que las muestras ω generadas según este tipo de procedimientos tendrán una función de densidad respecto al ángulo sólido asociada:

$$f_\omega(\omega) = \frac{\cos(\omega, n(P))}{A(\rho^*(\pi_P(F)))}, \forall \omega \in \pi_P(F)$$

donde $A(\rho^*(\pi_P(F)))$ es el área de $\rho^*(\pi_P(F))$. Supongamos que hemos generado R muestras $\{\omega^{(r)}\}_{r=1}^R$, siguiendo una distribución uniforme en $\rho^*(\pi_P(F))$, con F una fuente de luz, y que queremos estimar la radiancia saliente en una dirección ω_s a causa de la radiancia incidente procedente de F . El estimador Monte Carlo que aproxima dicho valor de radiancia es el siguiente:

$$\begin{aligned} \frac{1}{R} \sum_{i=1}^R & \frac{L_e(t(P, \omega^{(i)}), -\omega^{(i)}) f(P, \omega_s, \omega^{(i)}) \cos(n(P), \omega^{(i)})}{f_\omega(\omega^{(i)})} = \\ & \frac{1}{R} \sum_{i=1}^R \frac{L_e(t(P, \omega^{(i)}), -\omega^{(i)}) f(P, \omega_s, \omega^{(i)}) \cos(n(P), \omega^{(i)})}{\frac{\cos(\omega^{(i)}, n(P))}{A(\rho^*(\pi_P(F)))}} = \\ & \frac{1}{R} \sum_{i=1}^R L_e(t(P, \omega^{(i)}), -\omega^{(i)}) f(P, \omega_s, \omega^{(i)}) A(\rho^*(\pi_P(F))) \end{aligned}$$

Con lo que el estimador deja de tener el término coseno multiplicando. Por tanto, si las muestras siguen una distribución uniforme respecto al ángulo sólido proyectado, se reducirá la cantidad de muestras con poca aportación al estimador. Otra ventaja es que si el punto P que estamos

sombreado pertenece a una superficie difusa, es decir, si f es constante en P , y si la cantidad de radiancia que emite la fuente de luz es constante en todos puntos y en todas direcciones, entonces el estimador tendrá varianza nula incluso aunque tomemos una sola muestra, es decir, toma el valor real con una sola muestra. Basta ver la forma del estimador para comprobarlo.

Destacar que esta utilización de métodos de muestreo alternativos que reducen la varianza del estimador Monte Carlo es lo que se conoce como muestreo de importancia, tal y como se describió en el apartado 3.4.1.

En este capítulo presentaremos e implementaremos una serie de algoritmos de muestreo de fuentes de luz, alternativos al muestreo uniforme respecto al área, que pretenden reducir la varianza del estimador de Monte Carlo. En primer lugar, presentaremos un algoritmo que nos permitirá muestrear fuentes de luz rectangulares de manera uniforme respecto al ángulo sólido. A continuación, presentaremos dos algoritmos de muestreo de fuentes de luz con forma de disco, ambas generando muestras con distribución uniforme respecto al ángulo sólido. Por último, presentaremos dos algoritmos de muestreo de fuentes de luz esféricas tales que la función de densidad asociada a las muestras es uniforme respecto al ángulo sólido proyectado. Los algoritmos serán implementados en la versión 3 del renderizador fotorrealista de código libre pbrt, descrito en el libro [PJH16], que es accesible a través de [PJH]. De igual forma se mostrarán los resultados obtenidos. Todas las pruebas que aquí se muestren han sido realizadas en un ordenador con Ubuntu 20.04, procesador Intel Core i7 6500U y 8GB de RAM. El código implementado es accesible a través de [este repositorio de github](#).

5.1. Muestreo uniforme de rectángulos esféricos

Vamos a presentar un método de muestreo para fuentes de luz rectangulares, descrito en [UFK13]. Se denomina rectángulo esférico a la proyección de un rectángulo sobre una esfera unidad. Fijado un punto $P \in \mathbb{R}^3$ de la escena, consideramos π_P la proyección sobre la esfera unidad con centro P . Dada una fuente de luz rectangular F , nuestro objetivo es tomar muestras en el rectángulo esférico $\pi_P(F)$ utilizando una distribución uniforme respecto al ángulo sólido.

Con este objetivo describiremos una parametrización $M : [0, 1]^2 \rightarrow \pi_P(F)$ que preserva el área. Cuando decimos que preserva el área queremos decir que para cualquier subconjunto $U \subseteq [0, 1]^2$ se cumple que:

$$\frac{\mu(M(U))}{\mu(\pi_P(F))} = A(U)$$

donde $A(U)$ es el área del conjunto U . Por el teorema del cambio de

variable, tenemos que:

$$\mu(\pi_P(F))A(U) = \mu(M(U)) = \int_{M(U)} d\omega = \int_U |\det(J_M(x, y))| dx dy$$

y dado que la igualdad anterior se da para todo $U \subseteq [0, 1]^2$, deducimos que el mapa M preserva el área si y solo si $|\det(J_M(x, y))| = \mu(\pi_P(F))$ para todo $(x, y) \in [0, 1]^2$. En vista de esta propiedad, si V es un vector aleatorio con distribución uniforme en $[0, 1]^2$, entonces la función de densidad de $M(V)$ cumple que:

$$f_{M(V)}(M(v)) = \frac{f_V(v)}{|\det(J_M(v))|} \Rightarrow f_{M(V)}(M(v)) = \frac{1}{\mu(\pi_P(F))}, \forall v \in [0, 1]^2$$

Concluimos que basta aplicar el mapa M a una muestra siguiendo la distribución uniforme en $[0, 1]^2$ para obtener una muestra siguiendo la distribución uniforme en $\pi_P(F)$. Una importante característica común a todos los algoritmos aquí presentados es que están basados en parametrizaciones que parten de $[0, 1]^2$ y preservan el área, por lo que si las muestras iniciales en $[0, 1]^2$ están generadas utilizando estratificación, esta estratificación se trasladará a la región muestreada, reduciendo así la varianza de los estimadores. Pasamos por tanto a describir el mapa M .

5.1.1. Construcción de la parametrización M

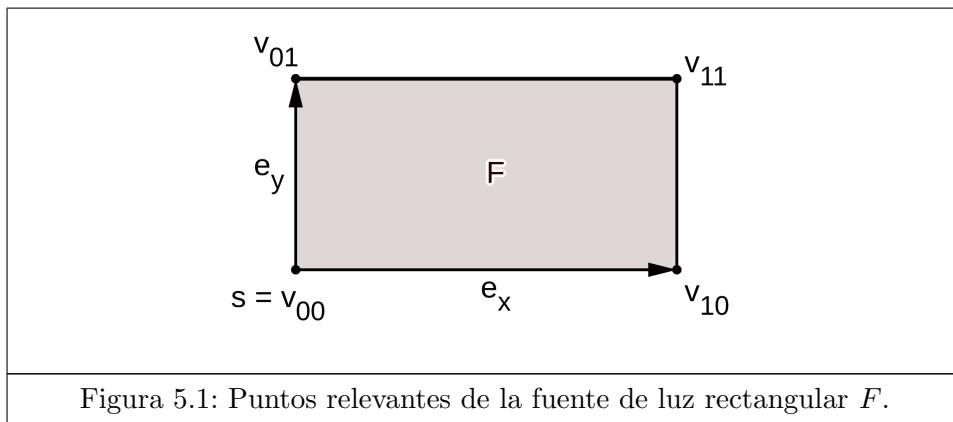


Figura 5.1: Puntos relevantes de la fuente de luz rectangular F .

En primer lugar, describiremos el sistema de referencia en el que vamos a trabajar. Fijemos $P \in \mathbb{R}^3$ el punto desde el cual queremos muestrear el ángulo sólido subtendido por el rectángulo F . Fijemos un vértice s del rectángulo F y consideremos los dos vectores e_x y e_y cuyo punto inicial es el vértice s y cuyos puntos finales son los vértices adyacentes a s (véase la

figura 5.1). Sea z el vector de longitud unidad perpendicular a e_x y e_y tal que $(s - P) \cdot z < 0$. Entonces, notando $x = \frac{e_x}{\|e_x\|}$, $y = \frac{e_y}{\|e_y\|}$, trabajaremos respecto al sistema de referencia $S = \{P; (x, y, z)\}$, por lo que todos los vectores de coordenadas serán respecto a S .

Consideramos ahora los siguientes valores:

$$x_0 = (s - P) \cdot x \quad y_0 = (s - P) \cdot y \quad z_0 = (s - P) \cdot z$$

$$x_1 = x_0 + \|e_x\| \quad y_1 = y_0 + \|e_y\|$$

Notaremos $v_{ij} := (x_i, y_j, z_0)$, $i = 0, 1$, $j = 0, 1$, a los cuatro vértices del rectángulo F . Consideramos la pirámide con ápice P y base F , y tomamos las normales a cada uno de los planos laterales de la pirámide:

$$\begin{aligned} n_0 &= \frac{v_{00} \times v_{10}}{\|v_{00} \times v_{10}\|} & n_1 &= \frac{v_{10} \times v_{11}}{\|v_{10} \times v_{11}\|} \\ n_2 &= \frac{v_{11} \times v_{01}}{\|v_{11} \times v_{01}\|} & n_3 &= \frac{v_{01} \times v_{00}}{\|v_{01} \times v_{00}\|} \end{aligned}$$

Notamos por $\gamma_i := \arccos(-n_i \cdot n_{i+1})$, $i = 0, 1, 2$, $\gamma_3 := \arccos(-n_3 \cdot n_0)$ al ángulo interior que forman los planos laterales de la pirámide. Se puede demostrar utilizando el teorema de Girard que el ángulo sólido subtendido por F desde P o, equivalentemente, el área de $\pi_P(F)$, mide:

$$\mu(\pi_P(F)) = \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 - 2\pi$$

Dado que F , que es la base de la pirámide descrita, tiene los lados paralelos a los ejes de coordenadas, se cumple que n_i tendrá al menos una coordenada nula para todo i . En efecto vemos que existen $\varphi_0, \dots, \varphi_3$ tales que:

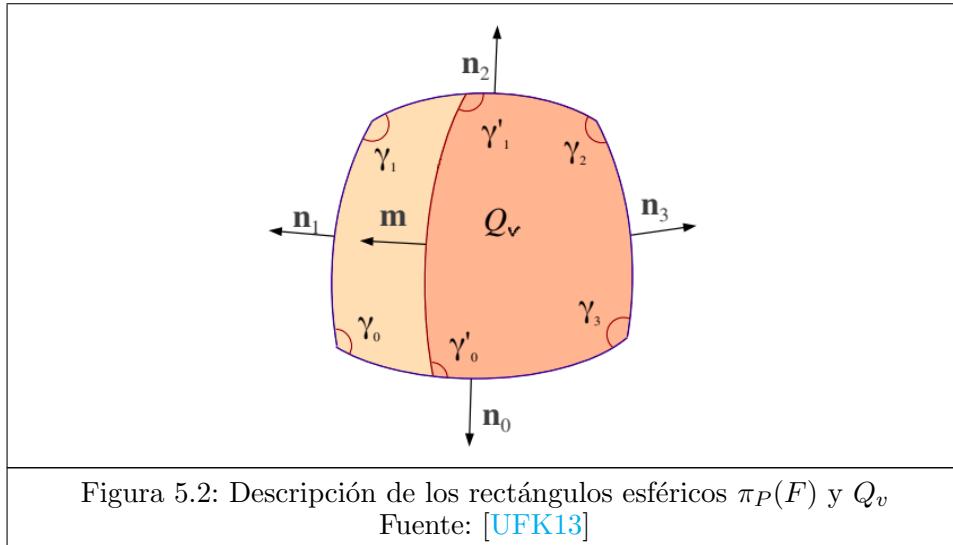
$$n_0 = \cos(\varphi_0)z + \sin(\varphi_0)y \quad n_1 = \cos(\varphi_1)z + \sin(\varphi_1)x$$

$$n_2 = \cos(\varphi_2)z + \sin(\varphi_2)y \quad n_3 = -\cos(\varphi_3)z - \sin(\varphi_3)x$$

Pasamos ya a construir la parametrización M que preserva el área. Tomamos $v = (v_1, v_2) \in [0, 1]^2$. Pretendemos encontrar (x_v, y_v, z_0) las coordenadas respecto al sistema de referencia S de un punto en F tal que su proyección sobre la esfera sea igual a $M(v)$. Para calcular x_v , establecemos un rectángulo esférico Q_v contenido en $\pi_P(F)$ tal que el área de Q_v cumpla que:

$$\mu(Q_v) = \mu(\pi_P(F))v_1$$

Para que M preserve el área tendremos que el rectángulo esférico Q_v es el resultante de proyectar sobre la esfera la parte del rectángulo F que cumple que su primera coordenada es menor o igual que x_v .



Como podemos ver en la figura 5.2, tres de los cuatro planos que delimitan el rectángulo $\pi_P(F)$ también delimitan Q_v . El cuarto plano delimitante es el plano que pasa por P y que es perpendicular a cierto vector m perteneciente al plano $y = 0$. Por tanto, para algún φ_v se tiene que:

$$m = \cos(\varphi_v)z + \sin(\varphi_v)x$$

El ángulo φ_v sólo dependerá de v_1 y varía desde φ_3 hasta φ_1 . El área de Q_v cumple que:

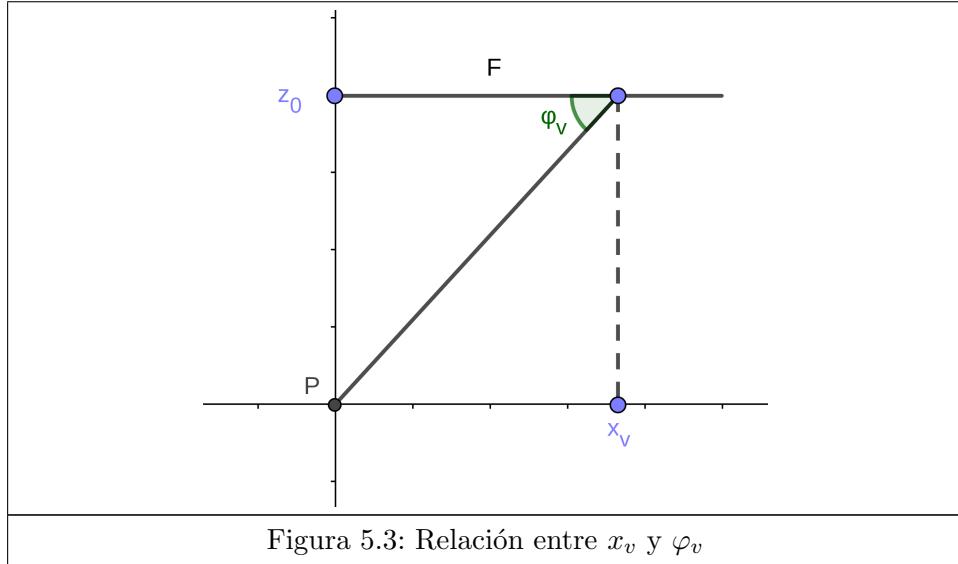
$$\begin{aligned} \mu(\pi_P(F))v_1 &= \mu(Q_v) = \arccos(-n_0 \cdot m) + \arccos(-m \cdot n_2) + \gamma_2 + \gamma_3 - 2\pi = \\ &= \arccos(-\cos(\varphi_0)\cos(\varphi_v)) + \arccos(-\cos(\varphi_2)\cos(\varphi_v)) + \gamma_2 + \gamma_3 - 2\pi \end{aligned}$$

Utilizando esta última igualdad y haciendo una serie de derivaciones (ver [UFK13]) obtenemos que:

$$\begin{aligned} \cos(\varphi_v) &= \frac{\operatorname{sign}(g(v_1))}{\sqrt{g(v_1)^2 + \cos(\varphi_0)}}, \\ g(v_1) &= \frac{\cos(\Phi(v_1))\cos(\varphi_0) - \cos(\varphi_2)}{\sin(\Phi(v_1))}, \\ \Phi(v_1) &= v_1\mu(\pi_P(F)) - \gamma_2 - \gamma_3 + 2\pi \end{aligned}$$

En vista de la figura 5.3, podemos obtener x_v de manera sencilla a partir de $\cos(\varphi_v)$:

$$x_v = -\cos(\varphi_v) \frac{z_0}{\sin(\varphi_v)} = -\cos(\varphi_v) \frac{z_0}{\sqrt{1 - \cos(\varphi_v)^2}}$$



Por último, calculemos y_v . Ya sabemos que la proyección de $M(v)$ sobre F estará en el segmento:

$$\{(x_v, ty_1 + (1 - t)y_0, z_0)/t \in [0, 1]\}$$

Para que M preserve el área, en vista de la igualdad 1.1, se tiene que cumplir que cambios iguales en v_2 generen cambios iguales en $\sin(\theta)$, con θ el ángulo entre los vectores y y $M(v)$. Para ello, podemos interpolar linealmente la segunda componente de $M(v)$, h_v , de la siguiente manera:

$$\begin{aligned} h_v &= h_0 + v_2(h_1 - h_0) \\ h_0 &= \frac{y_0}{\sqrt{x_v^2 + z_0^2 + y_0^2}} \\ h_1 &= \frac{y_1}{\sqrt{x_v^2 + z_0^2 + y_1^2}} \end{aligned}$$

Y concluimos que:

$$y_v = \frac{h_v \sqrt{x_v^2 + z_0^2}}{\sqrt{1 - h_v^2}}$$

Para finalizar, podemos normalizar el vector (x_v, y_v, z_0) , tomando:

$$(\hat{x}_v, \hat{h}_v, \hat{z}_0) = (x_v, y_v, z_0) \frac{1}{\sqrt{x_v^2 + y_v^2 + z_0^2}}$$

Y tenemos que la dirección $M(v)$ respecto al sistema de referencia usual cumple:

$$M(v) = \hat{x}_v \cdot x + \hat{h}_v \cdot y + \hat{z}_0 \cdot z$$

5.1.2. Resultados obtenidos

Este algoritmo ha sido implementado en [PJH]. Como vemos en la figura 5.4, con una fuente de luz pequeña ya se aprecia una importante mejora empleando el muestreo uniforme respecto al ángulo sólido en comparación al método de muestreo clásico que usa una función de densidad uniforme respecto al área de la fuente de luz. Destacar que todas las imágenes mostradas en este apartado han sido generadas aplicando muestreo de importancia múltiple, combinando así el muestreo de las fuentes de luz y de la función de distribución de dispersión, BSDF.

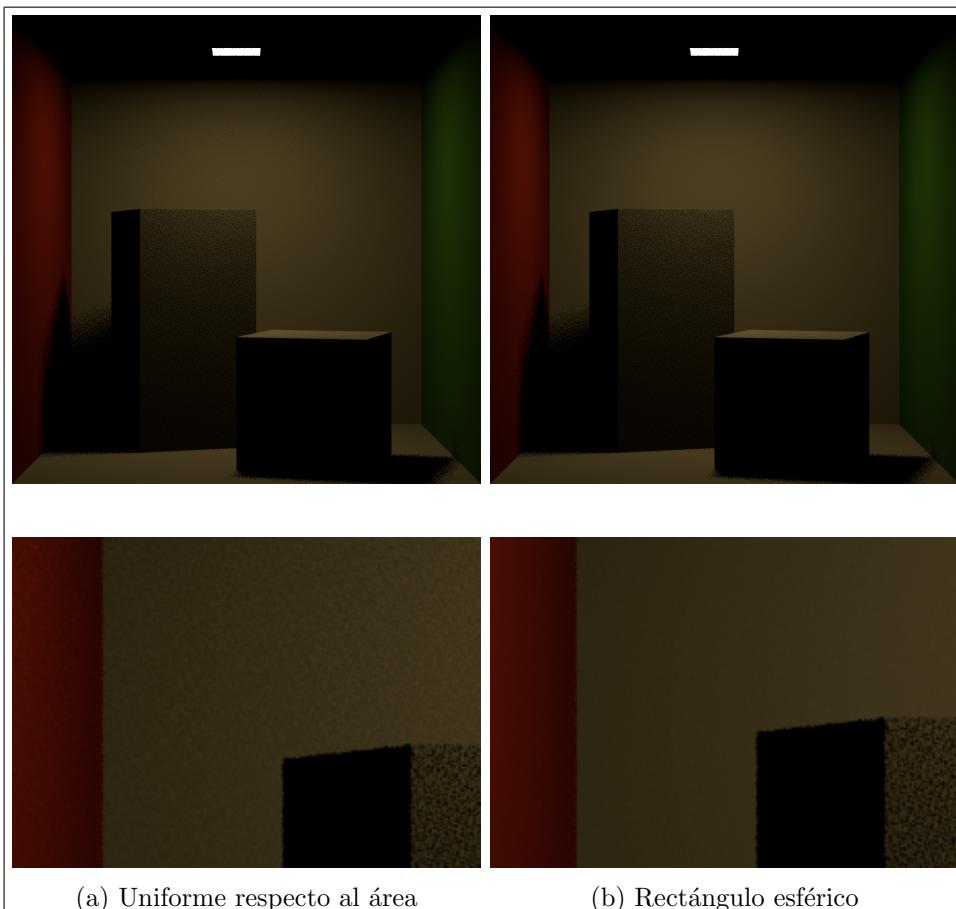


Figura 5.4: Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido. Imágenes generadas con una muestra por píxel y con el algoritmo de iluminación directa.

Si usamos una fuente de luz de mayor tamaño, la reducción de la varianza se aprecia más claramente (ver figura 5.5).

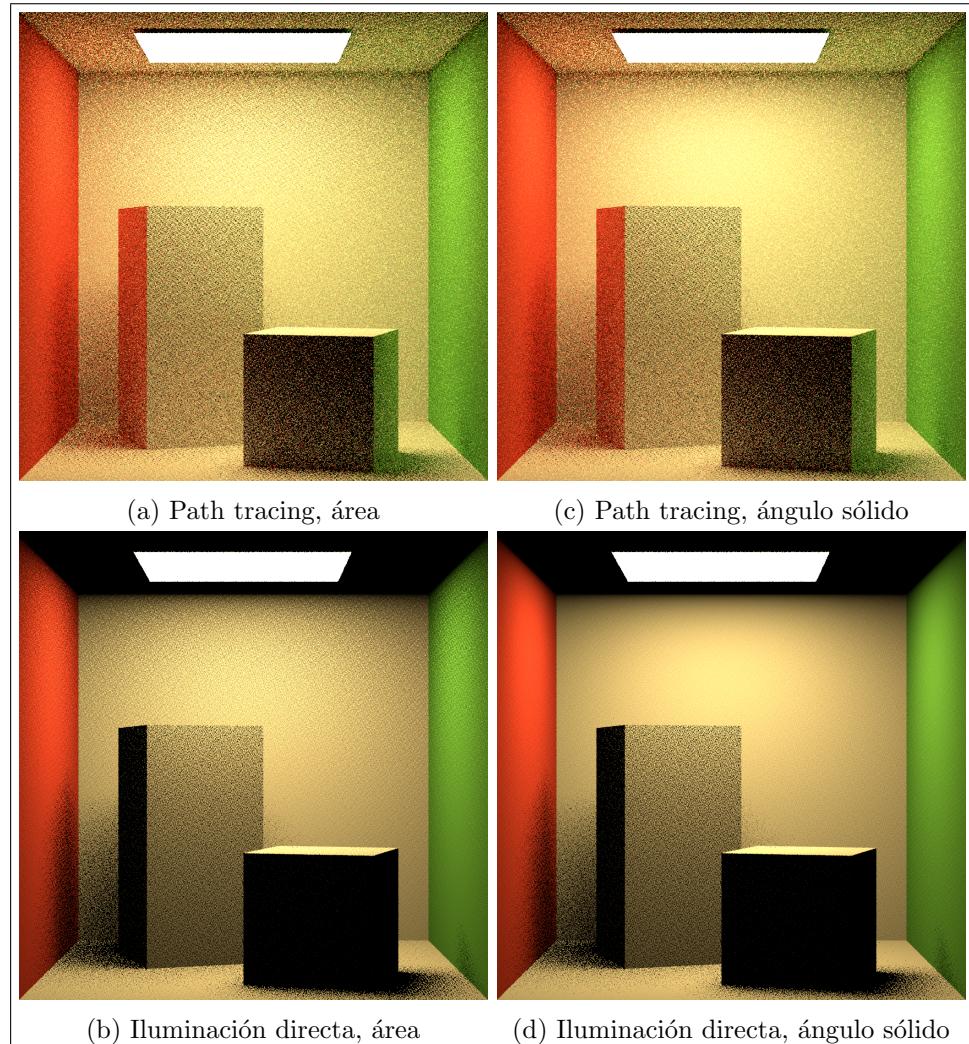


Figura 5.5: Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido. Imágenes generadas con una muestra por píxel.

Por otra parte, este algoritmo implica un aumento en el tiempo de ejecución, ya que la cantidad de operaciones y su complejidad es mucho mayor que en el caso del muestreo uniforme respecto al área. Sin embargo, como podemos observar en la figura 5.6, si generamos la imagen con dos muestras por píxel usando el muestreo uniforme respecto al área, el tiempo de ejecución es mayor que en el caso de una muestra por píxel utilizando el algoritmo descrito, pero en las superficies desde las que es completamente visible la fuente de luz sigue teniendo mejores resultados el muestreo del rectángulo esférico.

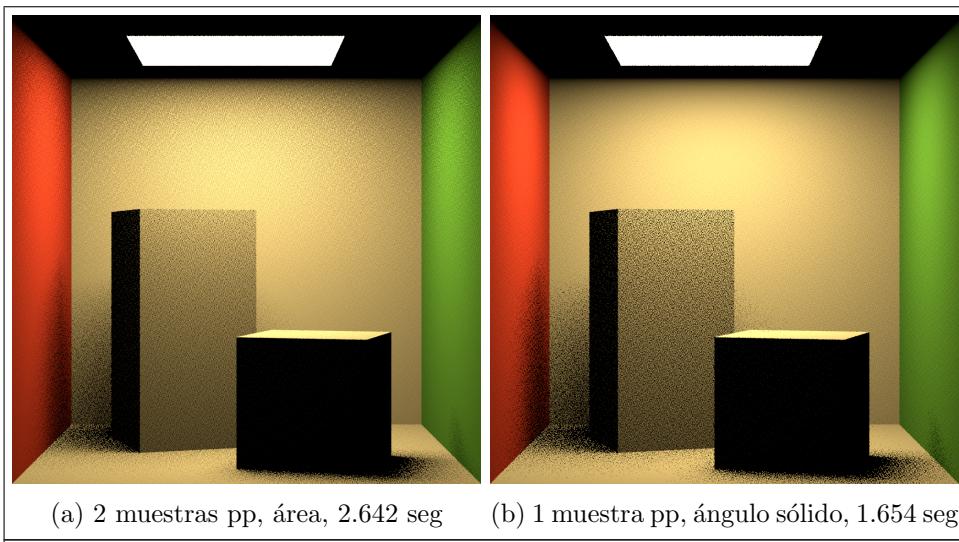


Figura 5.6: Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido.

Además podemos comprobar en la figura 5.7 que al utilizar este método en medios distintos al vacío la reducción de ruido es aún mayor.

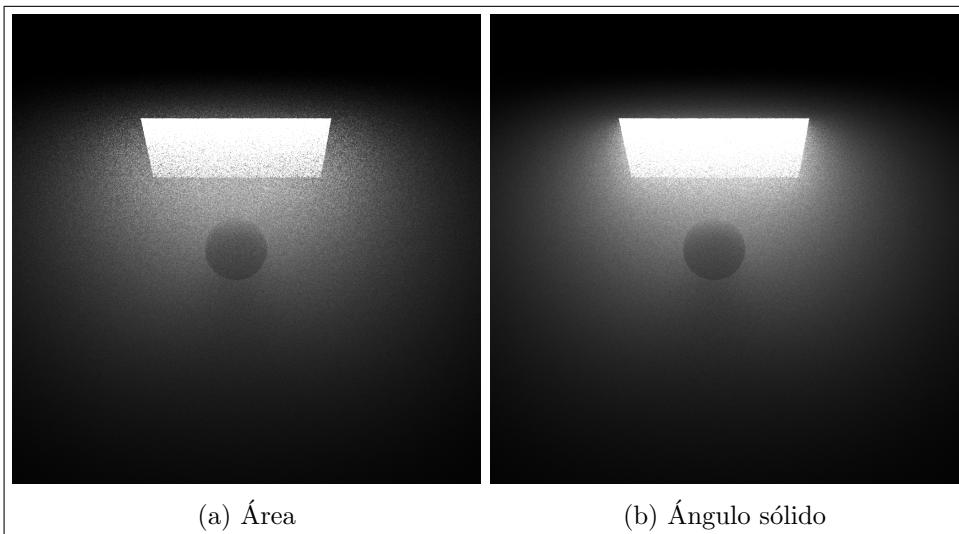


Figura 5.7: Comparativa entre muestreo uniforme respecto al área y muestreo uniforme respecto al ángulo sólido en medios distintos al vacío.

Por último, en la figura 5.8 vemos una comparativa de tiempo entre los dos algoritmos para generar la imagen de la caja de Cornell, tanto para iluminación directa como para path-tracing.

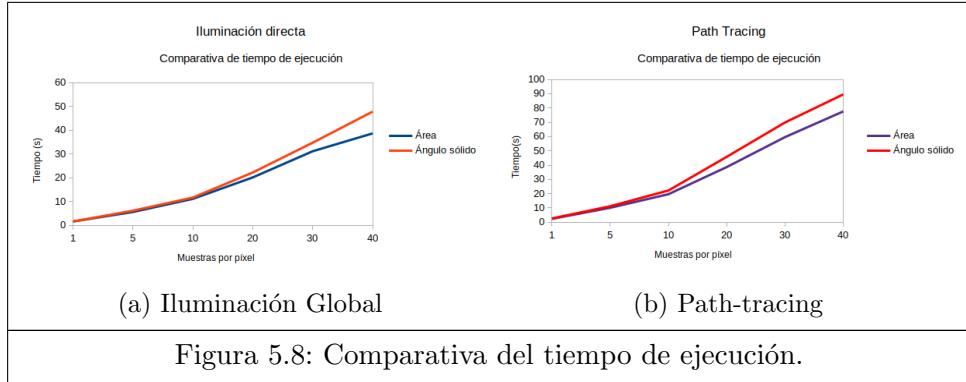


Figura 5.8: Comparativa del tiempo de ejecución.

Concluimos que, aunque este algoritmo incremente el tiempo de ejecución, reduce significativamente la varianza del estimador de la radiancia, especialmente en las superficies desde las que la fuente de luz es completamente visible y cuando la superficie de la fuente de luz es grande respecto a las dimensiones de la escena renderizada.

5.2. Muestreo uniforme de elipses esféricas

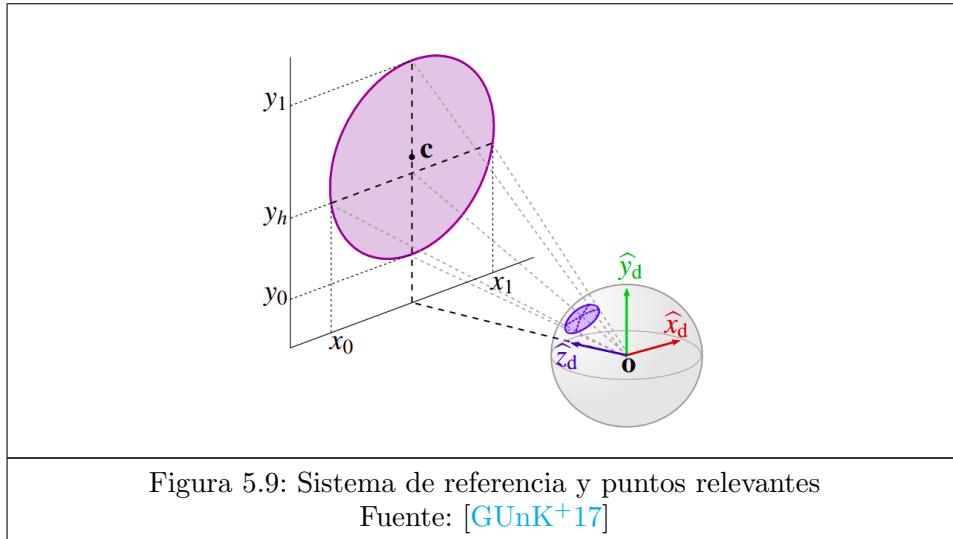
Procedemos de manera similar al caso de fuentes de luz rectangulares. El algoritmo descrito a continuación está recogido en [GUNK⁺17]. Una elipse esférica es la proyección sobre la esfera unidad de un disco. Consideramos una fuente de luz D cuya superficie es un disco, y fijamos un punto $o \in \mathbb{R}^3$ respecto al cual queremos muestrear las direcciones contenidas en la elipse esférica $\pi_o(D)$ siguiendo una distribución uniforme. En este caso construiremos dos parametrizaciones, $M_r : [0, 1]^2 \rightarrow \pi_o(D)$, $M_s : [0, 1]^2 \rightarrow \pi_o(D)$, que preservan el área.

5.2.1. Descripción del sistema de referencia y otros parámetros

Dada n_D la normal al disco D , c el centro del disco D , y consideremos los tres siguientes vectores:

$$z_d = -n_D$$

$$x_e = z_d \times \frac{c - P}{\|c - P\|}$$



$$y_d = x_e \times z_d$$

Tomamos el sistema de referencia $S_1 = \{o; (x_e, y_d, z_d)\}$ (ver figura 5.9) y todos los vectores de coordenadas aquí descritos serán respecto a S_1 . Consideramos los puntos y_0 e y_1 donde el disco toma el valor mínimo y máximo en la segunda coordenada. Si proyectamos y_0 e y_1 sobre la esfera unidad con centro o obtendremos dos puntos $(0, y'_0, z'_0)$, $(0, y'_1, z'_1)$. Entonces el centro de la elipse esférica $\pi_o(D)$ se puede calcular como:

$$z_e = (0, \frac{y'_0 + y'_1}{2}, \frac{z'_0 + z'_1}{2})$$

Notar que al reproyectar z_e en el disco obtenemos un punto $(0, y_h, z_h)$ que en general no tiene porque coincidir con el centro del disco c . La recta $y = y_h, z = z_h$ intersecciona con el disco en dos puntos, x_0 y x_1 . Proyectamos de nuevo el punto x_1 sobre la esfera unidad, y consideramos la primera coordenada del punto proyectado, que llamaremos x'_1 . Por último tomamos los siguientes valores:

$$\begin{aligned} a &= x'_1 & b &= \frac{1}{2} \sqrt{(y'_1 - y'_0)^2 + (z'_1 - z'_0)^2} \\ \alpha &= \sin^{-1}(a) & \beta &= \sin^{-1}(b) \\ a_t &= \tan(\alpha) & b_t &= \tan(\beta) \end{aligned}$$

donde a y b son los semi-ejes de la elipse esférica, α y β son sus semi-arcos. Tomando $y_e = z_e \times x_e$, de aquí en adelante cambiaremos el sistema de referencia y trabajaremos sobre $S = \{o; (x_e, y_e, z_e)\}$.

Antes de pasar a describir las dos parametrizaciones, destacar que estas están basadas en el siguiente resultado:

Proposición 5.1. *Sea \mathbb{S}^2 la esfera unidad de \mathbb{R}^3 , y sea C el cilindro de radio unidad cuyo eje está alineado con el eje X . Consideremos la transformación:*

$$T : \mathbb{S}^2 / \{(0, 0, 1), (0, 0, -1)\} \rightarrow C, \quad T(x, y, z) = \left(\frac{x}{\sqrt{x^2 + y^2}}, \frac{y}{\sqrt{x^2 + y^2}}, z \right)$$

Entonces dado $A \subseteq \mathbb{S}^2$ se cumple que:

$$\int_A dS = \int_{T(A)} dS$$

Demostración. Consideramos la parametrización $\Theta : [0, \pi] \times [0, 2\pi] \rightarrow \mathbb{S}^2$, con:

$$\Theta(\theta, \varphi) = (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta), \quad \forall (\theta, \varphi) \in [0, \pi] \times [0, 2\pi]$$

Consideramos también la parametrización $\Psi : [0, 2\pi] \times [-1, 1] \rightarrow C$, con:

$$\Psi(\varphi, z) = (\cos \varphi, \sin \varphi, z), \quad \forall (\varphi, z) \in [0, 2\pi] \times [-1, 1]$$

Por la igualdad 1.1, y haciendo un sencillo cálculo, vemos que:

$$\int_A dS = \int_{\Theta^{-1}(A)} \sin \theta d\theta d\varphi$$

$$\int_{T(A)} dS = \int_{\Psi^{-1}(T(A))} d\varphi dz$$

Consideramos la función:

$$R : [0, \pi] \times [0, 2\pi] \rightarrow [0, 2\pi] \times [-1, 1], R(\theta, \varphi) = (\varphi, \cos(\theta))$$

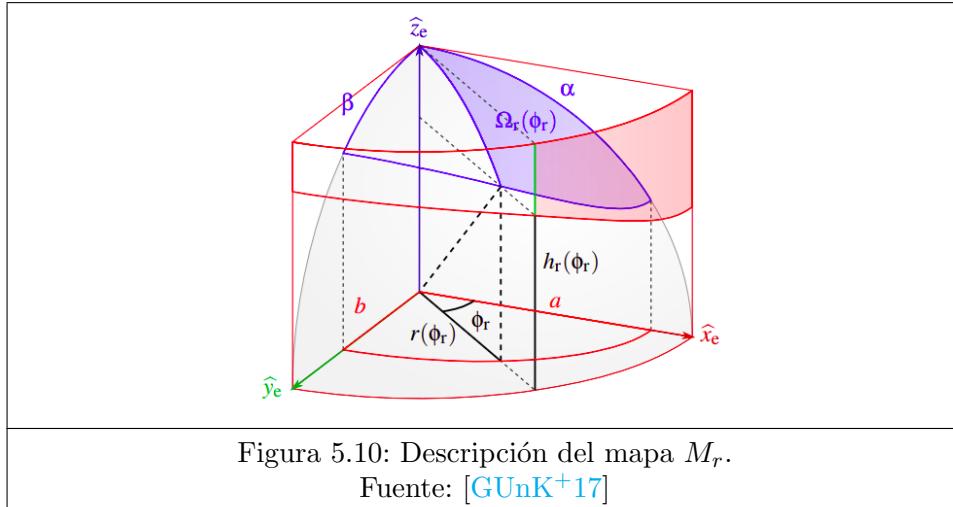
que cumple que $|det(J_R(\theta, \varphi))| = \sin(\theta)$. Por tanto, aplicando el teorema de cambio de variable, basta ver que $R(\Theta^{-1}(A)) = \Psi^{-1}(T(A))$ para obtener el resultado buscado, lo cuál se deduce fácilmente. ■

Este resultado nos dice que el ángulo sólido subtendido por una región de la esfera unidad es igual que el área de esa región proyectada sobre el cilindro de radio uno tal que su eje está alineado con un radio de la esfera.

5.2.2. Construcción de la parametrización M_r

Partimos de una muestra $v = (v_1, v_2) \in [0, 1]^2$, y vamos a describir como se obtiene la dirección $M_r(v)$ a partir de v . En este caso consideramos el

cilindro de radio unidad cuyo eje está alineado con z_e . Nos centraremos sólo en el caso de muestrear el primer cuadrante de la elipse, haciendo uso del hecho de que la elipse es radialmente simétrica. Para muestrear la elipse completa, basta dividir $[0, 1]$ en cuatro intervalos, un intervalo por cada cuadrante, y cambiar el sentido de x_e o y_e convenientemente en función del intervalo al que pertenece v_1 . Tras hacer esto hay que transformar v_1 de manera que el intervalo al que pertenece cubra todo $[0, 1]$.



Observando la figura 5.10, consideramos la función $\Omega_r : [0, \pi/2] \rightarrow \mathbb{R}_0^+$ que a cada ángulo ϕ_r respecto al eje x_e le asigna el ángulo sólido subtendido por la región azul de la elipse, o lo que es lo mismo, el área de la región roja del cilindro. Está claro que entonces $\mu(\pi_o(D)) = 4\Omega_r(\frac{\pi}{2})$. Es sencillo ver que:

$$\Omega_r(\phi_r) = \int_0^{\phi_r} [1 - h_r(x)] dx$$

donde h_r es la función que asigna a cada ángulo ϕ respecto al eje x_e la componente z_e del punto más bajo de la línea verde que delimita la región roja. Mediante una serie de derivaciones llegamos a:

$$\Omega_r(\phi_r) = \phi_r - \frac{b(1 - a^2)}{a\sqrt{1 - b^2}} \Pi(n; \varphi_r | m)$$

donde

$$\Pi(n; \varphi_r | m) = \int_0^{\varphi_r} \frac{dx}{(1 - n \sin^2(x)) \sqrt{1 - m \sin^2(x)}}$$

es la integral elíptica incompleta de Legendre de tercer tipo y donde

$$n = \frac{a^2 - b^2}{a^2(1 - b^2)} \quad m = \frac{a^2 - b^2}{1 - b^2} \quad \varphi_r = \arctan\left(\frac{a_t}{b_t} \tan(\phi_r)\right)$$

Dado que Ω_r no se puede invertir analíticamente, podemos utilizar el método de Newton-Raphson para encontrar ϕ_v cumpliendo que:

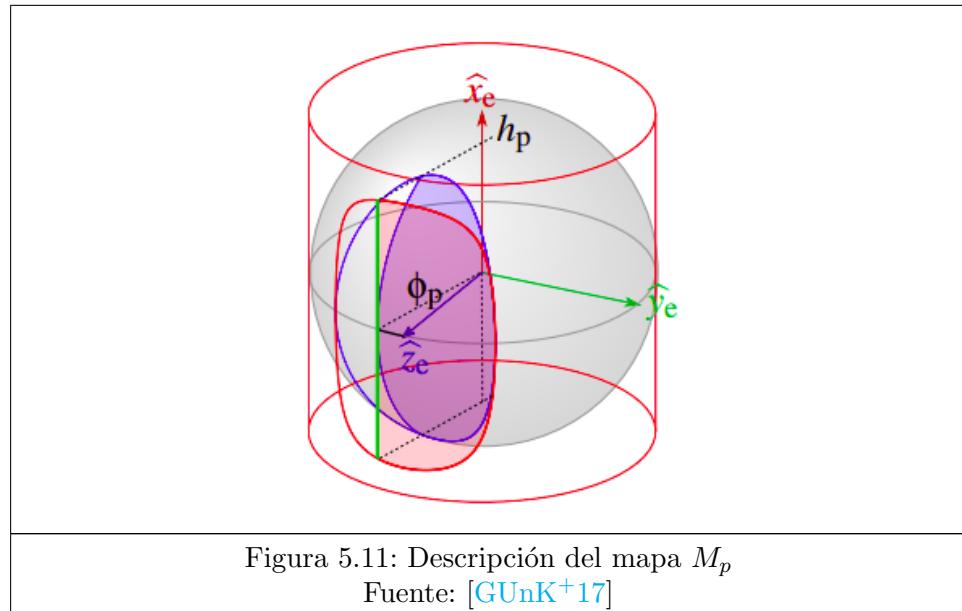
$$\Omega_r(\phi_v) - v_1 \Omega_r(\pi/2) = 0$$

Tomamos ahora $h = (1-v_2)h_r(\phi_v)+v_2$, y podemos definir $M_r(v)$ respecto al sistema de referencia usual como:

$$M_r(v) = \cos(\phi_v)\sqrt{1-h^2} \cdot x_e + \sin(\phi_v)\sqrt{1-h^2} \cdot y_e + h \cdot z_e$$

En efecto, M_r preserva el área.

5.2.3. Construcción de la parametrización M_p



En este caso, consideraremos el cilindro cuyo eje está alineado con el eje x_e . En este caso, considerando la imagen 5.11, el ángulo ϕ_p va desde $-\beta$ a β , y la integral del sector cuyo ángulo ϕ_p está comprendido entre $-\beta$ y ϕ_p (zona roja) se puede calcular como sigue:

$$\Omega_p(\phi_p) = \int_{-\beta}^{\phi_p} 2h_p(x)dx,$$

donde $(\phi_p, h_p(\phi_p)), (\phi_p, -h_p(\phi_p))$ son los extremos del segmento verde. Claramente el ángulo sólido subtendido por la elipse esférica cumple que:

$$\mu(\pi_o(D)) = \Omega_p(\beta)$$

Se puede derivar una expresión de la función h_p , obteniendo (ver [GUuK⁺¹⁷]):

$$h_p(\phi_p) = c_t \sqrt{\frac{1 - (p+1)\sin^2 \phi_p}{1 - (mp+1)\sin^2 \phi_p}},$$

donde:

$$p = \frac{1}{b_t^2} \quad m = \frac{a_t^2 - b_t^2}{a_t^2 + 1} \quad c_t = \frac{a_t}{\sqrt{1 + a_t^2}}.$$

Claramente por simetría la función h_p cumple que $h_p(x) = h_p(-x)$, $\forall x \in [0, \beta]$. Por tanto, tomando $\Omega_p^+(\phi_p) = \int_0^{\phi_p} 2h_p(x)dx$, vemos que:

$$\Omega_p(\phi_p) = \begin{cases} \Omega_p^+(\beta) + \Omega_p^+(\phi_p) & : \phi_p \geq 0 \\ \Omega_p^+(\beta) - \Omega_p^+(-\phi_p) & : \phi_p < 0 \end{cases}$$

Y sustituyendo la expresión de h_p en Ω^+ :

$$\Omega_p^+(\phi_p) = \frac{2c_t}{b_t} [(1-n)\Pi(n; \varphi_p|m) - F(\varphi_p|m)],$$

donde $n = -b_t^2$, $\varphi_p = \sin^{-1}(\frac{\tan \phi_p}{b_t})$, Π es la integral elíptica incompleta de Legendre de tercer tipo y F es la integral incompleta de Legendre de primer tipo, es decir:

$$F(\varphi_p|m) = \int_0^{\varphi_p} \frac{dx}{\sqrt{1 - m \sin(x)^2}}$$

Procedemos ahora igual que en el apartado anterior. Sea $v = (v_1, v_2) \in [0, 1]^2$, utilizamos el algoritmo de Newton-Raphson para encontrar un ϕ_v cumpliendo que:

$$\Omega_p(\phi_v) - v_1 \Omega_r(\beta) = 0$$

Tomamos $h = (-1 + v_2)h_p(\phi_v) + v_2h_p(\phi_v)$, y podemos definir $M_p(v)$ respecto al sistema de referencia usual como:

$$M_p(v) = h \cdot x_e + \sin(\phi_v) \sqrt{1 - h^2} \cdot y_e + \cos(\phi_v) \sqrt{1 - h^2} \cdot z_e$$

5.2.4. Resultados obtenidos

Los mapas recién presentados han sido implementados en el renderizador [PJH], habiéndose utilizado el código de la página [Bur] para aproximar el valor de las integrales elípticas. En la figura 5.13 vemos una comparativa

entre el muestreo clásico uniforme respecto al área de la fuente de luz y los dos algoritmos aquí presentados. En todas las imágenes aquí mostradas se ha aplicado muestreo de importancia múltiple.

Se puede apreciar que los dos mapas presentados en esta sección reducen la varianza, siendo el mapa paralelo el que mejores resultados ha dado para la escena mostrada. Como vemos, en la superficie de la esfera, que está más alejada de la fuente de luz, la varianza del muestreo uniforme respecto al área se ve reducida, aunque sigue dando mejores resultados el muestreo uniforme respecto al ángulo sólido. En las zonas en las que la fuente de luz es parcialmente visible, el método que peor parece comportarse es el mapa radial.

En la figura 5.12 podemos apreciar el comportamiento de estos algoritmos en medios distintos al vacío, apreciándose la reducción de ruido en los mapas radial y paralelo.

Por último, en la figura 5.14 vemos una comparativa del tiempo de ejecución de los diferentes métodos. Dado que en los dos métodos expuestos en esta sección se usan aproximaciones de las integrales elípticas, así como inversión numérica de funciones, el tiempo de ejecución es bastante superior al método de muestreo clásico, en especial en el caso del mapa paralelo, en el que se tienen que aproximar dos integrales elípticas cada vez que evaluemos la función Ω_p . Se podría disminuir el tiempo de ejecución de los dos mapas almacenando tablas de valores de las funciones que debemos invertir (Ω_p y Ω_r), y utilizando dichos valores para aproximar el valor buscado.

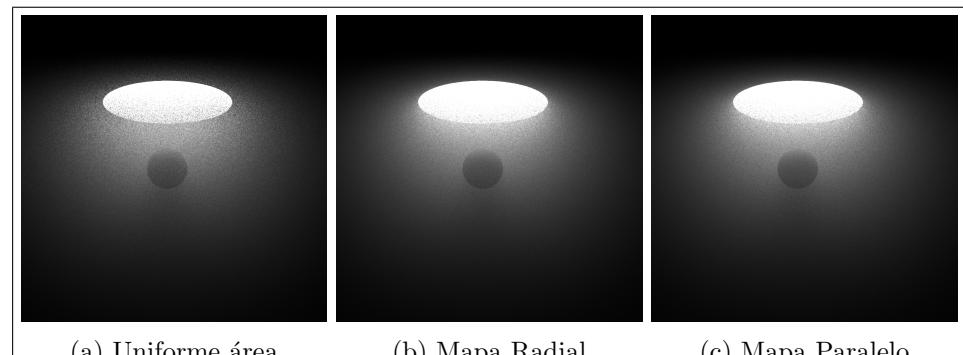


Figura 5.12: Comparativa entre muestreo uniforme respecto al área, mapa radial y paralelo en medios distintos al vacío.

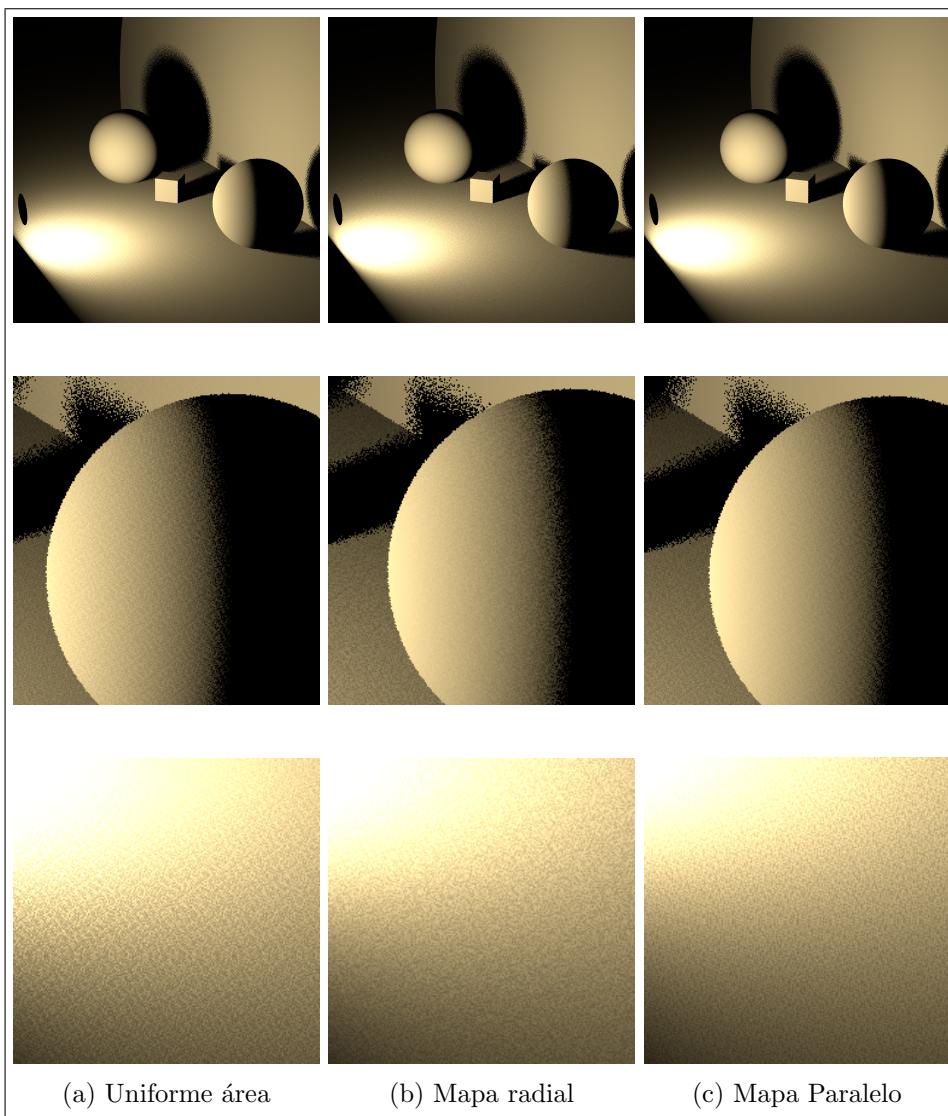
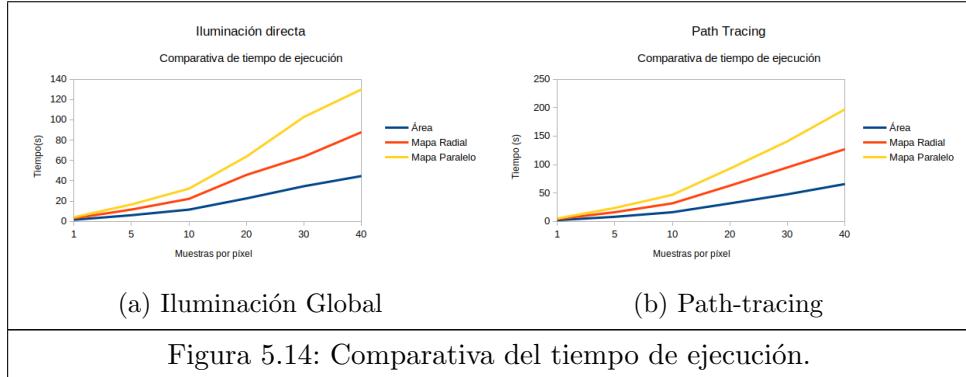


Figura 5.13: Comparativa entre muestreo uniforme respecto al área, mapa radial y paralelo. Imágenes generadas con una muestra por píxel y con el algoritmo de iluminación directa.



5.3. Muestreo de casquetes esféricos proyectados

En esta sección presentaremos un método de muestreo para fuentes de luz esféricas, descrito en [nG18]. Considerando la proyección ρ definida en 3.8, un casquete esférico proyectado es la imagen por ρ de un casquete esférico sobre la esfera unidad. Fijado un punto $o \in \mathbb{R}^3$ de la escena, consideramos π_o la proyección sobre la esfera unidad con centro o . Consideramos también ρ_o que es la proyección ρ respecto a un sistema de referencia cuyo vector z es la normal a la superficie en P , n_o . Dada una fuente de luz esférica E , nuestro objetivo es tomar muestras en el casquete esférico $\pi_o(E)$. Sin embargo, utilizaremos una distribución uniforme en el casquete esférico proyectado $\rho_o(\pi_o(E))$ para generar una muestra y dicha muestra se trasladará a $\pi_o(E)$ a través de ρ_o^{-1} . Pasamos ya a presentar los métodos de esta sección. Igual que en los casos anteriores, definiremos dos mapas $M_s : [0, 1]^2 \rightarrow \rho_o(\pi_o(E))$, $M_t : [0, 1]^2 \rightarrow \rho_o(\pi_o(E))$ que preserven el área.

5.3.1. Descripción del sistema de referencia y otros parámetros

Consideramos los siguientes tres vectores, con s el centro de la esfera E :

$$\hat{z} = n_o \quad y = \frac{(s - o) \times \hat{z}}{\|(s - o) \times \hat{z}\|} \quad x = \hat{z} \times y$$

En el caso en que el centro de la esfera esté en la misma dirección que \hat{z} , tomaremos y cualquier vector perpendicular a \hat{z} . Consideremos el siguiente punto, que es la proyección de s sobre la esfera unidad de centro o :

$$c = \frac{s - o}{\|s - o\|}$$

Y siendo r el radio de la esfera, consideramos los dos siguientes ángulos (ver figura 5.15, izquierda):

$$\alpha = \arcsin\left(\frac{r}{\|s - o\|}\right) \quad \hat{\beta} = \arcsin(\hat{z}(c - o))$$

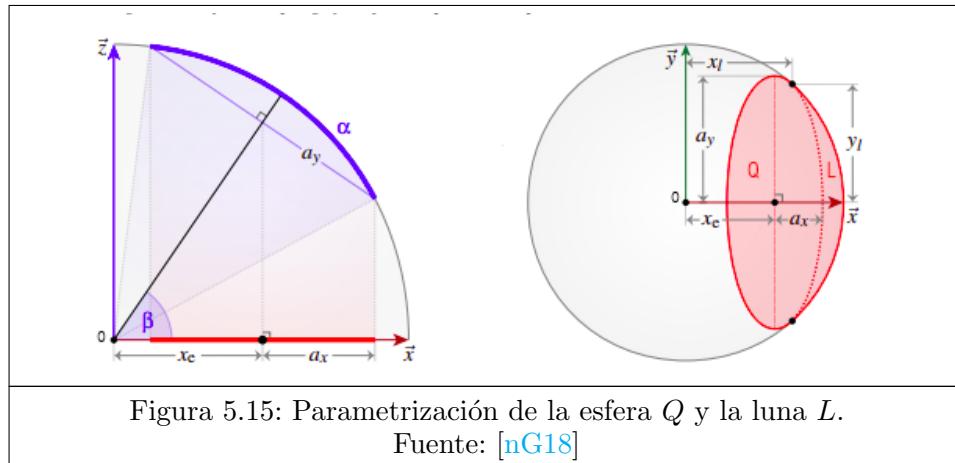
Vemos que $\hat{\beta} \in [-\pi/2, \pi/2]$. En el caso $\hat{\beta} < 0$, tomaremos:

$$z = -\hat{z} \quad \beta = -\hat{\beta}$$

Y en caso de que $\hat{\beta} \geq 0$, tomamos:

$$z = \hat{z} \quad \beta = \hat{\beta}$$

El sistema de referencia que usaremos será $\{o; (x, y, z)\}$.



Llegados a este punto tenemos que distinguir dos casos. Por un lado, si $0 \leq \alpha \leq \beta$, significará que el casquete esférico $\pi_o(E)$ está completamente contenido en el hemisferio superior de la esfera unidad cuyo polo norte es z . En este caso, $\rho_o(\pi_o(E))$ será una elipse. Por otro lado, si $0 \leq \beta < \alpha$, el casquete esférico tendrá una parte contenida en el hemisferio superior y otra en el hemisferio inferior. En este caso notaremos por $\rho_o(\pi_o(E))^+$ a la proyección de la parte de $\pi_o(E)$ contenida en el hemisferio superior, que tendrá forma de una elipse Q unida con una luna L (ver figura 5.15 derecha). Notaremos por $\rho_o(\pi_o(E))^-$ a la proyección de la parte de $\pi_o(E)$ contenida en el hemisferio inferior, que tendrá forma de luna L .

Vamos por tanto a calcular ciertos parámetros relacionados con la esfera Q y con la luna L (ver figura 5.15 derecha). Por un lado, la longitud de los

semiejes de la elipse, a_x , a_y , y la coordenada x del centro de la elipse, x_e , cumplen que:

$$a_y = \sin(\alpha) \quad a_x = \sin(\alpha) \sin(\beta) \quad x_e = \cos(\alpha) \cos(\beta)$$

Por otro lado, en el caso de que $0 \leq \beta < \alpha$ y la luna L esté definida, calcularemos los puntos de tangencia de la elipse con el disco, (x_l, y_l) , $(x_l, -y_l)$, haciendo uso de que cumplen las ecuaciones de la circunferencia y la elipse, obteniendo que:

$$x_l = \frac{\cos(\alpha)}{\cos(\beta)} \quad y_l = \sqrt{1 - x_l^2}$$

En el caso de que $0 \leq \beta < \alpha$, tenemos que decidir en qué hemisferio tomaremos muestras. Dado que partimos de una muestra uniforme en $[0, 1]^2$, podemos usar cualquiera de las componentes para determinar qué hemisferio muestrearemos, y la probabilidad de muestrear cada uno es proporcional al área relativa entre las dos proyecciones. Si el hemisferio muestreado es el inferior entonces cambiaremos el eje z por $-z$. Una vez determinado el hemisferio muestreado, tenemos que transformar la muestra utilizada para que vuelva a cubrir el intervalo $[0, 1]$.

5.3.2. Construcción de la parametrización M_s

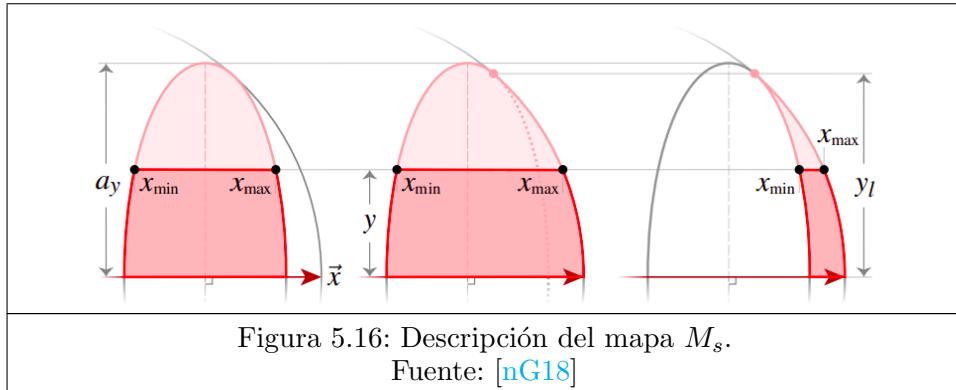
Por la simetría de los tres tipos de proyección, podemos muestrear sólo en la parte de la proyección con componente y positiva para luego ajustar la muestra en toda la proyección. En vista de la figura 5.16, definimos la siguiente función, que mide el área de una porción de la proyección de la esfera delimitada por un segmento paralelo al eje x :

$$A_p(s) = \int_0^s [x_{\max}(r) - x_{\min}(r)] dr$$

Distinguimos el caso 1, donde muestreamos solo una elipse Q , el caso 2, donde muestreamos una elipse y una luna, y el caso 3, donde solo muestreamos la luna L . Entonces definimos x_{\max} y x_{\min} como:

$$x_{\min}(s) = \begin{cases} x_e - a_x \sqrt{1 - \frac{s^2}{a_y^2}}, & \text{en el caso 3} \\ x_e + a_x \sqrt{1 - \frac{s^2}{a_y^2}}, & \text{en los casos 1 y 2} \end{cases}$$

$$x_{\max}(s) = \begin{cases} x_e + a_x \sqrt{1 - \frac{s^2}{a_y^2}}, & \text{si } \alpha \leq \beta \text{ o si } y_l < s \\ \sqrt{1 - s^2}, & \text{si } \beta < \alpha \text{ y } s \leq y_l \end{cases}$$



Entonces se cumple que, integrando la expresión anterior:

$$A_p(s) = \begin{cases} 2A_E(s), & \text{en el caso 1} \\ 2A_E(s) + A_C(s') - A_E(s'), & \text{en el caso 2} \\ A_C(s') - A_E(s'), & \text{en el caso 3} \end{cases}$$

donde $s' = \min\{y_l, s\}$, $I(u, w) = \frac{1}{2}(uw\sqrt{1-u^2} + \arcsin(u))$, y:

$$A_E(s) = a_x a_y I\left(\frac{s}{a_y}, 1\right)$$

$$A_C(s) = I(s, 1) - x_e s$$

Por último tomamos $v = (v_1, v_2) \in [0, 1]^2$, y definimos:

$$y_1 = \begin{cases} -y^* : A_p(y^*) - (1 - 2v_1)A_p(a_y) = 0, & \text{si } v_1 < 0.5 \\ y^* : A_p(y^*) - (2v_1 - 1)A_p(a_y) = 0, & \text{si } v_1 \geq 0.5 \end{cases}$$

$$x_1 = x_{\min}(y_1) + v_2(x_{\max}(y_1) - x_{\min}(y_1))$$

Para calcular el valor de y_1 , utilizamos el método de Newton-Raphson.
Por último la dirección buscada es:

$$M_s(v) = x_1 \cdot x + y_1 \cdot y + \sqrt{1 - x_1^2 - y_1^2} \cdot z.$$

M_s preserva el área por como la hemos construido.

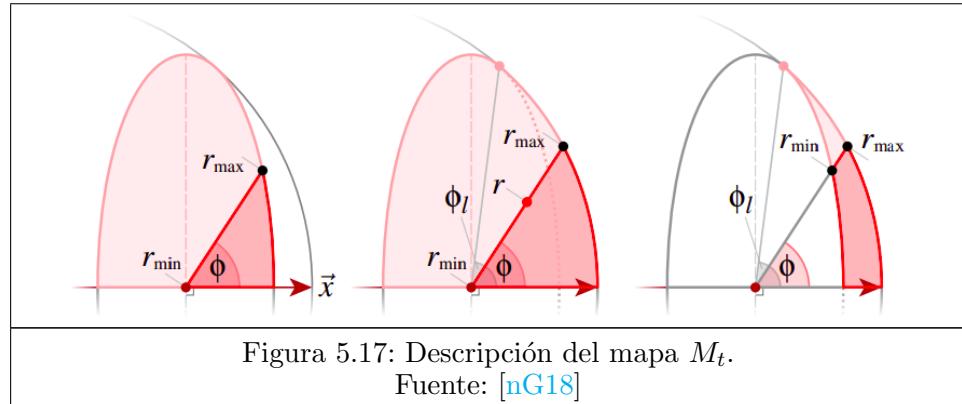
5.3.3. Construcción de la parametrización M_t

Fijamos $v = (v_1, v_2) \in [0, 1]^2$. Como en el caso anterior, por simetría, sólo consideramos la parte con y positiva del casquete esférico proyectado.

En vista de la figura 5.17, consideramos una región delimitada por una línea radial, tal que dicha región preserve el área respecto a la primera variable de v . Dado un ángulo $\phi \in [0, \pi]$, el área de la región asociada al radio cuyo ángulo con el eje x es ϕ viene dada por:

$$A_r(\phi) = \int_0^\phi \int_{r_{\min}(x)}^{r_{\max}(x)} r \, dr \, dx = \frac{1}{2} \int_0^\phi (r_{\max}^2(x) - r_{\min}^2(x)) \, dx$$

donde se ha usado que se trata una integral en coordenadas polares.



Al igual que en el mapa M_s , distinguimos tres casos, el caso 1, donde muestreamos solo una elipse Q , el caso 2, donde muestreamos una elipse y una luna, y el caso 3, donde solo muestreamos la luna L . Entonces definimos r_{\max} y r_{\min} como:

$$r_{\min}(\phi) = \begin{cases} 0, & \text{en los casos 1 y 2,} \\ \frac{a_x}{\sqrt{1-\cos^2 \beta \sin^2 \phi}}, & \text{en el caso 3} \end{cases}$$

$$r_{\max}(\phi) = \begin{cases} \frac{a_x}{\sqrt{1-\cos^2 \beta \sin^2 \phi}}, & \text{si } \alpha \leq \beta \text{ o si } \phi_l \leq \phi, \\ \sqrt{1-x_e^2 \sin^2 \phi} - x_e \cos \phi, & \text{si } \beta < \alpha \text{ y } \phi < \phi_l. \end{cases}$$

donde $\phi_l = \arctan\left(\frac{y_l}{x_l - x_e}\right)$. Integrando la expresión anterior obtenemos:

$$A_r(\phi) = \begin{cases} A'_E(\phi), & \text{caso 1,} \\ A'_E(\phi) - A'_E(\phi') + A'_C(\phi'), & \text{caso 2,} \\ A'_C(\phi') - A'_E(\phi'), & \text{caso 3} \end{cases}$$

donde $\phi' = \min(\phi, \phi_l)$ y:

$$A'_E(\phi) = \begin{cases} \frac{1}{2} a_x a_y \arctan\left(\frac{a_x}{a_y} \tan(\phi)\right), & \text{si } \phi \leq \pi/2 \\ \pi + \frac{1}{2} a_x a_y \arctan\left(\frac{a_x}{a_y} \tan(\phi)\right), & \text{si } \phi > \pi/2 \end{cases}$$

$$A'_C(\phi) = I(\sin(\phi), x_e^2) - I(x_e \sin(\phi), 1)$$

con I la función definida en el apartado anterior. Por último definimos:

$$\phi_1 = \begin{cases} -\phi^* : A_r(\phi^*) - (1 - 2v_1)A_p(\pi) = 0, & \text{si } v_1 < 0.5 \\ \phi^* : A_r(\phi^*) - (2v_1 - 1)A_p(\pi) = 0, & \text{si } v_1 \geq 0.5 \end{cases}$$

$$r_1 = \sqrt{r_{min}^2(\phi_1) + v_2(r_{max}^2(\phi_1) - r_{min}^2(\phi_1))}$$

Para calcular el valor de ϕ_1 , utilizamos el método de Newton-Raphson. r_1 está definida de esta forma para que el mapa preserve el área. Por último, tomando $x_1 = (x_e + r_1 \cos(\phi_1))$, $y_1 = (r_1 \sin(\phi_1))$, la dirección buscada es:

$$M_t(v) = x_1 \cdot x + y_1 \cdot y + \sqrt{1 - x_1^2 - y_1^2} \cdot z.$$

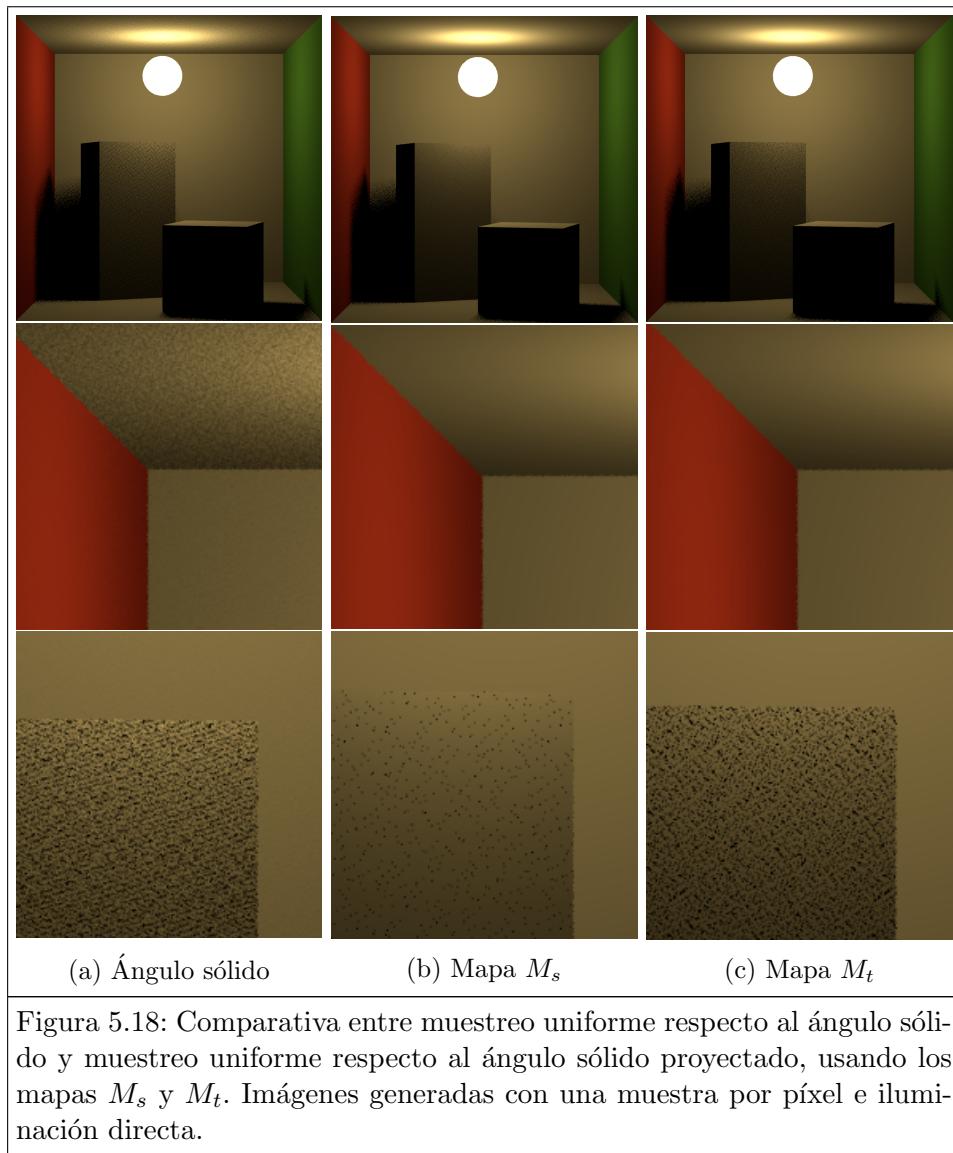
Sin embargo, en el caso 1 en que solo tenemos una elipse, el muestreo se puede optimizar tomando $x_1 = x_e + a_x \sqrt{v_1} \cos(2\pi v_2)$, $y_1 = a_y \sqrt{v_1} \sin(2\pi v_2)$, y definiendo el mapa M_t como:

$$M_t(v) = x_1 \cdot x + y_1 \cdot y + \sqrt{1 - x_1^2 - y_1^2} \cdot z.$$

5.3.4. Resultados obtenidos

En la figura 5.18 vemos una comparativa de los dos métodos presentados en este apartado y el método de muestreo clásico de fuentes de luz esféricas, que en este caso es el muestreo uniforme respecto al ángulo sólido. Se aprecia en estas imágenes la propiedad enunciada al inicio de este capítulo del muestreo uniforme respecto al ángulo sólido proyectado, con una sola muestra los puntos desde los que es completamente visible la fuente de luz tienen varianza nula. Además en los puntos en los que la fuente de luz no es completamente visible el mapa paralelo también presenta menos ruido que el muestreo uniforme respecto al ángulo sólido. El mapa radial no obtiene tan buenos resultados, probablemente debido a errores de aproximación debido al uso de números flotantes.

Por último, en la figura 5.19, podemos ver una gráfica comparativa de los tiempos de ejecución de los tres métodos.



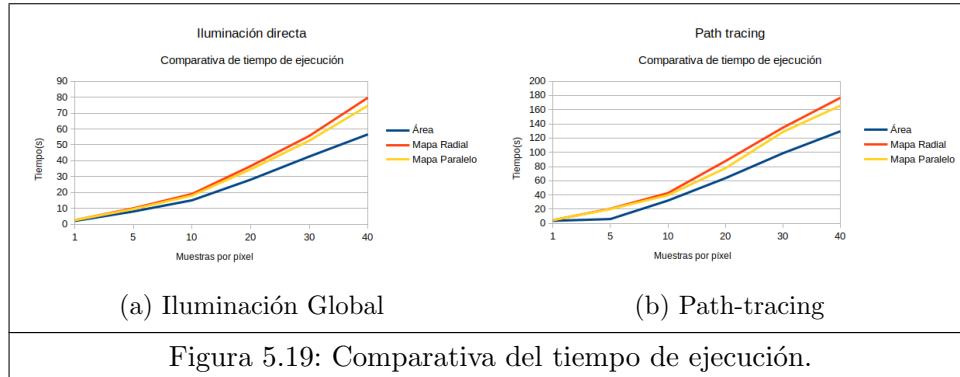


Figura 5.19: Comparativa del tiempo de ejecución.

5.4. Implementación en pbrt

Vamos a describir la implementación en pbrt. La implementación realizada para este trabajo es accesible desde [este enlace](#). Todos los caminos de archivos aquí indicados son relativos a la carpeta del repositorio `pbrt-v3/src`. En primer lugar, pbrt define una clase abstracta `Shape`, definida en los archivos `core/shape.cpp` y `core/shape.h`, que es una interfaz que implementarán todas las formas geométricas del sistema. Dentro de esta interfaz, son de vital importancia en lo referente a nuestro trabajo los siguientes métodos:

Listing 5.1: Métodos relevantes de la clase Shape

```
virtual Interaction Sample(const Interaction &ref, const
    Point2f &u, Float *pdf) const;

virtual Float Pdf(const Interaction &ref, const Vector3f &
    wi) const;
```

Estos dos métodos trabajan con funciones de densidad respecto al ángulo sólido subtendido por la forma geométrica desde el punto de referencia `ref`, que se le pasa como parámetro. La clase `Interaction` almacena todas las propiedades locales de un punto de la escena. El primero de los métodos toma un punto `ref` de la escena, que es el punto que está siendo sombreado, y `u` una muestra uniforme en $[0, 1]^2$, y devuelve una muestra en la superficie de la figura geométrica y el valor en el punto generado de la función de densidad respecto a la que se ha generado la muestra. El segundo toma un punto `ref` de la escena y una dirección `wi`, y devuelve el valor de la función de densidad evaluada en `wi`.

Las fuentes de luz de área, definidas por la clase `DiffuseAreaLight`, contienen un puntero a un objeto de tipo `Shape`, que define la geometría de la fuente de luz. Para generar muestras de una fuente de luz de área, la

clase `DiffuseAreaLight` tiene definido un método `Sample_Li`, que a su vez llama al correspondiente método `Sample` de su forma asociada y devuelve la muestra generada. Por otro lado, la función `Pdf` es necesaria para poder aplicar muestreo de importancia múltiple, ya que nos permite evaluar la función de densidad en direcciones generadas respecto a otras distribuciones.

En pbrt no hay definida una clase para instanciar rectángulos, y por tanto ha tenido que ser implementada de cero, y dicha implementación se encuentra en los archivos `project/rectangle.cpp` y `project/rectangle.h`. Por otro lado, pbrt sí que provee implementaciones de discos y de esferas, que se encuentran en `shapes/disk.cpp`, `shapes/disk.h`, `shapes/sphere.cpp` y `shapes/sphere.h`.

Por tanto, las implementaciones hechas en el sistema mayormente están contenidas en los métodos `Sample` y `Pdf` de las clases `Rectangle`, `Disk` y `Sphere`. En las tres clases se ha añadido un atributo `samplingMode` que determina qué tipo de muestreo será utilizado durante la ejecución del renderizador. A este atributo se le puede asignar un valor a través de los archivos de descripción de la escena, que están explicados en [este enlace](#). Por ejemplo, para instanciar un cuadrado de lado unidad que utilice el mapa presentado en este capítulo, basta escribir:

```
Shape "rectangle" "float height" [1] "float width"
      [1] "integer samplingMode" [2]
```

Para `samplingMode = 1`, el método de muestreo será el uniforme respecto al área en el caso de discos y rectángulos y el uniforme respecto al ángulo sólido en el caso de esferas. Para `samplingMode = 2`, usaremos el muestreo uniforme respecto al ángulo sólido en rectángulos, el mapa radial en discos y el mapa paralelo del ángulo proyectado en esferas. Para `samplingMode = 3`, usaremos el mapa paralelo en discos y el mapa radial en esferas.

En los archivos `project/newtonRaphson.cpp` y `project/newtonRaphson.h` se definen las funciones auxiliares para aplicar inversión numérica a las correspondientes funciones en los mapas paralelos y radiales. Los archivos `project/ellipticIntegral.cpp` y `project/ellipticIntegral.h` han sido descargados de [\[Bur\]](#), y contienen funciones que aproximan las integrales elípticas necesarias para la evaluación de las funciones Ω_p y Ω_r de los mapas descritos para discos. Además se ha definido una estructura para el almacenamiento de toda la información relevante de un rectángulo esférico:

```
struct SphRectangleData {
    Vector3f x, y, z;
    Float ex1, ey1;
    Point3f o, s;
    Float z0, z0sq;
    Float x0, y0, y0sq;
    Float x1, y1, y1sq;
    Float b0, b1, b0sq, k;
    Float solidAngle;
};
```

Así como también se ha definido una estructura para almacenar la información de las elipses esféricas:

```
struct SphEllipseData {
    Vector3f xd, yd, zd;
    Vector3f oc;
    Vector3f ye, ze;
    int signX, signY;
    double a, b, alpha, beta, at, bt;
    double solidAngle;
};
```

Además se han definido dos funciones auxiliares en el caso del disco que no estaban presentes por defecto en pbrt:

```
void ComputeSphEllipseData(SphEllipseData *data, Point3f o)
const;
```

```
Point3f ReprojectToDisk(const Vector3f &q, const
    SphEllipseData *data, const Point3f &o) const;
```

La primera inicializa los valores de **SphEllipseData**, y la segunda, una vez obtenemos la dirección muestreada, nos devuelve el punto en la superficie del disco asociada a dicha dirección, ya que como hemos mencionado anteriormente la función **Sample** devuelve el punto muestreado en la fuente de luz en lugar de la dirección.

Por último, los archivos de descripción de las escenas renderizadas están contenidos en la carpeta del repositorio **pbrt-v3/scenes**.

Capítulo 6

Conclusiones y trabajo futuro

En este trabajo hemos estudiado la base matemática de los métodos de Monte Carlo más utilizados en renderización, así como también hemos descrito formalmente los algoritmos más básicos de aproximación de la ecuación de renderización. Esto nos ha dado las herramientas para entender la importancia que tienen los métodos de muestreo de fuentes de luz en un renderizador fotorrealista, ya que utilizar muestras respecto de una distribución de probabilidad conveniente permite reducir significativamente la varianza de los estimadores, consiguiendo así una mejor aproximación con menos muestras.

Sin embargo, queda mucho por hacer. Los métodos de muestreo de fuentes de luz presentados, si bien reducen la varianza del estimador Monte Carlo que aproxima la ecuación de renderización, tienen sus inconvenientes. Por un lado, muchos de ellos incrementan en gran medida el tiempo de ejecución necesario para generar una imagen, por lo que se podrían analizar posibilidades de mejora en cuanto a eficiencia o reducción de la varianza. Además, los métodos que consisten en el muestreo uniforme del ángulo sólido subtendido por la fuente de luz, aún sufren el inconveniente de que su estimador asociado tiene un término coseno multiplicando, lo cuál hace que las muestras cercanas al horizonte no tengan apenas aportación. Una vía de mejora sería intentar buscar parametrizaciones del ángulo sólido proyectado asociado a fuentes de luz rectangulares o con forma de disco. También puede ser interesante investigar la parametrización del ángulo sólido o ángulo sólido proyectado subtendido por otras formas geométricas.

Por otra parte, se puede estudiar el uso y las características de otros renderizadores de producción, así como también queda pendiente el estudio de otras formas de aproximación de la ecuación de renderización, como los

métodos bidireccionales o el método de transporte de luz de Metrópolis, que no han sido tratados en este trabajo. Un aspecto también muy relevante que hemos pasado por alto es el comportamiento de la luz en medios distintos del vacío.

Bibliografía

- [Arv95] James Arvo. Stratified sampling of spherical triangles. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '95, page 437–438, New York, NY, USA, 1995. Association for Computing Machinery.
- [Bur] John Burkardt. Elliptic integrals. https://people.sc.fsu.edu/~jburkardt/cpp_src/elliptic_integral/elliptic_integral.html. [Online].
- [CH89] Richard Courant and David Hilbert. *Methods of Mathematical Physics*, volume 1. Wiley, New York, 1989.
- [DGP05] Ivan Dimov, Todor Gurov, and Anton Penzov. A monte carlo approach for the cook-torrance model. volume 3401, pages 257–265, 02 2005.
- [Fre] Integral equations. <http://www.et.byu.edu/~vps/ET502WWW/NOTES/CH7m.pdf>. [Online].
- [GT13] Carl Graham and Denis Talay. *Stochastic Simulation and Monte Carlo Methods. Mathematical Foundations of Stochastic Simulation.*, volume 68. 01 2013.
- [GUñK⁺17] Ibón Guillén, Carlos Ureña, Alan King, Marcos Fajardo, Ilian Georgiev, Jorge López-Moreno, and Adrian Jarabo. Area-preserving parameterizations for spherical ellipses. *Computer Graphics Forum*, 36(4):179–187, July 2017.
- [Kaj86] James T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150, August 1986.
- [LaL13] Scott M. LaLonde. The martingale stopping theorem. 2013.
- [MA14] Marcela Martins and Fernando Acero. Fredholm, ecuaciones integrales y álgebra lineal. 2014.

- [Mac98] D. J. C. Mackay. *Introduction to Monte Carlo Methods*, pages 175–204. Springer Netherlands, Dordrecht, 1998.
- [MB96] Antonio Martinon and Teresa Bermudez. On neumann operators. *Extracta mathematicae, ISSN 0213-8743, Vol. 7, N° 2-3, 1992, pags. 107-109*, 200, 06 1996.
- [MfRiMSS02] N.N. Madras, Fields Institute for Research in Mathematical Sciences, and American Mathematical Society. *Lectures on Monte Carlo Methods*. Fields Institute for Research in Mathematical Sciences Toronto: Fields Institute monographs. American Mathematical Society, 2002.
- [M.R44] Rev. James Booth LL.D. M.R.I.A. Iv. on the rectification and quadrature of the spherical ellipse. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 25(163):18–38, 1844.
- [nG18] Carlos Ureña and Iliyan Georgiev. Stratified sampling of projected spherical caps. *Computer Graphics Forum (Proceedings of EGSR)*, 37(4), 2018.
- [PJH] Matt Pharr, Wenzel Jakob, and Greg Humphreys. pbrt, version 3. <https://github.com/mmp/pbrt-v3>. [Online].
- [PJH16] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2016.
- [RR04] G. Roberts and J. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.
- [Shi20a] Peter Shirley. Ray tracing in one weekend, December 2020. <https://raytracing.github.io/books/RayTracingInOneWeekend.html>.
- [Shi20b] Peter Shirley. Ray tracing: The next week, December 2020. <https://raytracing.github.io/books/RayTracingTheNextWeek.html>.
- [Shi20c] Peter Shirley. Ray tracing: The rest of your life, December 2020. <https://raytracing.github.io/books/RayTracingTheRestOfYourLife.html>.
- [Tin] Samy Tindel. Conditional expectation. <https://www.math.purdue.edu/~stindel/teaching/ma539/cdt-expectation-2.pdf>. [Online].
- [UFK13] Carlos Ureña, Marcos Fajardo, and Alan King. An area-preserving parametrization for spherical rectangles. *Computer Graphics Forum*, 32(4):59–66, 2013.

- [Vea98] Eric Veach. *Robust Monte Carlo Methods for Light Transport Simulation*. PhD thesis, Stanford University, Stanford, CA, USA, 1998. AAI9837162.
- [Whi80] Turner Whitted. An improved illumination model for shaded display. *Commun. ACM*, 23(6):343–349, June 1980.
- [Wil91] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.