ANALYTIC METHODS FOR SIMULATED LIGHT TRANSPORT

James Richard Arvo

Yale University

1995

This thesis presents new mathematical and computational tools for the simulation of light transport in realistic image synthesis. New algorithms are presented for exact computation of *direct illumination* effects related to light emission, shadowing, and first-order scattering from surfaces. New theoretical results are presented for the analysis of *global illumination* algorithms, which account for all interreflections of light among surfaces of an environment.

First, a closed-form expression is derived for the *irradiance Jacobian*, which is the derivative of a vector field representing radiant energy flux. The expression holds for diffuse polygonal scenes and correctly accounts for shadowing, or partial occlusion. Three applications of the irradiance Jacobian are demonstrated: locating local irradiance extrema, direct computation of isolux contours, and surface mesh generation.

Next, the concept of irradiance is generalized to tensors of arbitrary order. A recurrence relation for irradiance tensors is derived that extends a widely used formula published by Lambert in 1760. Several formulas with applications in computer graphics are derived from this recurrence relation and are independently verified using a new Monte Carlo method for sampling spherical triangles. The formulas extend the range of non-diffuse effects that can be computed in closed form to include illumination from directional area light sources and reflections from and transmissions through glossy surfaces.

Finally, new analysis for global illumination is presented, which includes both direct illumination and indirect illumination due to multiple interreflections of light. A novel operator equation is proposed that clarifies existing deterministic algorithms for simulating global illumination and facilitates error analysis. Basic properties of the operators and solutions are identified which are not evident from previous formulations. A taxonomy of errors that arise in simulating global illumination is presented; these include perturbations of the boundary data, discretization error, and computational error. A priori bounds are derived for each category using properties of the proposed operators.

ANALYTIC METHODS FOR
SIMULATED LIGHT TRANSPORT

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
James Richard Arvo
December 1995

Dedicated to the memory of my parents

Mathilda Marie Arvo and Helmer Matthew Arvo

Acknowledgements

There are many people I wish to thank. First and foremost, I offer my sincere thanks to my advisor, Martin Schultz, for his unhesitating support. Without his encouragement and his steadfast confidence in me, this thesis would never have been completed. I am also deeply indebted to Ken Torrance, who served as an external committee member. Ken has been a wonderful influence and a guiding light. I am grateful to him for his kindness, his constant willingness to listen and to help, and the encouragement he offered when it was most needed. Because of Ken, I have learned to trust my own instincts.

I also wish to thank my other internal committee members, Ken Yip and Vladimir Rhoklin, who were a pleasure to interact with, and my external committee members, Al Barr, Pat Hanrahan, and John Hughes, who generously offered their time and expertise. Thanks also to Michael Fischer for kindly helping me to find my way to completion.

I am grateful to Don Greenberg for the opportunity to be part of the Program of Computer Graphics at Cornell University, where most of the research for this dissertation was conducted. Special thanks to Erin Shaw, Steve Westin, Pete Shirley, and John Hughes for carefully reading various portions of this work and offering detailed comments. Many thanks to my coauthors Julie Dorsey, Dave Kirk, Kevin Novins, David Salesin, François Sillion, Brian Smits, Ken Torrance, and Steve Westin, from whom I have learned so much over the years, and to Pete Shirley, Dani Lischinski, Bill Gropp, and Jim Ferwerda for enumerable stimulating discussions. Thanks also to Ben Trumbore and Albert Dicruttalo for modeling and software support, to Dan Kartch for all the help with document preparation, to Jonathan Corson-Rikert, Ellen French, Linda Stephens, and Judy Smith for administrative support, and to Hurf Sheldon for many years of cheerful and professional systems support. From my days at Apollo Computer, I'd like to thank Al Lopez, Fabio Pettinati, Ken Severson, and Terry Lindgren for all their encouragement. Many fellow students and assorted friends have also helped and inspired me along the way, including Lenny Pitt, Mukesh Prasad, Michael Monks, Ken Musgrave, Andrew Glassner, Mimi and Noel Mateo, and Susan Vonderheide.

The person to whom I am most indebted is my wife, Erin, whose love and understanding are truly without bound. It is difficult to imagine how I could have completed this work without her support at every step along the way.

This research was conducted at the Program of Computer Graphics, supported by the NSF/ARPA Science and Technology Center for Computer Graphics and Scientific Visualization (ASC-8920219), and performed on workstations generously provided by the Hewlett-Packard Corporation.

Much of the material in this thesis was presented at ACM SIGGRAPH 94 under the titles "The Irradiance Jacobian for Partially Occluded Polyhedral Sources" and "A Framework for the Analysis of Error in Global Illumination Algorithms", and at SIGGRAPH 95 under the titles "Applications of Irradiance Tensors to the Simulation of Non-Lambertian Phenomena" and "Stratified Sampling of Spherical Triangles". The copyright for this material is held by the ACM, and is included in this thesis by permission of the ACM.

Table of Contents

List of Figures

Nomenclature

Symbol	*Page*	*Meaning*
A	71	A subset of $\mathcal{S}^2$
$\mathcal{A}$	64	The algebra of polynomials over $\mathcal{S}^2$
c	28	Speed of light in a vacuum
$\mathbf{C}(a, b, c)$	41	The 3×3 corner matrix
$d\omega$	72	The 2-form corresponding to solid angle
$\mathcal{E}$	19	Particle measure
$f(\mathbf{r}, \mathbf{u})$	27	Radiance function, $\mathbb{R}^3 \times \mathcal{S}^2 \to \mathbb{R}$, [watts/m^2 sr]
$f_0(\mathbf{r}, \mathbf{u})$	136	Emission function, $\mathbb{R}^3 \times \mathcal{S}^2 \to \mathbb{R}$, [watts/m^2 sr]
$\widehat{f}(\mathbf{r}, \mathbf{r}')$	156	Transport intensity, $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, [watts/m^4]
$\widehat{f}_0(\mathbf{r}, \mathbf{r}')$	156	Transport emittance, $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, [watts/m^2]
$\mathcal{F}(A \times B)$	160	The radiant power leaving A that reaches B directly
$\widehat{g}(\mathbf{r}, \mathbf{r}')$	156	The geometry term, [m^{-2}]
$\mathbf{G}$	140	The field radiance operator
$\mathbf{H}$	151	Isomorphism taking surface radiance onto field radiance
h	27	Planck's constant
I	76	The multi-index $(i_1, \ldots, i_n)$
$\mathbf{I}$	141	The identity operator or matrix
$\mathbf{J}_{\mathbf{r}}(F)$	36	The Jacobian matrix of F at $\mathbf{r}$
$k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})$	137	Bidirectional reflectance distribution function, [sr^{-1}]
$\widehat{k}(\mathbf{r}, \mathbf{r}', \mathbf{r}'')$	156	Dimensionless scattering function
$\mathbf{K}$	140	The local reflection operator
L_p	64	The Lebesgue function spaces
m	18	Lebesgue measure over $\mathcal{M}$
$\mathcal{M}$	135	A collection of smooth 2-manifolds in $\mathbb{R}^3$
$\mathbf{M}$	141	The operator $\mathbf{I} - \mathbf{KG}$
$n(\mathbf{r}, \mathbf{u})$	23	Phase space density, $\mathbb{R}^3 \times \mathcal{S}^2 \to \mathbb{R}$, [m^{-3} sr^{-1}]
$\widehat{n}(\mathbf{r})$	26	Total phase space density, $\mathbb{R}^3 \to \mathbb{R}$, [m^{-3}]
$\mathbf{n}(\mathbf{r})$	30	Surface normal function, $\mathcal{M} \to \mathcal{S}^2$
P	29	A surface in $\mathbb{R}^3$, usually a polygon

Symbol	*Page*	*Meaning*
$\mathbb{P}$	17	Five-dimensional phase space $\mathbb{R}^3 \times \mathcal{S}^2$
$\mathcal{P}$	17	A σ-algebra of subsets of $\mathbb{P}$
$\mathbf{p}(\mathbf{r}, \mathbf{u})$	137	Ray casting function, $\mathcal{M} \times \mathcal{S}^2 \to \mathcal{M}$
$\mathbf{q}(\mathbf{r}, \mathbf{r}')$	137	Two-point direction function, $\mathcal{M} \times \mathcal{M} \to \mathcal{S}^2$
$\mathbf{Q}(p)$	41	The 3×3 cross product matrix p
r	79	The Euclidean lenght of the vector $\mathbf{r}$
$\mathbf{r}$	23	A point in $\mathbb{R}^3$ or in $\mathcal{M}$, depending on context
$\mathbb{R}$	17	The real numbers
$\mathcal{S}^2$	17	The unit sphere in $\mathbb{R}^3$
$\mathbf{T}^n(A)$	71	Integral of $\mathbf{u}^n$ over $A \subset \mathcal{S}^2$
$\mathbf{T}^{n,q}(A, \mathbf{w})$	111	Integral of $\mathbf{u}^n / (\mathbf{u} \cdot \mathbf{w})^q$ over $A \subset \mathcal{S}^2$
$u(\mathbf{r})$	62	Monochromatic energy density, $\mathbb{R}^3 \to \mathbb{R}$, [joules/m^3]
$\mathbf{u}$	74	A point in $\mathcal{S}^2$ (i.e. a direction), usually $\mathbf{r} / \lVert \mathbf{r} \rVert$
W_k	38	A vector normal to the kth face of a polygonal cone
X	135	A space of radiance functions
X_n	164	A finite-dimensional subspace of X
$\boldsymbol{\beta}_k^n$	118	A Monte Carlo estimator for $\mathbf{T}^n$
$\bar{\bar{\beta}}_k^n$	119	A Monte Carlo estimator for $\bar{\bar{\tau}}^{n,m}$
δ_{ij}	69	Kronecker delta function
Γ_k	35	The unit normal to the kth face of a polygonal cone
ε_{ijk}	70	Levi-Civita (permutation) symbol
$\eta(i, j, k)$	66	Integral of the monomial $x^i y^j z^k$ over $\mathcal{S}^2$
Θ_k	34	The angle subtended by the kth edge of a polygon
κ_{I}^i	76	Number of times i occurs in the multi-index I
μ	140	Cosine-weighted measure over $\mathcal{S}^2$
$\nu(\mathbf{r}, \mathbf{u})$	136	Distance to the manifold $\mathcal{M}$ from $\mathbf{r}$ along $\mathbf{u}$
$\boldsymbol{\xi}$	118	A random variable distributed over a subset of $\mathcal{S}^2$
$\boldsymbol{\Xi}^n$	76	An nth-order tensor related to $\eta(i, j, k)$
$\boldsymbol{\Pi}(S)$	29	Projection of the set $S \subset \mathbb{R}^3$ onto the unit sphere
$\rho_{\mathbf{r}}(\theta', \phi', \theta, \phi)$	137	Bidirectional reflectance distribution function, [sr^{-1}]

Symbol	*Page*	*Meaning*
ϱ	17	Phase space measure
σ	18	Lebesgue measure on the sphere $\mathcal{S}^2$
$\bar{\tau}^n$	88	nth-order axial moment
$\bar{\bar{\tau}}^{n,m}$	90	Double-axis moment
$\Upsilon(x, y)$	112	A two-parameter special function
$\Phi(\mathbf{r})$	30	Vector irradiance, $\mathbb{R}^3 \to \mathbb{R}^3$, [watts/m^2]
$\phi(\mathbf{r})$	30	Irradiance, $\mathcal{M} \to \mathbb{R}$, [watts/m^2]
$\varphi(\mathbf{r}, \mathbf{u})$	27	Phase space flux, $\mathbb{R}^3 \times \mathcal{S}^2 \to \mathbb{R}$, [m^{-2} sr^{-1}]
$\mathcal{X}_A(x)$	117	The characteristic function of the set A
$\Psi(\mathbf{r})$	62	Radiation pressure tensor, $\mathbb{R}^3 \to \mathbb{R}^9$, [joules/m^3]
ω	145	A bound on directional-hemispherical reflectivity
$\Omega_i(\mathbf{r})$	30	The incoming hemisphere of $\mathcal{S}^2$ at $\mathbf{r} \in \mathcal{M}$
$\Omega_o(\mathbf{r})$	138	The outgoing hemisphere of $\mathcal{S}^2$ at $\mathbf{r} \in \mathcal{M}$
$\Vert \cdot \Vert_p$	142	L_p-norm for radiance functions
$\langle \cdot \rangle$	118	Expected value
$\langle \cdot \mid \cdot \rangle$	152	Inner product
$[\, \cdot \mid \cdot \,]$	125	Normalized orthogonal component of a vector
,	79	Within a subscript, indicates partial derivatives
$\equiv$	17	Equal by definition
$\ll$	21	Absolute continuity (partial order on measures)
$\wedge$	68	Exterior product for differential forms
!!	67	Double factorial

Chapter 1

Introduction

One of the central problems of computer graphics is the creation of physically accurate synthetic images from complete scene descriptions. The computation of such an image involves the simulation of *light transport*, the large-scale interaction of light with matter. Within computer graphics, the problem of determining the appearance of an environment by simulating the transport of light within it is known as *global illumination.* While direct simulations of this type are easier in many respects than the inverse problems of computer vision, there remain many unsolved problems. The difficulties arise from complex geometries and surface reflections encountered in real scenes, and from the "mutual illumination" of objects in a scene by interreflected light. Fortunately, the predominant effects are time invariant and occur at scales much larger than the wavelength of visible light, which permits simplified models of light transport to be applied with little sacrifice in fidelity. As a result, the physical principles that underlie global illumination come primarily from geometrical optics, radiometry, and radiative transfer.

The current trend in image synthesis research toward increasing physical accuracy stems from a desire to make images that are not only realistic but are also *predictive.* Prediction is clearly a requirement for applications such as architectural and automotive design. One strategy for reliably predicting the appearance of a hypothetical scene from its physical description is to first simulate the physics of light transport, then approximate the visual stimulus of viewing the scene by mapping the result to a display device. Adhering to physical principles also makes the process of realistic image generation intuitive. It is far easier to control a simulation using familiar physical concepts than through arcane parameters.

The effectiveness of a synthetic image hinges on more than correct physics. Other factors include characteristics of the display device such as nonlinearities and limited dynamic range, the physiology of the eye, and even higher-level cognitive aspects of perception. Nevertheless, it is the physical and computational aspects of the problem that currently dominate the field. Consequently, global illumination is largely the study of algorithms for the simulation of visible light transport, which includes the processes of light emission, propagation, scattering, and absorption.

The central contributions of this thesis are 1) an improved theoretical foundation for the study of these processes, and 2) a collection of new computational tools for their simulation. In particular, methods of functional analysis are employed to quantify the process of light transport, which allows for analysis of error, and efficient algorithms are developed for computing various aspects of illumination and reflection in simplified settings.

1.1 Light Transport and Image Synthesis

This section summarizes the underlying physical principles and computational procedures of light transport and justifies their application to the problem of global illumination. Some historical context is also provided, both to clarify the origin of the fundamental ideas, and also to emphasize connections with other fields.

1.1.1 Physical Principles

Image synthesis involves the simulation of visual phenomena; those that are observable by the eye or some instrument capable of discerning light intensity and frequency. This level of description is sometimes referred to as *phenomenological* as it focuses on phenomena corresponding to the percepts of brightness and color.

There are many levels of physical description that can predict and explain the phenomenology of light. *Geometrical optics* adequately describes several large-scale behaviors of light, such as linear propagation, reflection, and refraction, but does not incorporate *radiometric* concepts that quantify light [78]. *Physical optics* is based on Maxwell's equations for electromagnetic radiation, which subsumes geometrical optics and includes additional phenomena such as dispersion, interference, and diffraction; these effects can dominate at scales near or below the wavelength of light. This level of description is also quantitative. However, physical optics is overly detailed at large scales when the radiation field is incoherent, which is true of macroscopic environments illuminated by common lighting instruments, or *luminaires*.

A third level of description is known as the *transport* level; in the context of electromagnetic radiation, the study of transport processes is known as *radiative transfer* [26,122]. Radiative transfer combines principles of geometrical optics and thermodynamics to characterize the flow of radiant energy at scales large compared with its wavelength and during time intervals large compared to its frequency [115, 111]. Central to the theory of radiative transfer are the principles of *radiometry*; the measurement of light. This level of description is compatible with the phenomenology of light without being overly detailed [122], which makes it appropriate to the task of global illumination and image synthesis.

Radiative transfer does not explain all phenomena at the level of electromagnetism or quantum mechanics, yet it may incorporate information from more detailed physical descriptions such as these. More complete theories that operate at microscopic scales can be used to predict first-order effects such as local scattering and absorption, which enter into simulations at larger scales as boundary or initial conditions. For instance, the reflection model of He et al. [63] employs physical optics to characterize reflection from rough surfaces; the model can then be incorporated into a global illumination algorithm operating at the transport level [149]. Another avenue by which physical optics can enter into the simulation is through physical measurement. For instance, the reflectance properties associated with boundaries (surfaces) may be obtained from samples of real materials using a gonioreflectometer [173].

Global illumination can also be placed within the much broader context of *transport theory* [20,65,40], a field encompassing all macroscopic phenomena that result from the interaction of infinitesimal particles within a medium. The macroscopic behaviors of photons, neutrons, and gas molecules are all within its purview. The unifying concepts of transport theory help to clarify the nature of global illumination and its relation to other physical problems [8]. For instance, radiative transfer, neutron transport, and gas dynamics all emphasize volume interactions, collisions involving the three-dimensional medium through which the particles migrate, while the boundary conditions are of secondary importance. In contrast, global illumination generally neglects volume scattering altogether but incorporates boundary conditions that are far more complex in terms of geometry and surface scattering distributions. This distinction gives the solution methods for global illumination a unique character. Nevertheless, there are points of contact with other problems; for example, when a participating medium is included in global illumination, the governing equation is virtually identical to that of neutron migration [79].

1.1.2 Computational Aspects

Many image synthesis techniques in use today are physically-based, yet none account for the entire repertoire of optical effects that are observable at large scales. The failure is generally not in the physical model, but rather in the computational methods. Nearly all the limitations result from the difficulty of faithfully representing the distributions of light scattering from real materials, and in accurately simulating the long-range effects of scattered light. To deal with this complexity, global illumination algorithms frequently incorporate idealized reflection models such as Lambertian (ideal diffuse), specular (mirror-like), or a combination of the two. Both of these extremes are easy to represent and correspond to well-understood computational procedures.

Ray tracing was introduced by Appel [5] and significantly extended by Whitted [178] to incorporate many of the principles of geometric optics; however the method neglects all multiple scatterings of light except those from mirror-like surfaces. The radiosity method, first applied to image synthesis by Goral et al. [52, 108], accounts for multiple diffuse interreflections using techniques from thermal engineering, yet it does not accommodate mirrored surfaces. Methods combining ray tracing and radiosity typically neglect more complex modes of reflection, such as glossy surfaces [150]. Other techniques accommodate more complex surface reflections [71,23,149,12] but either suffer from statistical errors or fail to handle extremely glossy surfaces; in addition, none account for light scattering by participating media such as smoke or fog. Of the methods that can account for participating media, none can accommodate surfaces with locally complex geometry or reflectance functions [134,135,73].

In studying the computational aspects of global illumination, it is often convenient to partition illumination into two components: *direct* and *indirect.* By direct illumination we mean the processes of light emission from luminaires (area light sources), propagation through space, and subsequent scattering at a second surface. By indirect illumination we mean light that undergoes multiple scatterings. Both of these aspects are *global* in that well-separated objects of a scene can influence one another's appearance, either by blocking light (casting shadows) in the case of direct lighting, or by scattering light (reflecting it) in the case of indirect lighting.

The direct and indirect components of illumination offer different computational challenges. Techniques for direct illumination include modeling the distribution of light emitted from luminaires [168,37], computing features of irradiance such as shadow boundaries [108] and gradients [175], and modeling surface reflections [177, 51]. Handling these effects is a prerequisite to simulating indirect illumination.

Indirect illumination is important to consider because a significant portion of the illumination in a room, for example, may come from multiply scattered light [101]. Both finite element methods [64,184,32] and Monte Carlo methods [71, 35,146] have been applied to the simulation of global illumination, which includes both direct and indirect illumination.

The most significant computational aspects of the finite-element approaches for global illumination are 1) discretizing, 2) computing element interactions, and 3) solving a linear system. Discretization includes surface mesh generation, which is one of the more challenging problems of global illumination. Thus far, the most effective meshing schemes have been based on adaptive refinement, such as the hierarchical scheme of Hanrahan et al. [62], *a priori* location of derivative discontinuities, as proposed by Heckbert [64] and Lischinski et al. [93], or a combination of the two, as proposed by Lischinski et al. [94].

Computing element interactions is perhaps the most costly aspect of global illumination. This is a significant computational challenge because it involves a potentially large number of multi-dimensional integrals over irregular domains. Each multi-dimensional integral represents the transfer of light among discrete elements and corresponds to a coupling term of the finite element matrix. The number of interactions can be large because they are non-local; any element may potentially interact with any other element. The regions are irregular because of occlusion and local geometric complexity of surfaces. The integrals are minimally two-dimensional, but may involve as many as six dimensions for non-diffuse surfaces; moreover, the integrands may be discontinuous due to changing visibility. Owing to these complexities, the problem is frequently approximated using Monte

Carlo integration. There are virtually no tools available for solving these problems analytically when non-diffuse reflection is involved.

Finally, every global illumination problem involves the solution of a linear system at some level. Because the finite element matrices can be large and dense, it is important to exploit some structure of the matrix to speed its construction and solution. For example, both block matrices [62] and wavelet representations [53] have been used for this purpose.

1.1.3 Historical Background

Global illumination draws from many fields, two of which are of particular relevance: radiative transfer and illumination engineering. The origins and underlying physical model of global illumination can be traced directly to the theory of radiative transfer, while the aims and methodologies of global illumination are today most closely aligned with those of illumination engineering.

The birth of radiative transfer theory is generally attributed to the astrophysical work of Schuster [142] and Schwarzschild [143] near the turn of the century. The theory was initially developed for the study of stellar atmospheres. Among the first problems encountered were the inverse problems of inferring physical properties of a stellar atmosphere from the light it emitted. To solve the inverse problems it was first necessary to obtain a governing equation for the direct process of radiative transfer through a hot atmosphere. The resulting governing equation, known today as the *equation of transfer*, is an integro-differential equation that describes the large-scale time-averaged interaction of light with matter and accounts for emission, absorption, and scattering. The emphasis in astrophysics, as in most engineering applications, is on modeling interactions with participating media. With the appropriate boundary conditions, however, the equation of transfer can also account for arbitrary surface reflections, which is the aspect that is emphasized in global illumination.

The concepts of radiative transfer have subsequently found application in areas such as illumination engineering [100], thermal engineering [148,68], hydrologic optics [123], agriculture [103,102], remote sensing [22,48], computer vision [180], and computer graphics [52]. The equation of transfer was also the starting point for the development of neutron transport theory, in which the diffusion of neutrons through matter is governed by essentially the same principles as the diffusion of photons through a participating medium [119,158]. This accounts for the many similarities between solution methods for global illumination and simulated neutron migration [151], particularly among Monte Carlo methods [8].

1.2 Thesis Overview

The contributions of this thesis are both theoretical and practical. The theoretical aspects of the work are an attempt to strengthen the mathematical foundation of realistic image synthesis. Although computer graphics has drawn from a number of related fields with more developed foundations, many of the features unique to this field are not yet well understood. For instance, numerous techniques have been adopted from finite element analysis, yet many fundamental questions concerning the nature of the solutions in the context of global illumination have not been addressed. New formalisms that begin to rectify this are presented.

The practical aspects of the work provide a number of useful computational tools for image synthesis. The focus is on new closed-form solutions for a variety of sub-problems relating to direct illumination; in particular, those involving non-diffuse emission and scattering. Previously, the few tools that existed for computing illumination from area light sources were limited to diffuse luminaires. New expressions are derived here that accommodate luminaires with directionally varying brightness; these expressions permit us to handle a much larger class of luminaires analytically.

Chapter 2 introduces the fundamental building blocks of radiometry that are used throughout the thesis. The concepts are developed in an untraditional manner using the formalism of measure theory. The approach is inspired by the axiomatic foundation of radiative transfer theory put forth by R. W. Preisendorfer [120]. This approach illustrates the origin of the most basic principles that underlie global illumination.

Chapter 3 introduces a new tool that is derived directly from a widely-used result attributed to Lambert. The tool is a closed-form expression for the derivative of the irradiance (a measure of incident energy) due to diffuse polygonal light sources. The interesting aspect of the problem is in correctly handling occlusions. The chapter describes the complication introduced by partial occlusion, provides a complete solution for diffuse polygonal luminaires, and demonstrates several applications.

Chapter 4 introduces the most powerful tools of the thesis. A tensor generalization of irradiance is proposed that leads to a number of new computational methods involving non-diffuse surfaces. The expressions derived in this chapter allow a number of fundamental computations to be done in closed-form for the first time. The new expressions and the algorithms for their efficient evaluation are independently verified in chapter 5 by comparison with Monte Carlo estimates.

Chapter 6 introduces new theoretical tools for the study of global illumination. The focus is on a new way to express the governing equation for global illumination in terms of linear operators with very convenient properties. These properties lend themselves to standard methods of error analysis, which are described in chapter 7. This final chapter also discusses the sources of error in solving the new operator equation, and relates these to existing global illumination algorithms.

Throughout the thesis the emphasis is on physical rather than perceptual aspects of light transport. This bias is reflected in the consistent use of radiometric rather than photometric units, which take the response of the human visual system into account. Thus, the admittedly difficult issues of display and perception are largely ignored. Moreover, this work does not address participating media, transparent surfaces, time-dependent or transient solutions, or probabilistic solution methods other than those used for independent verification. The characteristics of the problems considered here may be summarized as follows:

- Monochromatic radiometric quantities (radiance, reflectivities, emission)
- A linear model of radiative transfer
- Direct radiative exchange among opaque surfaces
- Steady-state solutions
- Deterministic boundary element formulations

Many applications meet these restrictions because environments that we wish to simulate frequently come from everyday experience, where these assumptions tend to be valid. For instance, at habitable temperatures re-emission of absorbed light by most materials is effectively zero at visible wavelengths, which permits us to simulate different wavelengths independently and also leads to a linear model of light. Atmospheric effects over tens of meters are insignificant under normal circumstances, so we may assume that surfaces exchange energy directly with no attenuation. Furthermore, because changes in our surroundings take place slowly with respect to the speed of light, transient solutions are of little interest. Also, since common sources of light are incoherent, effects related to phase, such as interference, are usually masked. While diffraction and interference can be observed in diffraction gratings as well as in natural objects such as butterfly wings and thin films [152], these do not dominate architectural settings and other scenes that we commonly wish to simulate.

Finally, this work focuses on deterministic algorithms primarily to limit the scope, but also because deterministic methods are vital to Monte Carlo. Frequently both the accuracy and efficiency of Monte Carlo methods can be vastly improved by first solving a nearby problem deterministically and estimating the difference stochastically. Thus, it is hoped that fundamental advances made for deterministic algorithms will also be useful in the context of Monte Carlo, which usually has much broader applicability.

1.3 Summary of Original Contributions

This section summarizes the original contributions of the thesis, which are organized into three major categories: I) new derivations of basic physical quantities such as radiance II) new methods for direct illumination involving both diffuse and non-diffuse surfaces, and III) new tools for the analysis of global illumination. Parts I and III are primarily of theoretical interest, while part II presents a number of practical algorithms.

I: Derivations of Fundamental Physical Quantities [Chapter 2]

The most fundamental quantity of radiative transfer, known as *radiance*, is derived from macroscopic properties of abstract particles using formalisms from measure theory. The new approach clarifies some of the assumptions hidden within classical definitions and demonstrates deep connections with other physical quantities. Well-known properties of radiance, such as constancy along rays in free space, are shown to hold using the new formalism.

II: Methods for Direct Illumination [Chapters 3, 4, and 5]

Derivatives of Irradiance: A closed-form expression is derived for the *Irradiance Jacobian* (the derivative of vector irradiance) that holds in the presence of occlusions. The new expression makes possible a number of computational techniques for diffuse polygonal scenes, such as generating isolux contours and finding local irradiance extrema. A new meshing scheme based on isolux contours is also demonstrated.

Irradiance Tensors: A natural generalization of irradiance is presented that embodies high-order moments of the radiation field. The new concept is a useful tool for deriving formulas involving moments of radiance distribution functions.

Extending Lambert's Formula: Irradiance tensors are shown to satisfy a recurrence relation that is a natural generalization of a fundamental formula for irradiance derived by Lambert in the 18th century. The new formula extends Lambert's formula to non-diffuse phenomena.

Moment Methods for Non-Lambertian Phenomena: Closed-form solutions for moments of irradiance from polygonal luminaires are presented, along with efficient algorithms for their evaluation. The expressions are derived using irradiance tensors. The new algorithms are demonstrated by simulating three different non-diffuse phenomena, each computable in closed-form for the first time: glossy reflection, glossy transmission, and directional area luminaires.

Results on Non-polygonal and Inhomogeneous Luminaires: The problem of computing irradiance from spatially varying luminaires is reduced to that of integrating rational polynomials over the sphere. A generalization of irradiance tensors is introduced to handle rational moments. It is shown that these integrals generally cannot be evaluated in terms of elementary functions, even in polygonal environments.

Stratified Sampling of Solid Angle: A direct Monte Carlo sampling algorithm for spherical triangles is derived. The method allows stratified sampling of the solid angle subtended by a polygon. The algorithm is used to construct a low-variance estimator for irradiance tensors and related expressions, providing independent validation of the closed-form expressions.

III: Analysis of Global Illumination [Chapters 6 and 7]

Linear Operator Formulation: A new operator formulation of the well-known governing equation for global illumination is presented, and is shown to be equivalent to the *rendering equation*. The new formulation cleanly separates notions of geometry and reflection, allowing these aspects to be studied independently. Several related operators, such as adjoints, are easily analyzed using the new operators.

Theoretical Bounds on Operators and Radiance Functions: Bounds are computed for the new operators using principles of thermodynamics and basic tools of functional analysis. Based on this analysis, it is shown that the process of global illumination is closed with respect to all L_p spaces.

Taxonomy of Errors with a priori Bounds: Using the proposed linear operators and standard methods of analysis, *a priori* bounds are computed for three categories of error: perturbed boundary data, discretization, and computation.

Chapter 2

Particles and Radiometry

In this chapter we define a number of fundamental *radiometric quantities*, that is, concepts pertaining to the measurement of light that are essential to the study of radiative transfer. Rather than enumerating standard definitions, we shall instead deduce the important concepts starting from "observable" behaviors of particle-based phenomena that we take as axioms. This approach will expose some of the assumptions that are hidden within the classical definitions.

The simulation of any real process or phenomenon involves simplifying assumptions that affect both the physical model and its mathematical representation. Physical assumptions are introduced to limit the scope of the problem, usually by ignoring or altering known physical laws. This always involves a compromise between what is desired and what can be effectively simulated. Mathematical assumptions include those in which structure is imposed that is not necessarily present in reality. For instance, a problem may be embedded in a richer mathematical framework so that well-understood methods of analysis can be applied. Assumptions of both types are made in defining radiometric quantities.

The most common and far reaching mathematical assumption in radiometry is that light and matter can be faithfully represented by continuous analogues. The utility of this assumption is that it allows us to express new quantities in terms of limiting processes that can only be approximated by actual experiments. By embedding the phenomena we wish to study in a continuum, we may employ all the usual machinery of mathematical analysis, such as measure theory and differential geometry. In the physical world, of course, the corresponding limiting processes cease to make sense below a certain scale [80]. Consequently, the mathematical abstractions generally extend beyond the phenomena they model.

We shall begin with macroscopic time-averaged properties of light that are theoretically observable by an eye or a laboratory instrument, and deduce the basic concepts of photometry and radiometry; that is, quantitative aspects of light that are based on a continuum. We shall adopt the common simplifying assumption that light may be adequately modeled as a flow of non-interacting neutral particles [169]. From elementary principles that operate at the macroscopic scale we then construct functions that correspond to *radiance* and related quantities, which will be essential building blocks in the chapters that follow.

Much of the large-scale behavior of particles can be understood in terms of abstract particles with minimal semantics; that is, particles with only those features common to photons, neutrons, and other neutral particles. This observation typifies the point of view taken in *transport theory*, which is the study of abstract particles and their interaction with matter [40]. Transport theory applies classical notions of physics at the level of discrete particles to predict large-scale statistical behaviors. The emphasis on statistical explanations differentiates transport theory from other classical theories such as electromagnetism. Essential to the theory are several fundamental simplifying assumptions about the particles. For example, it is assumed that 1) the particles are so small and numerous that their statistical distribution can be treated as a continuum, and 2) at any point in time a particle is completely characterized by its position, velocity, and internal states such as polarization, frequency, charge, or spin [88,179].

These assumptions lead naturally to the concept of *phase space*, an abstraction used in representing the spatial and directional distribution and internal states of a collection of particles. We shall employ a continuous phase space whose points specify possible particle states. A distribution of particles can then be represented as a density function defined over phase space. The aim of transport theory is to determine such a density function, commonly describing a distribution of particles at equilibrium, based on individual particle behaviors as well as the geometrical and physical properties of the medium through which the particles travel [25,40].

The setting of abstract particles is sufficient to deduce a primitive version of radiance, as well as other radiometric quantities, using only measure-theoretic concepts and elementary facts about the behavior of groups of particles. Each distinct quantity corresponds to a different form of channeling of the particles, with radiance being the most fundamental for radiative transfer. To deduce the properties of radiance, which will be needed in subsequent chapters, we must ultimately impose further semantics on the abstract particles that will distinguish them as photons.

2.1 Phase Space Measure

Phase space is an abstraction used in many fields to represent configurations of discrete particles. Frequently the term connotes a $6N$-dimensional Euclidean space with each point encoding the position and velocity of N distinct particles. A single point of such a phase space completely specifies the configuration of all N particles, and the time-evolution of the particles defines a space-curve [89].

For many physical problems a phase space of far fewer dimensions suffices, which has both conceptual and computational advantages. For example, when the particles cannot influence one another, their aggregate behavior may be determined by characterizing the behavior of a single particle. Each phase space dimension then represents one physical degree of freedom of a particle, with the space as a whole representing all possible particle states. We shall assume that the particles are independent, which implies that particles may interact with their surroundings but not with each other. This assumption is valid to very high accuracy for rarefied gases, neutrons, and photons when interference is ignored.

We shall further assume that particle-matter interaction, or *scattering*, consists of collisions that are either perfectly elastic and perfectly inelastic. That is, a particle either retains its original internal state after scattering, such as energy or wavelength, or is completely absorbed. This precludes processes by which particles migrate from one energy band to another; for photons this may happen when a particle is absorbed and re-emitted, as in blackbody radiation, phosphorescence, and fluorescence.

In transport theory this simplification is known as the *one-speed* assumption, since speed is proportional to energy for most particles, and energy is assumed to remain constant despite multiple collisions. When applied to photons, however, the term implies constant wavelength rather than speed. Hence, this simplification is also known as the *gray* assumption in the context of radiative transfer [158].

Finally, we assume that the particles possess no internal states other than energy or wavelength, which is fixed for all particles. Thus, we assume monochromatic radiation and ignore phenomena such as polarization. Under these assumptions, each particle has only five degrees of freedom: three for position and two for direction. The corresponding 5-dimensional phase space is

$$\mathbb{P} \equiv \mathbb{R}^3 \times \mathcal{S}^2, \tag{2.1}$$

where $\mathbb{R}^3$ is Euclidean 3-space and $\mathcal{S}^2$ is the unit sphere in $\mathbb{R}^3$.

The abstraction of a distribution of particles in phase space requires some additional structure for the purpose of analysis. In order to quantify collections of particles and define various transformations, it is necessary to endow the space with a *measure*. We therefore introduce the concept of *phase space measure* consisting of the triple $(\mathbb{P}, \mathcal{P}, \varrho)$, where ϱ is a positive set function, or measure, defined on the elements of a set $\mathcal{P}$, which is a σ-algebra of subsets of $\mathbb{P}$. That is, $\mathcal{P}$ is a collection of subsets that contains $\mathbb{P}$ and is closed under complementation and countable unions [132]. The elements of $\mathcal{P}$ are called the *measurable* sets.

We shall construct the phase space measure $(\mathbb{P}, \mathcal{P}, \varrho)$ from the natural Lebesgue measures on $\mathbb{R}^3$ and $\mathcal{S}^2$, which correspond to standard notions of volume and surface area; we shall denote these measures by v and σ respectively. We then define ϱ to be the product measure $v \times \sigma$ and form the σ-algebra $\mathcal{P}$ from the two component σ-algebras. More formally, we define $\mathcal{P}$ be the *completion* of the cartesian product of the two σ-algebras, which is again a σ-algebra [132, p. 29]. Completion simply extends the collection of measurable sets to include those formed by a cartesian product of an unmeasurable set with a set of measure zero.

Lebesgue measure is the appropriate abstraction to adopt here since it is invariant under rigid transformation and permits the consistent definition of a family of useful function spaces. However, the use of such formalisms requires that the domain of ϱ be restricted to $\mathcal{P}$, the σ-algebra of subsets of $\mathbb{P}$. This restriction is necessary in Euclidean spaces of three or more dimensions, where measures that are defined on all subsets cannot possess certain essential properties. For example, such measures cannot be invariant under rigid motions, a property we require of space and the particles that travel through it. The Banach-Tarski paradox [14] is a more colorful example of why such a restriction on the domain is necessary. This famous result states that a measure defined on all subsets of Euclidean 3-space must either assign the same measure to all volumes, or allow unacceptable consequences [144]; for example, one unit sphere could be decomposed into a finite number of pieces and reassembled into two identical spheres of the same volume [127].

2.2 Particle Measures

Phase space measure applies only to the space on which the particles are defined and not the distributions themselves. To quantify the distributions and relate them to phase space measure we require several additional properties that apply to collections of non-interacting particles. We shall identify four such properties, which we take to be axiomatic; that is, we shall not provide explanations in terms of more basic phenomena. These axioms closely parallel macroscopic properties of real particles, although the abstractions transcend their physical counterparts in several respects. The axioms lead naturally to the concept of a *particle measure*, from which we may deduce basic properties of neutral particle migration that underlie radiative transfer and other particle-based phenomena.

2.2.1 Particle Axioms

Phenomena such as radiation that arise from microscopic particles can exhibit large-scale features only through the correlated motion of collections or *ensembles* of particles. We may therefore express properties of particles that are relevant to large-scale transport in terms of these ensembles. We shall consider four such properties that are motivated by physical intuition and correspond to real phenomena that can be measured at the macroscopic scale, at least in principle [122]. From these properties we may deduce the existence of a density function over phase space that exactly represents the configuration of particles.

We begin by assuming the existence of a field of non-interacting one-speed particles in space. Every region of space then has a *particle content*, a number indicating how many particles exist within that region. This number may be further partitioned according to the directions in which the particles move. By associating values to subsets of space and direction we are defining a real-valued *set function* over phase space, which we shall denote by $\mathcal{E}$. We assert that $\mathcal{E}$ obeys the following physically plausible axioms:

(A1) $\mathcal{E} : 2^{\mathbb{P}} \to [0, \infty)$

(A2) $A \cap B = \emptyset \implies \mathcal{E}(A \cup B) = \mathcal{E}(A) + \mathcal{E}(B)$

(A3) $\varrho(A) < \infty \implies \mathcal{E}(A) < \infty$

(A4) $\varrho(A) = 0 \implies \mathcal{E}(A) = 0$

Here A and B are subsets of $\mathbb{P}$ and, in the case of the last two axioms, A and B must also be measurable with respect to ϱ. Axiom A1 states that $\mathcal{E}$ is a positive set function defined on all subsets of the phase space $\mathbb{P}$; to every subset there corresponds a non-negative value indicating its particle content.

Axiom A2 states that the set function $\mathcal{E}$ is additive. By repeated application of this axiom $\mathcal{E}$ is clearly additive over any finite collection of disjoint sets. When applied to photons, this property illustrates a divergence of the phenomenological formulation from physical optics. The electromagnetic description of light includes effects that can violate the additivity property; for instance, distinct volumes in phase space may interact via interference. Thus, in the context of photon transport axiom A2 restricts us to incoherent sources of light, as interference can arise whenever the illumination has some degree of coherence [13].

Axioms A3 and A4 establish the only connection between the set function $\mathcal{E}$ and the phase space through which the particles migrate. Specifically, axiom A3 states that every finite volume of phase space has a finite particle content, and axiom A4 states that a region of phase space with zero volume cannot contain a meaningful number of particles. The latter axiom has important implications as it simultaneously disallows two physically implausible situations; point sources and perfectly collimated thin beams. Because ϱ is a product measure, axiom A4 implies that $\mathcal{E}(D \times \Omega) = 0$ when $v(D) = 0$, which is true at isolated points, or when $\sigma(\Omega) = 0$, which is true of any perfectly collimated beam. These properties greatly influence the nature of radiometric calculations since any meaningful transfer of energy will entail both spatial and directional integration.

2.2.2 Extension to a Measure

Axioms A1-A4 express macroscopic properties of particles by means of a set function $\mathcal{E}$ that is similar to a measure. With slight alterations to two of the axioms, $\mathcal{E}$ can be made a genuine measure. Doing so will allow us to study radiance functions as elements of the standard L_p or *Lebesgue* function spaces, which are defined in terms of the Lebesgue integral. Toward this end, we weaken A1 by restricting its domain, and strengthen A2 to apply to infinite sums over mutually disjoint sets. Specifically, we replace axioms A1 and A2 with

(A1*) $\mathcal{E} : \mathcal{P} \to [0, \infty)$

(A2*) $\mathcal{E}\left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathcal{E}(A_i)$ for mutually disjoint $\{A_i\} \subset \mathcal{P}$.

Axiom A1* only requires $\mathcal{E}$ to be defined on the σ-algebra $\mathcal{P}$, while axiom A2* states that $\mathcal{E}$ is countably additive. In summary, axioms A1*, A2*, A3, and A4 taken together state that the function $\mathcal{E}$ is a positive countably additive set function defined on a σ-algebra. By definition then, $\mathcal{E}$ is a *positive measure* [59].

With this interpretation of $\mathcal{E}$, axiom A4 now has additional significance; given that $\mathcal{E}$ and ϱ are both measures over the same σ-algebra, axiom A4 states that the particle measure is *absolutely continuous* with respect to the phase space measure. Absolute continuity defines an order relation on the set of measures over the same σ-algebra, which is usually denoted by "$\ll$" [132, p. 128]. The notation $m_1 \ll m_2$ simply means that $m_1(E) = 0$ whenever $m_2(E) = 0$.

Note that $\mathcal{E}(\mathbb{P})$ need not be finite since an infinite number of particles may exist within an infinite volume of phase space without violating basic physical principles. However, a particle measure will always possess the weaker property of being σ-finite; that is, it is finite on each element of some countable partition of the domain [132]. This follows immediately from axiom A3 and the fact that the phase space measure ϱ is σ-finite.

Given the abstraction of independent particles in a phase space $\mathbb{P}$ equipped with a measure ϱ, axioms A1*, A2*, A3, and A4 imply that the concept of particle content coincides with the concept of a positive measure. We summarize this point, which is largely a matter of definitions, in the following theorem.

Theorem 1 (Existence of Particle Measures) *Given a configuration of non-interacting one-speed particles in an abstract phase space* $\mathbb{P}$ *with measure* ϱ, *if the particle content function* $\mathcal{E}$ *satisfies axioms A1*, A2*, A3, and A4, then* $\mathcal{E}$ *defines a positive* σ*-finite measure over* $\mathbb{P}$ *with* $\mathcal{E} \ll \varrho$.

We call the three-tuple $(\mathbb{P}, \mathcal{P}, \mathcal{E})$ the *particle measure space*, and $\mathcal{E}$ the *particle measure*. The existence of measures analogous to $\mathcal{E}$ were taken as axiomatic by Preisendorfer [120,121]; here we have identified the physical and mathematical assumptions from which they may be deduced.

2.3 Phase Space Density

To characterize the distribution of particles in a continuous phase space, we require a function known as *phase space density* defined over the space. We now show that such a density function emerges naturally from the concept of particle measure; the central observations parallel the measure-theoretic development of radiative transfer due to Preisendorfer [120].

By theorem 1 the particle content of any $A \in \mathcal{P}$ is given by a σ-finite measure that is absolutely continuous with respect to ϱ. This characterization of particle content is sufficient to deduce the existence of phase space density; the final step is provided by the Radon-Nikodym theorem [183, p. 93] [132, p. 130].

Theorem 2 (Radon-Nikodym) *Let* $(B, \mathcal{B}, m)$ *be a measure space over the set* B, *and let* ν *be a* σ*-finite positive measure defined on the* σ*-algebra* $\mathcal{B}$. *If* $\nu \ll m$, *then there exists a non-negative* m*-integrable function* p *such that*

$$\nu(E) = \int_E p \, dm \tag{2.2}$$

for all $E \in \mathcal{B}$ *with* $m(E) < \infty$. *Moreover, the function* p *is unique to within a set of* m*-measure zero.*

The function p in the above theorem is commonly referred to as a *density function.* While any non-negative function p can be used to define a new measure ν by equation (2.2), the Radon-Nikodym theorem shows that under rather mild conditions the converse also holds; in particular, the density function p can be recovered almost uniquely from the two measures m and ν, provided that $\nu \ll m$. Theorems 1 and 2 together imply the following theorem.

Theorem 3 (Existence of Phase Space Density) *For every configuration of non-interacting one-speed particles in an abstract phase space* $\mathbb{P}$ *with measure* ϱ*, there exists a* ϱ*-measurable function* $n : \mathbb{P} \to \mathbb{R}^+$*, which is unique to within a set of* ϱ*-measure zero, satisfying*

$$\mathcal{E}(A) = \int_A n \, d\varrho, \tag{2.3}$$

where $\mathcal{E}(A)$ *denotes the particle content, and* $A \subset \mathbb{P}$ *is a measurable subset.*

The function n, whose existence is guaranteed by the above theorem, is called the *phase space density.* By virtue of equation (2.3), this function is also referred to as the *Radon-Nikodym derivative* of $\mathcal{E}$ with respect to ϱ, since it exhibits algebraic properties analogous to a standard derivative [59]. To emphasize the similarity with differentiation, it is typically denoted by

$$n = \frac{d\mathcal{E}}{d\varrho}, \tag{2.4}$$

which also suggests dimensional relationships. Since $\mathcal{E}(A)$ is dimensionless, n must have the dimensions of inverse phase space volume, or $\mathrm{m}^{-3}\mathrm{sr}^{-1}$. We shall treat n as a function of two variables and write $n(\mathbf{r}, \mathbf{u})$, where $\mathbf{r} \in \mathbb{R}^3$ and $\mathbf{u} \in \mathcal{S}^2$.

Phase space density makes no mention of material attributes such as mass or energy; these concepts do not enter until we assign physical meaning to the particles. This abstract nature of phase space density makes it quite universal; in one form or another it underlies virtually all particle transport problems including radiative transfer, neutron migration, and gas dynamics. When further semantics

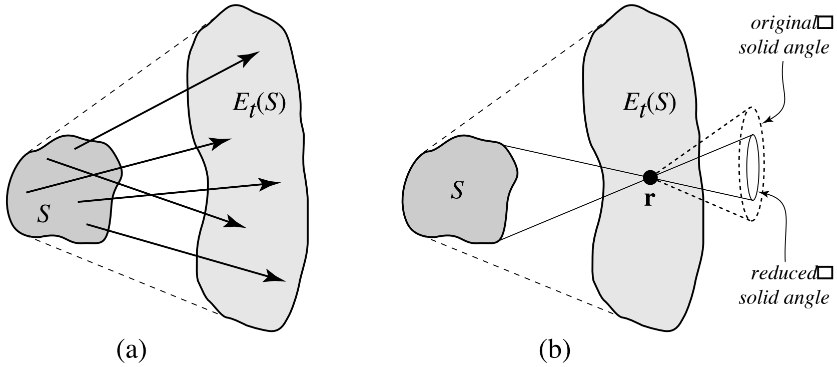

Figure 2.1: *Phase space density remains constant along straight lines in empty space, but total phase space density decreases with the time-evolution of a region of phase space.*

such as energy content are added to the particles, the meaning is inherited by the resulting density function and is also reflected in its physical units.

Part of the semantics of a particle is its behavior with time. We have assumed neutral non-interacting particles, which implies that the particles travel in straight lines, except when interacting with matter. This behavior is consistent with the notions of geometrical optics. Another time-related semantic property is that particles travel at a fixed speed, which is true of monochromatic photons within a uniform medium. These properties give rise to a crucial property of phase space density that is inherited by concepts such as radiance.

Theorem 4 (Invariance of Phase Space Density) *In steady-state, the phase space density of one-speed particles is constant along straight lines in empty space.*

<u>Proof:</u> We consider the time-evolution of a collection of neutral non-interacting one-speed particles in a region of phase space. Let S be an arbitrary subset of $\mathbb{P}$, and define the time-evolution operator E_t by

$$E_t(S) \equiv \{\ (\mathbf{r} + tv\mathbf{u}, \mathbf{u}) : (\mathbf{r}, \mathbf{u}) \in S\ \}\,, \tag{2.5}$$

where v is the constant speed of the particles. This function defines a new set by expanding the spatial component of S in the direction of travel of its particles, leaving the set of directions unchanged. Thus, any particle whose position and direction of travel are within the set S at time t_0 will also be in the set $E_t(S)$ at time $t_0 + t$. Similarly, by conceptually reversing the direction of travel of the particles we see that any particle in $E_t(S)$ at time t_1 must have been in the set S at time $t_1 - t$. Note that particles arriving from other directions may exist within the spatial extent of $E_t(S)$ at time t_1, but these do not contribute to the content of $E_t(S)$. It follows that $E_t(S)$ contains exactly those particles previously existing in S. See Figure 2.1a. In steady state this equality holds at all times. Thus,

$$\mathcal{E}(S) = \mathcal{E}(E_t(S)) \tag{2.6}$$

for all $t > 0$ such that the particles of S encounter no matter. Consequently,

$$\int_S n(\mathbf{r}, \mathbf{u}) \, d\varrho \;=\; \int_{E_t(S)} n(\mathbf{r}, \mathbf{u}) \, d\varrho \;=\; \int_S n(\mathbf{r} + tv\mathbf{u}, \mathbf{u}) \, d\varrho \tag{2.7}$$

for all $S \in \mathcal{P}$ and $t > 0$ provided that the evolution of S into $E_t(S)$ takes place in empty space. When these conditions are met, equation (2.7) yields

$$\int_S \left[n(\mathbf{r}, \mathbf{u}) - n(\mathbf{r} + tv\mathbf{u}, \mathbf{u}) \right] \, d\varrho = 0, \tag{2.8}$$

which implies that $n(\mathbf{r}, \mathbf{u}) = n(\mathbf{r} + tv\mathbf{u}, \mathbf{u})$ for almost all $\mathbf{r} \in \mathrm{I\!R}^3$ and $\mathbf{u} \in \mathcal{S}^2$ such that the points $\mathbf{r}$ and $\mathbf{r} + t\mathbf{u}$ are mutually visible; that is, separated by empty space. □□

The classical demonstration of the above invariance principle proceeds by equating the radiant energy passing consecutively though two differential surfaces or portals [98,68]. Here we have reached the same conclusion by considering the time evolution of arbitrary ensembles of particles within a phase space.

This invariance principle has important implications. For example, under the given assumptions, the directional derivative of phase space density must be zero in every direction and at all points in empty space. This is equivalent to

$$\mathbf{u} \cdot \nabla n(\mathbf{r}, \mathbf{u}) \;=\; 0, \tag{2.9}$$

where the gradient is with respect to the spatial variable. As we shall see, this fact also causes derivatives of related quantities to vanish.

The invariance of phase space density along lines is not shared by other closely related quantities. For instance, consider the quantity known as *total phase space density*, which is defined by

$$\widehat{n}(\mathbf{r}) \equiv \int_{\mathcal{S}^2} n(\mathbf{r}, \mathbf{u}) \, d\sigma(\mathbf{u}), \tag{2.10}$$

where σ is the canonical measure on the sphere [19, p. 276]. Given a bounded set $S \in \mathcal{P}$, the total phase space density must everywhere decrease in the set $E_t(S)$ when t is sufficiently large. This follows from that fact that $\widehat{n}(\mathbf{r})$ is bounded above by the product of the solid angle subtended by S at $\mathbf{r}$ and the maximum phase space density attained within S. See Figure 2.1b. A number of quantities related to $\widehat{n}$ are introduced in chapter 4, where we investigate moments of radiance.

2.4 Phase Space Flux

It is frequently more convenient to characterize the density of particles by their rate of flow across a real or imaginary surface. To develop this idea we shall now proceed using traditional heuristic arguments based on infinitesimal quantities.

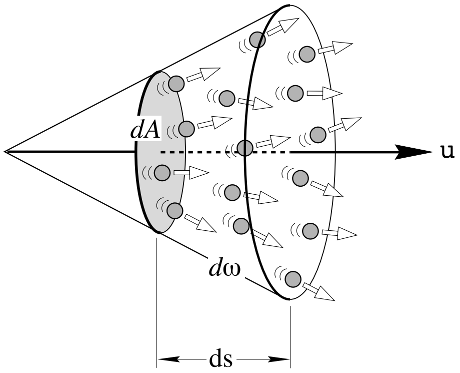

Figure 2.2: *Phase space flux is the number of particles crossing the surface dA perpendicular to* **u** *per unit area, per unit solid angle, per unit time.*

Consider the particles that pass through a differential area dA in time dt with directions of travel within a differential solid angle $d\omega$ about the surface normal. All of the particles are contained within the volume $dA\,ds$, where $ds = v\,dt$, as shown in Figure 2.2. Assuming that $\mathbf{r}$ is a point within this volume, the particle content of the volume is given by

$$n(\mathbf{r}, \mathbf{u})\ dA\ ds\ d\omega. \tag{2.11}$$

But the particles can also be counted using the rate at which they cross the surface dA. That is, the particle content of the volume $dA\,ds$ is also given by

$$v\,n(\mathbf{r}, \mathbf{u})\ dA\ d\omega\ dt, \tag{2.12}$$

since $ds = v\,dt$. This method of counting the particles, which substitutes a temporal dimension for a spatial dimension, motivates the notion of *phase space flux*, denoted φ, which is defined by

$$\varphi(\mathbf{r}, \mathbf{u}) \equiv v\,n(\mathbf{r}, \mathbf{u}) \qquad \left[\frac{1}{\mathrm{m}^2\ \mathrm{sr}\ \mathrm{s}}\right]. \tag{2.13}$$

The concept of phase space flux can be used to express virtually any quantity relating to the macroscopic distribution of particles.

2.5 Radiance

Radiance is the physical counterpart to the physiological concept of brightness or intensity, and is the most ubiquitous concept in radiometry. We denote the radiance at the point $\mathbf{r}$ and in the direction $\mathbf{u}$ by $f(\mathbf{r}, \mathbf{u})$. At the macroscopic scale, this function completely specifies the distribution of incoherent monochromatic radiant energy in the medium; consequently, all radiometric quantities may be defined in terms of radiance.

To define radiance from phase space flux we need only introduce the concept of energy per particle, given by $h\nu$, where h is Planck's constant and ν is frequency.

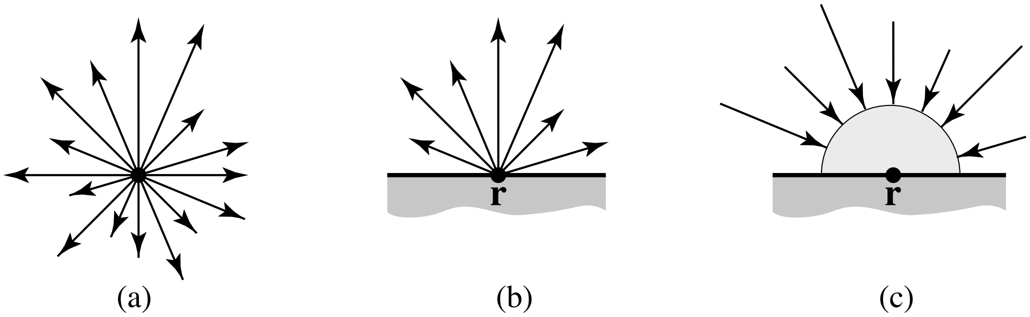

Figure 2.3: *There is no distinction between incoming and outgoing rays at points in space. At surfaces the radiance distribution is naturally partitioned into surface radiance (outgoing) and field radiance (incoming).*

We then have

$$\begin{aligned} f(\mathbf{r},\mathbf{u}) &\equiv h\nu\, \varphi(\mathbf{r},\mathbf{u}) \\ &= ch\nu\, n(\mathbf{r},\mathbf{u}), \end{aligned} \tag{2.14}$$

where c is the speed of light. Radiance therefore has the units of joules/m^2 sr sec, or equivalently, watts/m^2 sr.

Radiance is defined at all points in space and in all directions, as is phase space flux. However it is convenient to define several restrictions of the radiance function. We shall adopt the terminology of Preisendorfer [122] in naming these restrictions. When the point $\mathbf{r} \in \mathrm{I\!R}^3$ is fixed, the function of direction $f(\mathbf{r}, \cdot)$ is called the *radiance distribution function* at the point $\mathbf{r}$; this function is defined at all points in space as well as on surfaces. Note that a radiance distribution function is distinct from a *radiant intensity distribution*, which is a mathematical construct, with the units of watts/sr, that characterizes a point light source. We shall not make use of the latter concept here.

When $\mathbf{r}$ is constrained to lie on a surface, $\mathbf{r} \in \mathcal{M}$, the surface tangent plane naturally partitions the radiance distribution function at each point into two components corresponding to incoming and outgoing radiation. We define *surface radiance* to be the radiance function restricted to a given surface $\mathcal{M}$ and to direc-

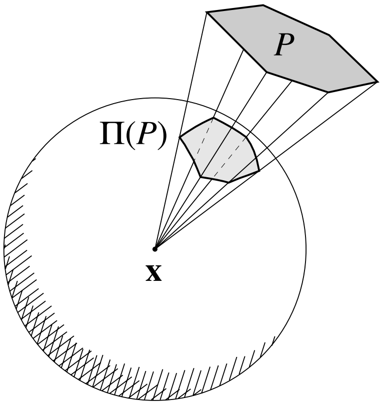

Figure 2.4: *The surface P and its spherical projection $\mathbf{\Pi}(P)$.*

tions pointing away from the surface; by definition it is zero elsewhere. Similarly, we define *field radiance* to be the other half of the radiance function restricted to $\mathcal{M}$; that is, the component with directions pointing toward the surface. See Figure 2.3. The directions that are exactly tangent to the surface are to be ignored in both cases, as they comprise a set of measure zero are physically meaningless.

2.6 Irradiance

Additional radiometric quantities can be defined in terms of radiance by means of weighted integrals over various domains. To define *irradiance* and related quantities, we require a convenient notation for solid angle. Let $P \subset \mathbb{R}^3$ be a surface and consider its projection onto the unit sphere centered at the point $\mathbf{x}$, as shown in Figure 2.4. Formally, we define its *spherical projection* $\mathbf{\Pi}_{\mathbf{x}}(P)$ by

$$\mathbf{\Pi}_{\mathbf{x}}(P) \equiv \{\, \mathbf{u} \in \mathcal{S}^2 : \mathbf{x} + t\mathbf{u} \in P \text{ for some } t > 0 \,\}. \tag{2.15}$$

The solid angle subtended by the surface P can be defined in terms of this operator. If σ is Lebesgue measure on the sphere, then $\sigma(A)$ is the surface area of any measurable subset $A \subset \mathcal{S}^2$. The *solid angle* subtended by the surface P with respect to the point $\mathbf{x} \in \mathbb{R}^3$ is then $\sigma(\mathbf{\Pi}_{\mathbf{x}}(P))$, the surface area of its spherical projection. Since we may always select a coordinate system in which the center of projection $\mathbf{x}$ is at the origin, we shall omit the point $\mathbf{x}$ without loss of generality, which will simplify notation.

The quantity known as *irradiance* is a density function defined at surfaces; it corresponds to the radiant power per unit area reaching the surface at each point, which is the most basic characterization of the illumination reaching a surface. We denote irradiance by $\phi(\mathbf{r})$, where $\mathbf{r} \in \mathcal{M}$. By definition

$$\phi(\mathbf{r}) \equiv \int_{\Omega_i} f(\mathbf{r}, \mathbf{u}) \, \cos\theta \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{watts}}{\text{m}^2}\right], \tag{2.16}$$

where Ω_i is the hemisphere of incoming directions with respect to the surface at the point $\mathbf{r}$, and $f(\mathbf{r}, \mathbf{u})$ denotes a monochromatic field radiance function. The angle θ within the integral is the incident angle of the vector $\mathbf{u}$ with respect to the surface at $\mathbf{r}$. Thus, $\cos\theta = |\mathbf{u} \cdot \mathbf{n}(\mathbf{r})|$, where $\mathbf{n}(\mathbf{r})$ is the unit normal to the surface at $\mathbf{r}$. The presence of the $\cos\theta$ accounts for the fact that an incident pencil of radiation is spread over larger areas near grazing, thereby decreasing the density of its radiation. This fact is sometimes called the *cosine law* [125].

Given the role of the surface normal in the above definition, it is convenient to introduce a new vector-valued function $\Phi(\mathbf{r})$ defined by

$$\Phi(\mathbf{r}) \equiv \int_{\mathcal{S}^2} \mathbf{u} \, f(\mathbf{r}, \mathbf{u}) \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{watts}}{\text{m}^2}\right]. \tag{2.17}$$

The new function is called the *vector-irradiance* [122] at the point $\mathbf{r}$; other common names for the same quantity are *light vector* [46] and *net integrated flux* [80]. This vector quantity does not depend on surface orientation and is defined at all points in $\mathbb{R}^3$. It corresponds to the time-averaged Poynting vector of the electromagnetic field [46,122]. The resulting vector field over $\mathbb{R}^3$ is known as the *light field* [46]. It follows from equations (2.16) and (2.17) that

$$\phi(\mathbf{r}) = -\mathbf{n}(\mathbf{r}) \cdot \Phi(\mathbf{r}), \tag{2.18}$$

when only field radiance is considered; that is, when the vector irradiance does not include light reflected or emitted from the surface. Furthermore, from equation (2.9) it follows that

$$\nabla \cdot \Phi(\mathbf{r}) \;=\; \int_{\mathcal{S}^2} \mathbf{u} \cdot \nabla f(\mathbf{r}, \mathbf{u}) \, d\sigma(\mathbf{u}) \;=\; 0. \tag{2.19}$$

A function with zero divergence is said to be *solenoidal* [46].

Both *radiant exitance* and *radiosity* have definitions similar to that of irradiance and share the same units; that is, watts/m^2. The difference is that the domain of integration is the outgoing hemisphere rather than the incident hemisphere. The semantic difference between these two quantities is that radiant exitance refers to emitted light, while radiosity includes both emitted and reflected light.

Chapter 3

Derivatives of Irradiance

A perennial problem of computer graphics is the accurate representation of light leaving a surface. In its full generality, the problem entails modeling both the local reflection phenomena and the distribution of light reaching the surface. Frequently the problem is simplified by assuming polyhedral environments and Lambertian (diffuse) emitters and reflectors. With these simplifications the remaining challenges are to accurately model the illumination and shadows resulting from area light sources and occluders, and to simulate interreflections among surfaces.

Many aspects of surface illumination have been studied in order to accurately model features of the incident or reflected light. In previous work, Heckbert [64] and Lischinski et al. [93] identified derivative discontinuities in irradiance and used them to construct effective surface meshes. Nishita and Nakamae [107,108] located penumbrae due to occluders in polyhedral environments, while Teller [160] performed an analogous computation for a sequence of portals to locate antipenumbrae. Ward and Heckbert [175] and Vedel [166] estimated irradiance gradients by Monte Carlo ray tracing, and used them to improve the interpolation of irradiance functions. Vedel [167] also computed the gradient analytically in unoccluded cases. Salesin et al. [136] and Bastos et al. [16] employed gradients to construct higher-order interpolants for irradiance functions. Drettakis and Fiume [39] estimated gradients as well as isolux contours from a collection of discrete samples and used them to guide subsequent sampling, placing more samples where the curvature of the isolux contours was large, for example. Drettakis and Fiume [38] and Stewart and Ghali [157] proposed methods for incrementally computing visible portions of a luminaire to increase the efficiency of many of the above methods.

This chapter introduces a new computational tool that applies to diffuse polyhedral environments and is useful in any application requiring derivatives of irradiance. The central contribution is a closed-form expression for the derivative of vector irradiance (as defined in chapter 2), which is termed the *irradiance Jacobian*. The new expression properly accounts for occlusion and subsumes the irradiance gradient as a special case.

3.1 Introduction

The irradiance at a point on a surface due to a polyhedral luminaire of uniform brightness can be computed using a well-known formula due to Lambert. In this chapter we derive the corresponding closed-form expression for the *irradiance Jacobian*, the derivative of the vector representation of irradiance. Although the result is elementary for unoccluded luminaires, within penumbrae the irradiance Jacobian must incorporate more information about blockers than is required for the computation of either irradiance or vector irradiance. The expression presented here holds for any number of polyhedral blockers and requires only a minor extension of standard polygon clipping to evaluate. To illustrate its use, three related applications are briefly described: direct computation of isolux contours, finding local irradiance extrema, and iso-meshing. Isolux contours are curves of constant irradiance across a surface that can be followed using a predictor-corrector method based on the irradiance Jacobian. Similarly, local extrema can be found using a gradient descent method. Finally, iso-meshing is a new approach to surface mesh generation that incorporates families of isolux contours.

In Section 3.2 we derive the irradiance Jacobian for polygonal luminaires of uniform brightness starting from a closed-form expression for the vector irradiance. In previous work the same closed-form expression was used in scalar form by Nishita and Nakamae [107] to accurately simulate polyhedral sources, and by Baum et al. [18] for the computation of form factors, which relate to energy transfers among surface patches. Section 3.3 introduces a method for characterizing changes in the apparent shape of a source due to differential changes in the receiving point, which is the key to handling occlusions. In Section 3.4 basic properties of the irradiance Jacobian are discussed, including existence and the connection with gradients.

To illustrate the potential uses of the irradiance Jacobian, Section 3.5 describes several computations that employ irradiance gradients. We describe a method for direct computation of isolux contours, which are curves of constant irradiance on a surface. Each contour is expressed as the solution of an ordinary differential equation which is solved numerically using a predictor-corrector method. The resulting contours can then be used as the basis of a meshing algorithm. Finding local extrema is a related computation that can be performed using a descent method.

3.1.1 Lambert's Formula

For sources of uniform brightness, Φ can be expressed analytically for a number of simple geometries including spheres and infinite strips [46]. Polygonal sources are another important class with known closed-form expressions, and are the focus of this chapter. Suppose P is a simple planar polygon in $\mathbb{R}^3$ with vertices $v_1, v_2, \ldots, v_n$. If P is a diffuse source with constant radiant exitance M [watts/m^2], then the light field due to P is given by

$$\Phi(\mathbf{r}) = \frac{M}{2\pi} \sum_{i=1}^{n} \Theta_i(\mathbf{r})\, \Gamma_i(\mathbf{r}), \tag{3.1}$$

where $\Theta_1, \ldots, \Theta_n$ are the angles subtended by the n edges as seen from the vantage point $\mathbf{r}$, or equivalently, the arclengths of the edges projected onto the unit sphere

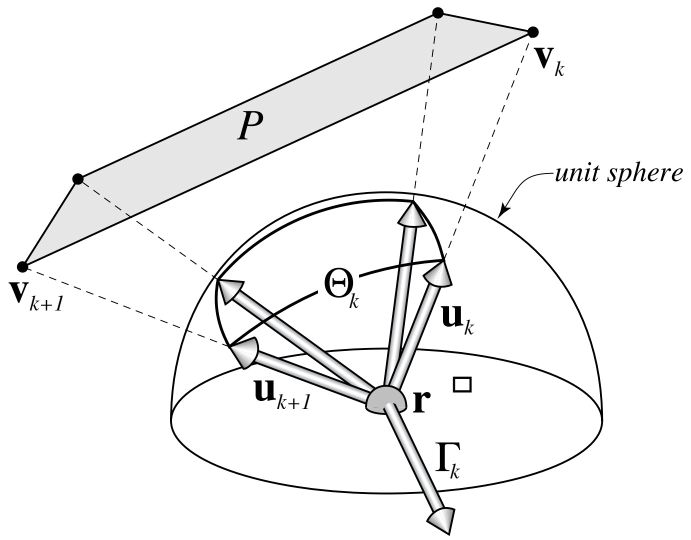

Figure 3.1: *The vector irradiance at point* $\mathbf{r}$ *due to a diffuse polygonal luminaire* P *can be expressed in closed form. The contribution due to edge* k *is the product of the angle* Θ_k *and the unit vector* Γ_k*.*

about $\mathbf{r}$. The vectors $\Gamma_1, \ldots, \Gamma_n$ are unit normals of the polygonal cone with cross section P and apex at $\mathbf{r}$, as shown in Figure 3.1. For any $1 \leq k \leq n$ the functions Θ_k and Γ_k can be written

$$\Theta_k(\mathbf{r}) = \cos^{-1}\left(\frac{v_k - \mathbf{r}}{||\, v_k - \mathbf{r}\, ||} \cdot \frac{v_{k+1} - \mathbf{r}}{||\, v_{k+1} - \mathbf{r}\, ||} \right), \tag{3.2}$$

and

$$\Gamma_k(\mathbf{r}) = \frac{(v_k - \mathbf{r}) \times (v_{k+1} - \mathbf{r})}{||\, (v_k - \mathbf{r}) \times (v_{k+1} - \mathbf{r})\, ||}, \tag{3.3}$$

where $||\cdot||$ is the Euclidean norm and $v_{n+1} \equiv v_1$. Equation (3.1) most commonly appears in scalar form, which includes the dot product with the surface normal [18, 43,107]. With $M = 1$, the corresponding expression $-\Phi(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r})$ is the form factor between a differential patch at $\mathbf{r}$ and the polygonal patch P. Equivalent expressions have been independently discovered by Yamauti [181], Fok [43], and Sparrow [155]. However, as noted by Schröder and Hanrahan [140], the scalar expression appeared much earlier in Lambert's treatise on optics, published in 1760 [85]. Henceforth, equation (3.1) shall be referred to as Lambert's formula.

The light field defined by Gershun [46], and described in chapter 2, is a true vector field. That is, the vector irradiance due to multiple sources may be obtained by summing the vector irradiance due to each source in isolation. Thus, polyhedral sources can be handled by applying equation (3.1) to each face and summing the resulting irradiance vectors. Alternatively, when the faces have equal brightness, equation (3.1) can be applied to the outer contour of the polyhedron as seen from the point **r**, as described by Nishita and Nakamae [107]. Partially occluded sources can be handled similarly, by summing the contributions of all the visible portions. Determining the visible portions of the sources in polyhedral environments is analogous to clipping polygons for hidden surface removal [176].

The closed-form expression for vector irradiance in equation (3.1) provides an effective means of computing related expressions, such as derivatives. In the remainder of the chapter we derive closed-form expressions for derivatives of the irradiance and vector irradiance due to polyhedral sources in the presence of occluders, and describe several applications.

3.2 The Irradiance Jacobian

If the function $F : \mathrm{I\!R}^3 \rightarrow \mathrm{I\!R}^3$ is differentiable, then its derivative DF is represented by a 3×3 Jacobian matrix. We shall denote the Jacobian matrix of F at $\mathbf{r}$ by $\mathbf{J_r}(F)$. That is,

$$\mathbf{J_r}(F) \equiv DF(\mathbf{r}) = \left[\frac{\partial F_i(\mathbf{r})}{\partial x_j} \right]. \tag{3.4}$$

In this section we derive the Jacobian matrix of the vector irradiance, which we shall call the *irradiance Jacobian*, when the illumination is due to a diffuse polygonal luminaire. The obvious approach to obtaining an expression for $\mathbf{J_r}(\Phi)$ in this case is to simply differentiate equation (3.1) with respect to the point **r**; this is a

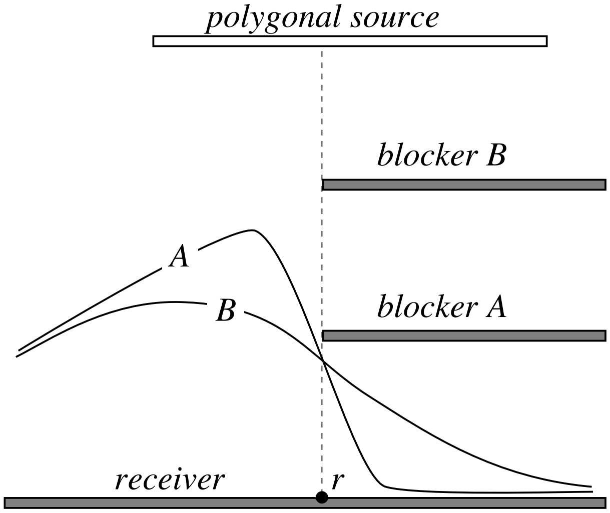

Figure 3.2: *The irradiance at point* **r** *due to source P is the same with either blocker, but the slopes of the irradiance curves are different.*

straightforward exercise that can be performed by a symbolic differentiation package, for example. The resulting expression will be valid for unoccluded polygonal luminaires, but not for partially occluded luminaires.

To see why the irradiance Jacobian is more difficult to compute when blockers are present, consider the arrangement in Figure 3.2. First, observe that the irradiance at the point **r** can be computed by applying equation (3.1) to the visible portion of the source, which is the same in the presence of either blocker A or blocker B. Although the common expression can be differentiated, the resulting derivative does not correspond to the irradiance Jacobian. To see this, observe that the two blockers produce irradiance functions with different slopes at **r**; the blocker closest to the receiving plane produces the sharper shadow, which implies a larger derivative. Consequently, the irradiance Jacobians must also differ in the two cases to account for blocker position.

To derive an expression that applies within penumbrae, we express $\Phi(\mathbf{r})$ in terms of *vertex vectors*, which correspond to vertices of the spherical projection of

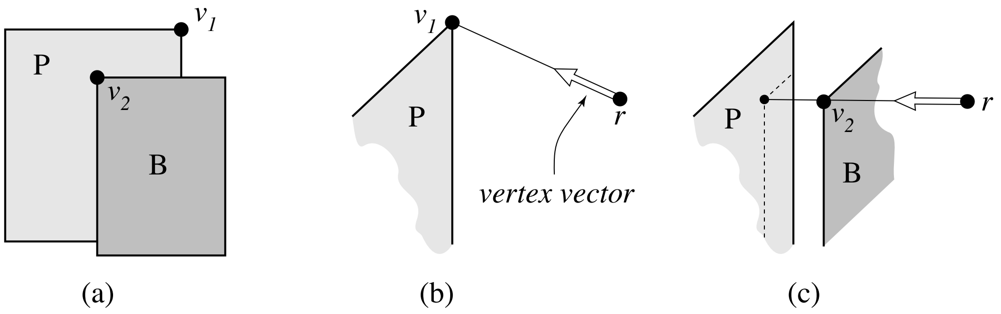

Figure 3.3: *(a) The view from* $\mathbf{r}$ *of the two types of "intrinsic" vertices. (b) The vertex vector for the unoccluded source vertex* v_1*. (c) The vertex vector for the blocker vertex* v_2*, whose projection falls within the interior of* P*.*

the polygon, as depicted in Figure 3.1. Vertex vectors may point toward vertices of two distinct types: *intrinsic* and *apparent*. An intrinsic vertex exists on either the source or the blocker, as shown in Figure 3.3. An apparent vertex results when the edge of a blocker, as seen from $\mathbf{r}$, crosses the edge of the source or another blocker, as shown in Figure 3.4. We shall express $\mathbf{J_r}(\Phi)$ in terms of derivatives of the vertex vectors. Since a vertex vector is a mapping from points in $\mathbb{R}^3$ to unit vectors, its derivative is a 3×3 matrix, which we call the *vertex Jacobian*. The vertex Jacobians account for the geometric details of the vertices, which yields a relatively simple closed-form expression for $\mathbf{J_r}(\Phi)$, even in the presence of occluding objects.

Let $v'_1, v'_2, \ldots, v'_m$ be the vertices of P', the source P after clipping away portions that are occluded with respect to the point $\mathbf{r}$. Without loss of generality, we may assume that P' is a single polygon; if it is not, we simply iterate over the pieces. The vertex vectors $u_1(\mathbf{r}), u_2(\mathbf{r}), \ldots, u_m(\mathbf{r})$ are defined by

$$u_k(\mathbf{r}) \equiv \frac{v'_k - \mathbf{r}}{||\, v'_k - \mathbf{r}\, ||}. \tag{3.5}$$

To simplify notation we also let $w_1(\mathbf{r}), \ldots, w_m(\mathbf{r})$ denote the cross products

$$w_k(\mathbf{r}) \equiv u_k(\mathbf{r}) \times u_{k+1}(\mathbf{r}). \tag{3.6}$$

Henceforth, we assume that u_k and w_k are functions of position and omit the

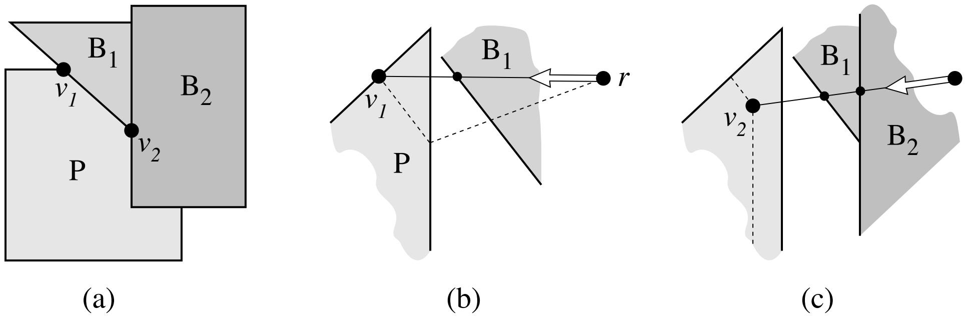

Figure 3.4: *(a) The view from* **r** *of the two types of "apparent" vertices. (b) The vertex vector for* v_1 *results from a blocker edge and a source edge. (c) The vertex vector for* v_2 *results from two blocker edges.*

explicit dependence on **r**. Expressing Θ_k and Γ_k in terms of w_k, we have

$$\Theta_k = \sin^{-1} \| \, w_k \, \| \, , \tag{3.7}$$

and

$$\Gamma_k = \frac{w_k}{\| \, w_k \, \|}. \tag{3.8}$$

Note that equation (3.7) is equivalent to equation (3.2) only for acute angles; that is, only when $u_k \cdot u_{k+1} \geq 0$. Nevertheless, because the new expression for Θ_k simplifies some of the intermediate expressions that appear below, it will be retained for most of the derivation. The restriction to acute angles will eventually be removed, however, so that the final result will apply in all cases.

To compute $\mathbf{J}(\Phi)$ in terms of the vertex Jacobians $\mathbf{J}(u_1), \ldots, \mathbf{J}(u_m)$ we first consider the kth term of the summation in equation (3.1). Differentiating, we have

$$\mathbf{J}(\Theta_k \Gamma_k) = \Gamma_k \, \nabla\Theta_k + \Theta_k \mathbf{J}(\Gamma_k), \tag{3.9}$$

where $\Gamma_k \, \nabla\Theta_k$ is the outer product of the vector Γ_k and the gradient $\nabla\Theta_k$. We now compute $\nabla\Theta_k$ and $\mathbf{J}(\Gamma_k)$. For brevity, we shall denote the vertex vectors u_k

and u_{k+1} by a and b respectively, and the cross product $a \times b$ by w. Then the gradient of Θ_k with respect to $\mathbf{r}$ is

$$\begin{aligned}
\nabla\Theta_k &= \nabla \sin^{-1} \|\, w \,\| \\
&= \frac{1}{\sqrt{1 - w^{\mathrm{T}} w}} \left(\frac{w^{\mathrm{T}}}{\|\, w \,\|} \right) \mathbf{J}(w) \\
&= \left(\frac{w^{\mathrm{T}}}{a^{\mathrm{T}} b} \right) \frac{\mathbf{J}(w)}{\|\, w \,\|}. \qquad (3.10)
\end{aligned}$$

Similarly, differentiating Γ_k with respect to $\mathbf{r}$ we have

$$\begin{aligned}
\mathbf{J}(\Gamma_k) &= D\left(\frac{w}{\|\, w \,\|} \right) \\
&= \frac{\mathbf{J}(w)}{\|\, w \,\|} - \frac{w w^{\mathrm{T}}}{\|\, w \,\|^3} \mathbf{J}(w) \\
&= \left(\mathbf{I} - \frac{w w^{\mathrm{T}}}{w^{\mathrm{T}} w} \right) \frac{\mathbf{J}(w)}{\|\, w \,\|}. \qquad (3.11)
\end{aligned}$$

From Equations (3.7)-(3.11), we obtain an expression for $\mathbf{J}(\Theta_k \Gamma_k)$ in terms of $\mathbf{J}(w)$ and the vertex vectors a and b:

$$\mathbf{J}(\Theta_k \Gamma_k) = \left[\frac{w}{\|\, w \,\|} \left(\frac{w^{\mathrm{T}}}{a^{\mathrm{T}} b} \right) + \sin^{-1} \|\, w \,\| \left(\mathbf{I} - \frac{w w^{\mathrm{T}}}{w^{\mathrm{T}} w} \right) \right] \frac{\mathbf{J}(w)}{\|\, w \,\|}.$$

If the factor of $\sin^{-1} \|\, w \,\|$ is now replaced by the angle between a and b, or $\cos^{-1} a^{\mathrm{T}} b$, then the expression will hold for all angles, removing the caveat noted earlier. The above expression may be written compactly as

$$\mathbf{J}(\Theta_k \Gamma_k) = \mathbf{E}(a, b)\, \mathbf{J}(a \times b), \qquad (3.12)$$

where the function $\mathbf{E}$ is the *edge matrix* defined by

$$\mathbf{E}(a, b) \equiv \left(\frac{1}{a^{\mathrm{T}} b} \right) \frac{w w^{\mathrm{T}}}{w^{\mathrm{T}} w} + \frac{\cos^{-1} a^{\mathrm{T}} b}{\|\, w \,\|} \left(\mathbf{I} - \frac{w w^{\mathrm{T}}}{w^{\mathrm{T}} w} \right). \qquad (3.13)$$

In equation (3.13) we have retained w as an abbreviation for $a \times b$. Because the edge matrix contains no derivatives, it can be computed directly from the vertex vectors a and b. To simplify the Jacobian of $a \times b$, we define another matrix-valued function $\mathbf{Q}$ by

$$\mathbf{Q}(p) \equiv \begin{bmatrix} 0 & p_z & -p_y \\ -p_z & 0 & p_x \\ p_y & -p_x & 0 \end{bmatrix}. \tag{3.14}$$

Then for any pair of vectors p and q, we have $p \times q = \mathbf{Q}(p)q$. Writing the cross product as a matrix multiplication leads to a convenient expression for the Jacobian matrix of $F \times G$, where F and G are vector fields in $\mathrm{I\!R}^3$. Thus,

$$\mathbf{J}(F \times G) = \mathbf{Q}(F)\mathbf{J}(G) - \mathbf{Q}(G)\mathbf{J}(F). \tag{3.15}$$

Applying the above identity to equation (3.12), summing over all edges of the clipped source polygon P', and scaling by $M/2\pi$, we arrive at an expression for the irradiance Jacobian due to the visible portion of polygonal source P:

$$\mathbf{J}(\Phi) = \frac{M}{2\pi} \sum_{i=1}^{m} \mathbf{E}(u_i, u_{i+1}) \left[\mathbf{Q}(u_i)\mathbf{J}(u_{i+1}) - \mathbf{Q}(u_{i+1})\mathbf{J}(u_i) \right]. \tag{3.16}$$

This expression can be simplified somewhat further by collecting the factors of each $\mathbf{J}(u_i)$ into a single matrix. We therefore define the *corner matrix* $\mathbf{C}$ to be the matrix-valued function

$$\mathbf{C}(a, b, c) \equiv \mathbf{E}(a, b)\mathbf{Q}(a) - \mathbf{E}(b, c)\mathbf{Q}(c). \tag{3.17}$$

Then the final expression for the irradiance Jacobian can be written as a sum over all the vertex Jacobians transformed by corner matrices:

$$\mathbf{J}(\Phi) = \frac{M}{2\pi} \sum_{i=1}^{m} \mathbf{C}(u_{i-1}, u_i, u_{i+1})\ \mathbf{J}(u_i), \tag{3.18}$$

where we have made the natural identifications $u_0 \equiv u_m$ and $u_{m+1} \equiv u_1$. Note that each corner matrix $\mathbf{C}$ depends only on the vertex vectors, and not their derivatives. All information about changes in the visible portion of the luminaire due to changes in the position $\mathbf{r}$ is embodied in the vertex Jacobians $\mathbf{J}(u_1), \ldots, \mathbf{J}(u_m)$, which we now examine in detail.

3.3 Vertex Jacobians

To apply equation (3.18) we require the vertex Jacobians, which we now construct for both unoccluded and partially occluded polygonal sources. First, observe that each vertex vector $u(\mathbf{r})$ is a smooth function of $\mathbf{r}$ almost everywhere; that is, $u(\mathbf{r})$ is differentiable at all $\mathbf{r} \in \mathrm{I\!R}^3$ except where two or more edges of distinct polygons appear to coincide, as described in section 3.4. Differentiability follows from the smoothness of the Euclidean norm and the fact that the apparent point of intersection of two skew lines varies quadratically in $\mathbf{r}$ along each of the lines [116]. From this it is evident that the vertex Jacobian exists whenever the real or apparent intersection of two edges exists and is unique.

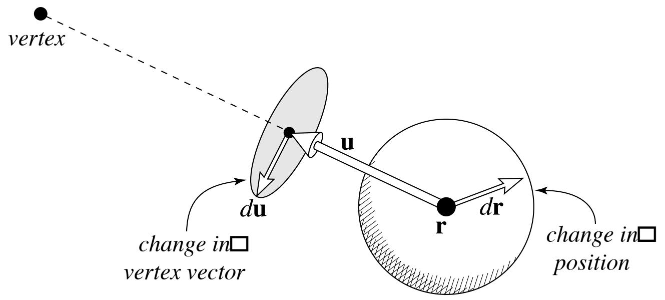

Figure 3.5: *A differential change in the position* $\mathbf{r}$ *results in a change in the unit vertex vector u. The locus of vectors du forms a disk, or more generally, an ellipse in the plane orthogonal to u.*

When the vertex Jacobian exists, it can be constructed by determining its action on each of three linearly independent vectors; that is, by determining the instantaneous change in the vertex vector u as a result of moving $\mathbf{r}$. Differential changes in u are orthogonal to u and collectively define a disk, or in the case of partial occlusion, an ellipse. See Figure 3.5. The rates of change that are easiest obtain are those along the major and minor axes of the ellipse, which are the eigenvectors of the vertex Jacobian. We first treat intrinsic vertices and then generalize to the more difficult case of apparent vertices.

3.3.1 Intrinsic Vertices

Suppose that u is the vertex vector associated with an unoccluded source vertex, as shown in Figure 3.3b. In this case the vertex Jacobian $\mathbf{J}(u)$ is easy to compute since it depends solely on the distance between $\mathbf{r}$ and the vertex, which we denote by α. Moving $\mathbf{r}$ in the direction of the vertex leaves u unchanged, while motion perpendicular to u causes an opposing change in u. The changes in u are inversely proportional to the distance α. This behavior completely determines the vertex Jacobian. Thus, we have

$$\mathbf{J}(u) = -\frac{1}{\alpha}\left(\mathbf{I} - uu^{\mathrm{T}}\right), \tag{3.19}$$

where the matrix $\mathbf{I} - uu^{\mathrm{T}}$ is a projection onto the tangent plane of $\mathcal{S}^2$ at the point u, which houses all differential motions of the unit vector u. The same reasoning applies to vertex vectors defined by a blocker vertex, as in Figure 3.3c. In this case α is the distance along u to the blocker vertex.

3.3.2 Apparent Vertices

Within penumbrae, apparent vertices may be formed by the apparent crossing of non-coplanar edges. The two distinct cases are depicted in Figure 3.4. Let u be the vertex vector associated with such a vertex, where the determining edges are segments of skew lines $\mathcal{L}_1$ and $\mathcal{L}_2$. Let s and t be vectors parallel to $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively, as depicted in Figure 3.6. As in the case of intrinsic vertices, moving $\mathbf{r}$ toward the apparent vertex leaves u unchanged, so $\mathbf{J}(u)u = 0$. To account for other motions, we define the vectors $\widehat{s}$ and $\widehat{t}$ by

$$\begin{aligned} \widehat{s} &\equiv \left(\mathbf{I} - uu^{\mathrm{T}}\right) s \\ \widehat{t} &\equiv \left(\mathbf{I} - uu^{\mathrm{T}}\right) t, \end{aligned}$$

which are projections of s and t onto the plane orthogonal to u. Now consider the change in u as $\mathbf{r}$ moves parallel to $\widehat{s}$, as shown in Figure 3.6a. In this case the apparent vertex moves along $\mathcal{L}_1$ while remaining fixed on $\mathcal{L}_2$. Therefore, the change in u is parallel to $\widehat{s}$ but opposite in direction to the change in $\mathbf{r}$. If α_t is the distance to $\mathcal{L}_2$ along u, we have

$$\mathbf{J}(u)\,\widehat{s} = -\frac{\widehat{s}}{\alpha_t}. \tag{3.20}$$

Evidently, $\widehat{s}$ is an eigenvector of $\mathbf{J}(u)$ with associated eigenvalue $-1/\alpha_t$. A similar argument holds when $\mathbf{r}$ moves along $\widehat{t}$, as shown in Figure 3.6b. Here the apparent vertex moves along $\mathcal{L}_2$ while remaining fixed at $\mathcal{L}_1$. If α_s is the distance to $\mathcal{L}_1$ along u, we have

$$\mathbf{J}(u)\,\widehat{t} = -\frac{\widehat{t}}{\alpha_s}, \tag{3.21}$$

which provides the third eigenvector and corresponding eigenvalue. Collecting these relationships into a matrix equation, we have

$$\mathbf{J}(u)\,\left[\begin{array}{ccc} \widehat{s} & \widehat{t} & u \end{array}\right] = \left[\begin{array}{ccc} -\dfrac{\widehat{s}}{\alpha_t} & -\dfrac{\widehat{t}}{\alpha_s} & 0 \end{array}\right]. \tag{3.22}$$

It follows immediately that whenever the lines $\mathcal{L}_1$ and $\mathcal{L}_2$ are distinct and non-colinear as viewed from the point $\mathbf{r}$, then

$$\mathbf{J}(u) = \mathbf{A}\left[\begin{array}{ccc} -1/\alpha_t & & \\ & -1/\alpha_s & \\ & & 0 \end{array}\right]\mathbf{A}^{-1} \tag{3.23}$$

where $\mathbf{A} \equiv \left[\begin{array}{ccc} \widehat{s} & \widehat{t} & u \end{array}\right]$. Note that equation (3.23) reduces to equation (3.19) when $\alpha_s = \alpha_t$. Equation (3.23) therefore suffices for all vertex vectors, but the special case for intrinsic vertices can be used for efficiency.

3.3.3 Polygon Depth-Clipping

To compute the irradiance or vector irradiance at a point $\mathbf{r}$, it suffices to clip all sources against all blockers, as seen from $\mathbf{r}$, and apply equation (3.1) to the resulting vertex lists. This operation is also sufficient to compute the corner matrices and

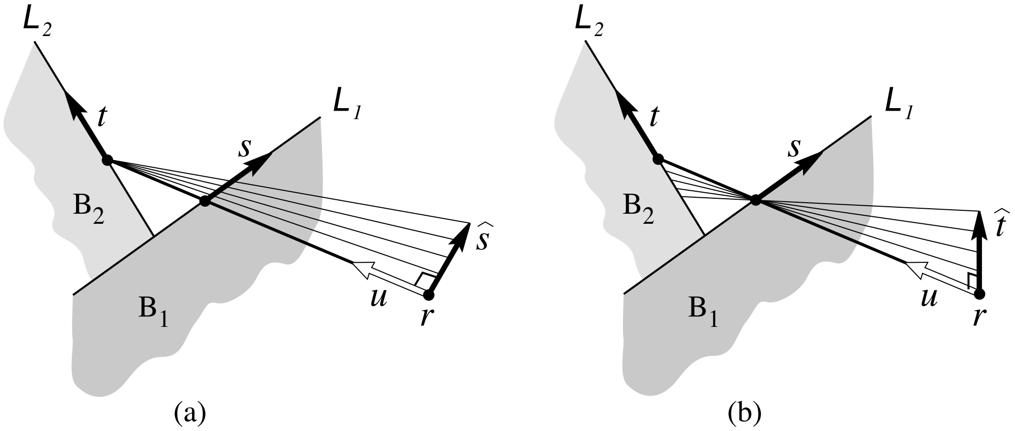

Figure 3.6: *The vertex Jacobian* $\mathbf{J}(u)$ *with respect to two skew lines* $\mathcal{L}_1$ *and* $\mathcal{L}_2$ *is found by determining how the vertex vector* u *changes as* $\mathbf{r}$ *moves parallel to (a) the vector* $\widehat{s}$*, and (b) the vector* $\widehat{t}$*.*

the vertex Jacobians at unoccluded source vertices. However, the vertex Jacobians for the cases illustrated in Figures 3.3c, 3.4b, and 3.4c all require information about the blockers that is missing from traditionally-clipped polygons. Specifically, the distances to the blocker edges that define each vertex are needed to form the matrices in Equations (3.19) and (3.23).

Thus, additional depth information must be retained along with the clipped polygons for use in computing vertex Jacobians. Here we propose a simple mechanism, called *depth clipping*, by which the required information appears as additional vertices. The idea is to construct the clipped polygon using segments of source and blocker edges and joining them by segments called *invisible edges*, which cannot be seen from the point $\mathbf{r}$. See Figure 3.7. The resulting non-planar contour is identical to that of the traditionally-clipped polygon when viewed from $\mathbf{r}$. Each invisible edge defines a vertex Jacobian of the form in equation (3.23); the end points of such an edge encode the two distances from $\mathbf{r}$ while the adjacent edges provide the two directions. Each vertex that is not adjacent to an invisible edge produces a vertex Jacobian of the form in equation (3.19).

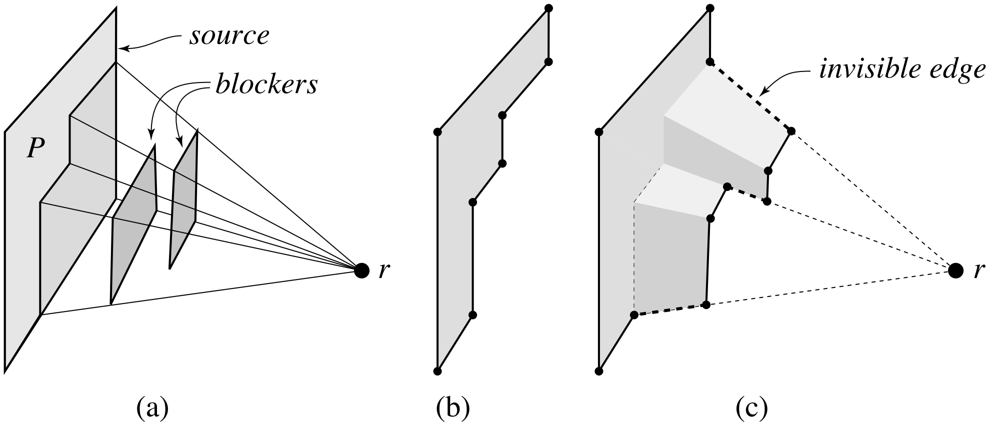

Figure 3.7: *(a) Source P is partially occluded by two blockers as seen from* $\mathbf{r}$*. (b) The vector irradiance at* $\mathbf{r}$ *due to* P *can be computed using the simply-clipped polygon. (c) The irradiance Jacobian at* $\mathbf{r}$ *requires the depth-clipped polygon.*

The depth-clipped polygon and the radiant exitance M completely specify the irradiance Jacobian. Most polygon clipping algorithms can be extended to generate this representation using the plane equation of each blocker. The depth-clipped polygon also illustrates the geometric information required for the computation of irradiance Jacobians.

3.4 Properties of the Irradiance Jacobian

In this section we list some of the basic properties of the irradiance Jacobian, beginning with existence. By definition, the Jacobian $\mathbf{J_r}(\Phi)$ exists wherever Φ is differentiable, which requires the existence of each directional derivative at $\mathbf{r}$. Because we consider only area sources, the variation of Φ is *continuous* along any line except when a blocker is in contact with the receiving surface. Instantaneous occlusion causes discontinuous changes in the vector irradiance. In the absence of contact occlusions, the variation of Φ is not only continuous but *differentiable* everywhere except along lines where edges appear to coincide; that is, points at which a source or blocker edge appears to align with another blocker edge [93].

For instance, when both blockers are present simultaneously in Figure 3.2, the irradiance curve coincides with curve B to the left of $\mathbf{r}$, and with curve A to the right. Therefore, the irradiance at $\mathbf{r}$ has a discontinuity in the first derivative. Only contact occlusion and edge-edge alignments cause the Jacobian to be undefined; other types of events cause higher-order discontinuities in the vector irradiance, but are first-order smooth.

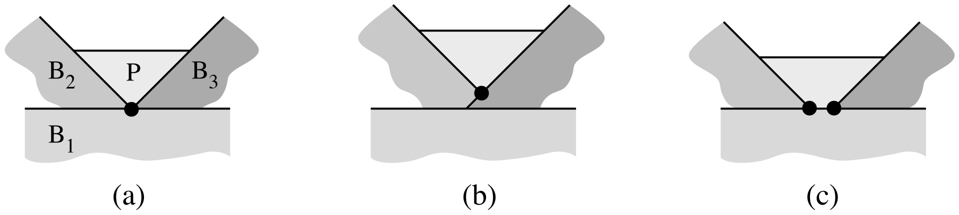

Figure 3.8: *(a) The vertex Jacobian does not exist at the intersection of three edges. A small change can produce (b) a single apparent vertex, or (c) two apparent vertices.*

From equation (3.18) it would appear that the irradiance Jacobian does not exist if any one of the vertex Jacobians fails to exist; this is not always so. A vertex Jacobian may be undefined because the vertex lies at the intersection of three edges, as shown in Figure 3.8a. In cases such as this, a minute change in $\mathbf{r}$ can lead to several possible configurations with different vertex Jacobians. See Figures 3.8b and 3.8c. However, the unoccluded area of the source still changes smoothly despite such a difficulty at a single vertex. To ensure that equation (3.18) is valid wherever Φ is differentiable, we simply restrict the edges that are used in computing the vertex Jacobians to those that actually bound the clipped source. Thus, in Figure 3.8, blocker B_1 is ignored until it makes its presence known by the addition of a new edge, as in Figure 3.8c.

The most basic property of the Jacobian matrix is its connection with directional derivatives. For any $\xi \in \mathcal{S}^2$, the directional derivative of Φ at $\mathbf{r}$ in the direction ξ is

$$D_\xi \Phi(\mathbf{r}) = \mathbf{J}_{\mathbf{r}}(\Phi)\, \xi. \tag{3.24}$$

Although directional derivatives of Φ may be approximated to second order with central differences, using the irradiance Jacobian has several advantages. First, all directional derivatives of $\Phi(\mathbf{r})$ are easily obtained from the irradiance Jacobian at $\mathbf{r}$, which requires a single global clipping operation. That is, sources need only be clipped against blockers once per position $\mathbf{r}$ when computing the irradiance Jacobian according to equation (3.18). In contrast, finite difference approximations require a minimum of two clipping operations to compute a single directional derivative. More importantly, directions of maximal change follow immediately from the Jacobian but require multiple finite differences to approximate.

A final property, which we build upon in the next section, is the connection with the rate of change of surface irradiance. Differentiating equation (2.18) with respect to position, we have

$$\nabla\phi = \Phi^{\mathrm{T}}\,\mathbf{J}(\mathbf{n}) + \mathbf{n}^{\mathrm{T}}\,\mathbf{J}(\Phi), \tag{3.25}$$

which associates the irradiance gradient with the irradiance Jacobian. Note that $\mathbf{J}(\mathbf{n})$ is the curvature tensor of the surface at each point $\mathbf{r} \in \mathcal{M}$. For planar surfaces $\mathbf{J}(\mathbf{n}) \equiv 0$, so equation (3.25) reduces to

$$\nabla\phi = \mathbf{n}^{\mathrm{T}}\,\mathbf{J}(\Phi), \tag{3.26}$$

which is the form we shall use to compute isolux contours on polygonal receivers. When evaluating equation (3.26) several optimizations are possible by distributing the vector multiplication across the terms of equation (3.18), which changes the summation of matrices into a summation of row vectors.

3.5 Applications of the Irradiance Jacobian

Thus far we have derived a closed-form expression for the irradiance Jacobian and described the geometrical computations required to evaluate it. By means of equation (3.18), the irradiance Jacobian can be computed analytically at any point $\mathbf{r}$ within a diffuse polyhedral environment illuminated by diffuse luminaires. The most difficult aspect of the computation is in determining the visible portions of the luminaires, although this is also a requirement for computing irradiance. The steps are summarized as follows.

```
Matrix IrradianceJacobian( Point r )
  Matrix J ← 0
  for each source $P$ with radiant exitance $M$
      begin
      $\widehat{P} \leftarrow P$ depth-clipped against all blockers, as seen from $\mathbf{r}$
      for each $i$: $\mathbf{J}_i \leftarrow$ vertex Jacobian for the $i$th vertex of $\widehat{P}$
      for each $i$: $\mathbf{E}_i \leftarrow$ edge matrix for the $i$th edge of $\widehat{P}$
      for each $i$: $\mathbf{C}_i \leftarrow$ corner matrix using $\mathbf{E}_{i-1}$ and $\mathbf{E}_i$
      $\mathbf{J} \leftarrow \mathbf{J} + \frac{M}{2\pi}$ (sum of all $\mathbf{C}_i\,\mathbf{J}_i$)
      end
  return J
```

Here the inner loops all refer to the vertices as seen from $\mathbf{r}$; pairs of vertices associated with invisible edges are counted as one. Gradients can then be computed using equation (3.25) or equation (3.26). The procedure above is a general-purpose tool with many applications, several of which are described in the remainder of this section.

3.5.1 Finding Local Extrema

The first application we examine is that of locating irradiance extrema on surfaces, which can be used in computing bounds on the transfer of energy between surfaces [92]. Given the availability of gradients, the most straightforward approach to locating a point of maximal irradiance is with an ascent method of the form

$$\mathbf{r}^{i+1} \equiv \mathbf{r}^i + \gamma_i\,[\,\mathbf{I} - \mathbf{n}(\mathbf{r})\,\mathbf{n}^{\mathrm{T}}(\mathbf{r})\,]\,\nabla\phi^{\mathrm{T}}(\mathbf{r}^i), \tag{3.27}$$

where $\mathbf{r}^0$ is a given starting point, and the factor γ_i is determined by a line search that insures progress is made toward the extremum. For example, the line search may simply halve γ_i until an increase in irradiance is achieved. The extremum has been found when no further progress can be made. Minima are found similarly. The principle drawbacks of this method are that it finds only local extrema, and convergence can be very slow when the irradiance function is flat. In the absence of a global method for locating all extrema, seed points near each of the relevant extrema must be supplied.

3.5.2 Direct Computation of Isolux Contours

Curves of constant irradiance over surfaces are known as *isolux contours* [163]. Within computer graphics, isolux contours have been used for depicting irradiance distributions [168,108], volume rendering [9], simplifying shading computations [33], and identifying *isolux areas*, which are regions of approximately constant illumination [39]. In computer vision isolux contours have been used to perform automatic image segmentation [90]. Because of their utility in visualizing the illumination of surfaces, illumination engineers had developed methods for plotting isolux contours by hand for one or two unoccluded sources long before the advent of the computer [163,67]. In this section we show how isolux contours can be computed directly using the irradiance Jacobian.

Every isolux contour on a surface $\mathcal{M}$ can be represented by a function of a single variable, $\mathbf{r} : [0, \infty) \to \mathcal{M}$, that satisfies

$$\phi(\mathbf{r}(s)) = \phi(\mathbf{r}(0)) \tag{3.28}$$

for all $s \geq 0$. To compute such a curve we construct a first-order ordinary differential equation (ODE) to which it is a solution, and solve the ODE numerically.

The direction of most rapid increase in $\phi(\mathbf{r})$ at a point $\mathbf{r} \in \mathcal{M}$ is given by the gradient $\nabla\phi(\mathbf{r})$, which generally does not lie in the tangent plane of the surface. The projection of the gradient onto the surface is a tangent vector that is orthogonal to the isolux curve passing through its origin. See Figure 3.9a. If the projected gradient is rotated by 90 degrees, we obtain a direction in which the irradiance remains constant to first order. The rotated gradients define a vector field whose

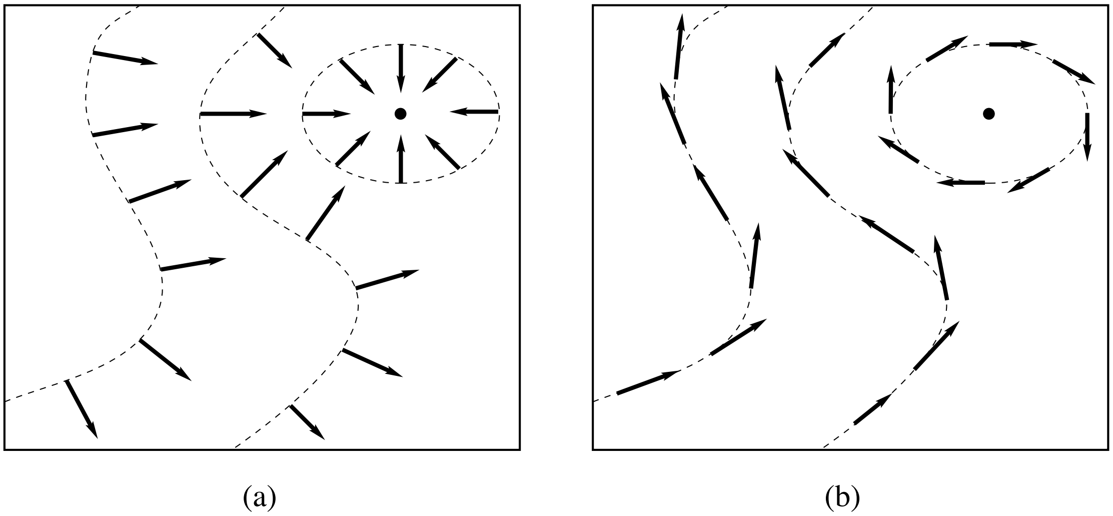

Figure 3.9: *(a) Projecting the gradient onto a surface defines a 2D vector field everywhere orthogonal to the level curves. (b) Rotating the projected gradients by $-\pi/2$ creates a vector field whose flow lines are isolux contours. Local maxima are then encircled by clockwise loops.*

flow lines are isolux contours. See Figure 3.9b. Combining these observations, we define the *isolux differential equation* by

$$\dot{\mathbf{r}} = \mathbf{P}(\mathbf{r})\, \nabla\phi^{\mathrm{T}}(\mathbf{r}), \tag{3.29}$$

with the initial condition $\mathbf{r}(0) = \mathbf{r}_0$, where

$$\mathbf{P}(\mathbf{r}) \equiv \mathbf{R}(\mathbf{n}(\mathbf{r}))\left[\mathbf{I} - \mathbf{n}(\mathbf{r})\;\mathbf{n}^{\mathrm{T}}(\mathbf{r})\right], \tag{3.30}$$

and $\mathbf{R}(z)$ is a rotation by $-\pi/2$ about the vector z. The matrix $\mathbf{P}(\mathbf{r})$ is constant for planar surfaces. The solution of this ODE is an isolux contour with irradiance $c = \phi(\mathbf{r}_0)$.

3.5.3 Solving the Isolux Differential Equation

Any technique for solving first-order ordinary differential equations can be applied to solving the isolux differential equation. The overriding consideration in selecting an appropriate method is the number of irradiance values and gradients used in taking a step along the curve. Obtaining this information involves a global clipping operation, which is generally the most expensive part of the algorithm.

Multistep methods are particularly appropriate for solving the isolux ODE since they make efficient use of the recent history of the curve. For example, Milne's predictor-corrector method is a multistep method that predicts the point $\mathbf{r}_{k+1} \equiv \mathbf{r}(s_{k+1})$ by extrapolating from the three most recent gradients and function values using a parabola. When the matrix $\mathbf{P}$ is fixed, Milne's predictor is given by

$$\mathbf{r}^0_{k+1} \equiv \mathbf{r}_{k-3} + \frac{4h}{3}\mathbf{P}\left(2g_{k-2} - g_{k-1} + 2g_k\right), \tag{3.31}$$

where g_k denotes the gradient at the point $\mathbf{r}_k$, and h is the step size [2]. Given the predicted value, a corrector is then invoked to find the nearest point on the curve. Because the contour is the zero set of the function $\phi(\mathbf{r}) - c$, the correction can be performed very efficiently using Newton's method. Beginning with the predicted point $\mathbf{r}^0_k$, a Newton corrector generates the sequence $\mathbf{r}^1_k, \mathbf{r}^2_k, \ldots$ by

$$\mathbf{r}^{i+1}_k \equiv \mathbf{r}^i_k + \left[c - \phi(\mathbf{r}^i_k)\right] \frac{\nabla\phi^{\mathrm{T}}(\mathbf{r}^i_k)}{||\,\nabla\phi^{\mathrm{T}}(\mathbf{r}^i_k)\,||^2}, \tag{3.32}$$

which converges quadratically to a point on the curve. The iteration is repeated until

$$\left|c - \phi(\mathbf{r}^i_k)\right| \leq \epsilon, \tag{3.33}$$

where ϵ is a preset tolerance. With this corrector, accurate polygonal approximations can be generated for arbitrarily long isolux contours. This would not be possible with the traditional Milne corrector, for example, which would eventually drift away from the curve. With a good predictor, very few correction steps are required, which saves costly gradient evaluations.

3.5.4 Examples of Isolux Contours

The predictor-corrector method described above was used to compute isolux contours for simple test cases with both unoccluded and partially occluded sources.

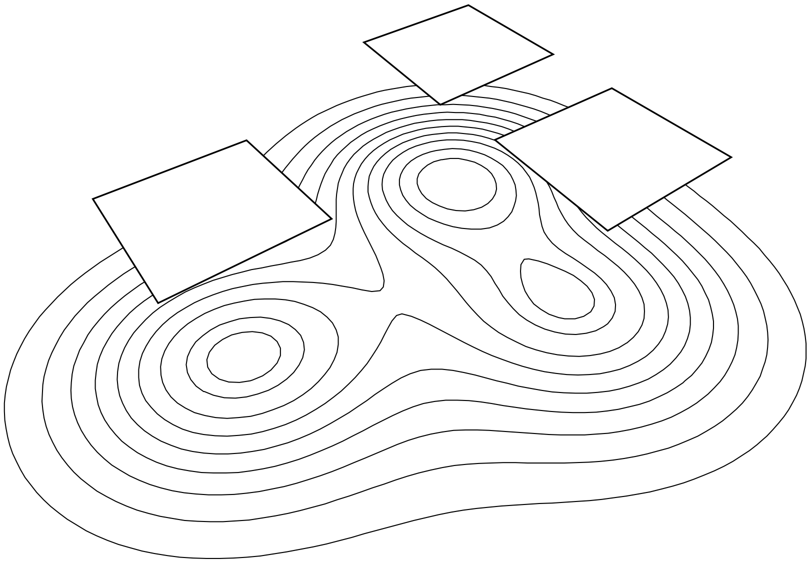

Figure 3.10: *A family of isolux contours for three unoccluded sources.*

The step size h and the tolerance for the corrector were user-supplied parameters. Use of the Newton corrector made the curve follower fairly robust; even abrupt turns at or near derivative discontinuities in the irradiance function were automatically compensated for.

To generate a family of curves depicting equal steps in irradiance, similar to a topographic map, we must find starting points for each curve with the desired irradiance values $c_1 > c_2 > \cdots > c_k$. The Newton corrector can be used to find a point on the $(k+1)$st curve by finding a root of the equation $\phi(\mathbf{r}) - c_{k+1}$ beginning at any point on the kth curve. The curve families in Figures 3.10 and 3.11 were automatically generated in this way. Figure 3.10 shows a family of isolux contours resulting from three unoccluded rectangular sources. Three distinct families were generated, starting at each of the three local maxima, which were found by the ascent method described in section 3.5.1. Figure 3.11 shows a family of isolux contours resulting from a rectangular source and a simple blocker. These contours surround both a peak and a valley.

Because distinct isolux contours cannot cross, any collection of closed contours has an obvious partial ordering defined by containment. To display filled contours, as shown in Figures 3.12a and 3.12b, the regions can be painted in back-to-front order after sorting according to the partial order.

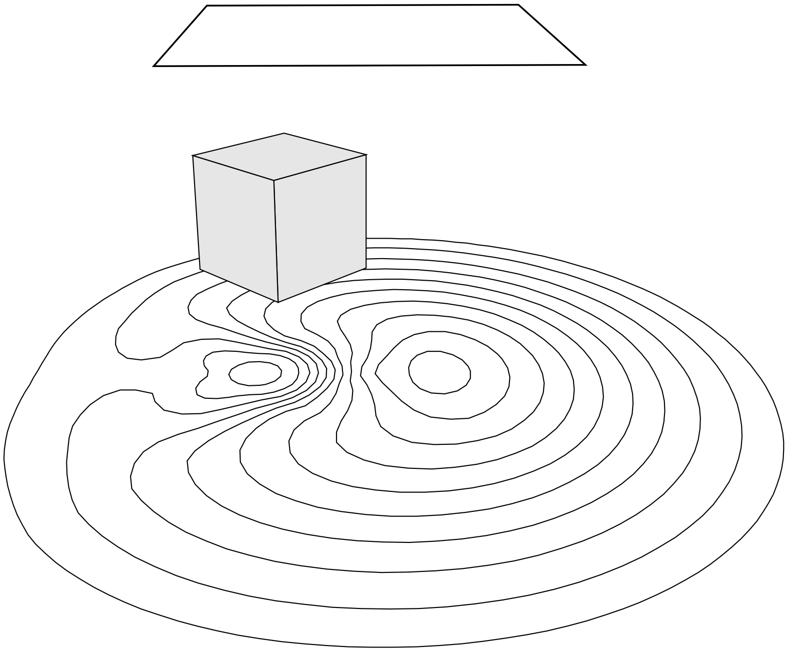

Figure 3.11: *Isolux contours on a planar receiver due to a rectangular source and simple blocker above the plane of the receiver.*

Because the isolux contours described in the previous section are generated by direct computation rather than by post-processing an image, they may be used in the image generation process. For example, isolux contours can be used to drive a meshing algorithm for global illumination.

The idea is similar to that of discontinuity meshing [64,94], which can identify important discontinuities in the radiance function over diffuse surfaces. Isolux contours provide additional information about radiance functions, and can be employed for mesh generation either in a preprocessing step for modeling direct illumination, or as part of a radiosity post-process to create a high-quality mesh for rendering a final image [94].

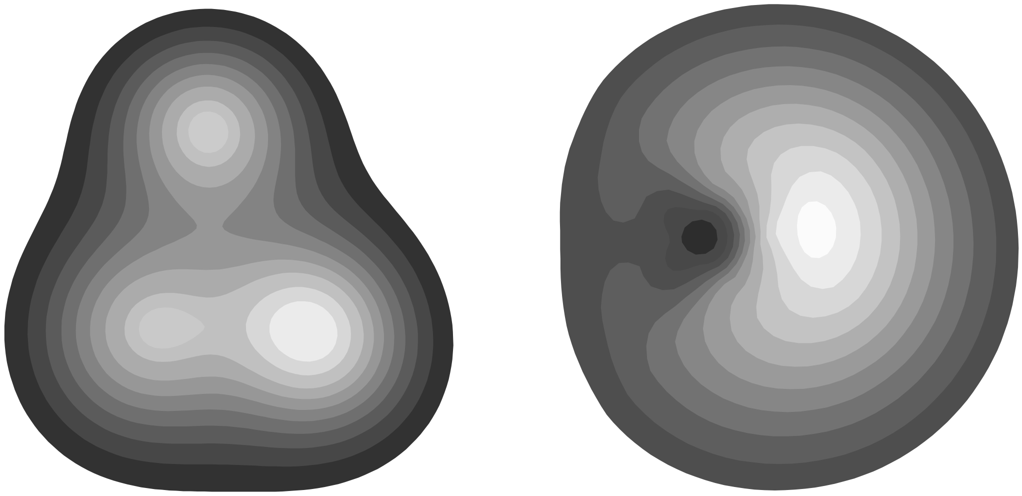

Figure 3.12: *Filled isolux contours corresponding to the previous figures. Each region is shaded according to the constant irradiance of its contour.*

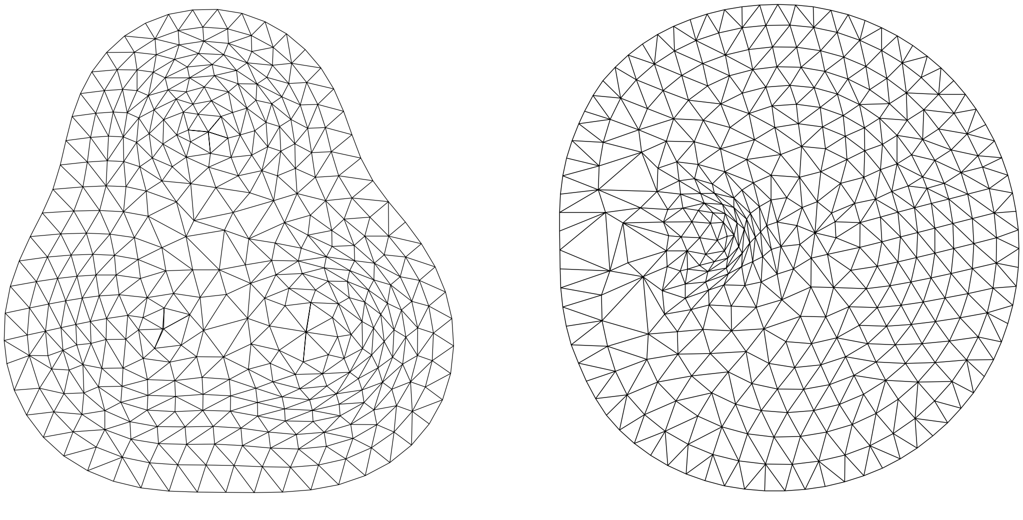

Figure 3.13: *Iso-meshes generated from families of isolux contours using constrained Delaunay triangulation.*

To best exploit the information in the contours, the mesh elements of an *isomesh* must follow the contours. Constrained Delaunay triangulation may be used to generate a mesh with this property from a sequence of segments that approximate isolux contours [27]. In earlier work, Lischinski et al. [94] used the same technique to generate meshes that incorporated irradiance discontinuities. In the present work, the constraints force the edges of the mesh elements to coincide with the isolux contours rather than crossing them. Another advantage of Delaunay triangulation is that it creates triangles with good aspect ratios. Figure 3.13 shows the result of applying this algorithm to the families of isolux contours shown in Figure 3.12. Meshes of varying coarseness can be generated by selecting subsets of the points along the contours.

Chapter 4

Irradiance Tensors

In this chapter we generalize the concept of irradiance to higher-order forms and derive an extended version of Lambert's formula that applies to non-diffuse surfaces. The generalization of irradiance consists of an infinite sequence of tensors that includes both the scalar and vector forms of irradiance as special cases. Each tensor in the sequence, called an *irradiance tensor*, consists of moments of a radiance distribution function $f(\mathbf{r}, \cdot)$. A large class of emission and scattering distributions can be characterized by combinations of these moments, which makes them useful in the simulation of non-diffuse phenomena. Applications include the computation of irradiance due to directionally-varying area light sources, reflections from glossy surfaces, and transmission through glossy surfaces.

The techniques developed in this chapter apply to any emission, reflection, or transmission distribution expressed as a polynomial over the unit sphere. As a concrete example, we derive expressions for a simple but versatile subset of these polynomial functions, called *axial moments*, which are closely related to the Phong reflection model [113]. Using the extended version of Lambert's formula, we derive closed-form expressions for moments of all orders in polygonal environments. Complete algorithms for efficient evaluation of the new expressions are presented and demonstrated by simulating Phong-like emission and scattering effects.

4.1 Non-Lambertian Phenomena

Rendering algorithms are frequently quite limited in the surface reflectance functions and luminaires they can accommodate, particularly when they are based on purely deterministic methods. To a large extent, this limitation stems from the difficulty of computing multi-dimensional integrals associated with non-diffuse phenomena, such as reflections from surfaces with directional scattering. While numerous closed-form expressions exist for computing the radiative exchange among uniform Lambertian (diffuse) surfaces with simple geometries [69,108,140], these expressions rarely apply in more general settings. Currently, the only approaches capable of simulating non-diffuse phenomena are those based on Monte Carlo [34, 147,173,174], hierarchical subdivision [12], or numerical quadrature [28,141,36].

Few methods exist at present for computing semi-coherent reflections of a scene in a nearly-specular or *glossy* surface. The earliest examples of glossy reflection in computer graphics are due to Amanatides [3] and Cook [34]. Amanatides used cone tracing to simulate glossy reflections for simple scene geometries and reflectance functions. Cook introduced a general Monte Carlo method for simulating such effects that was later extended to path tracing by Kajiya [71] and applied to realistic surfaces by Ward [173]. Wallace et al. [170] approximated Phong-like directional scattering by rendering through a narrow viewing aperture using a z-buffer. Aupperle et al. [12] devised the first general deterministic method using three-point transfers coupled with view-dependent hierarchical subdivision.

This chapter presents the first *analytic* method for computing direct lighting effects that involve both directional emission or scattering and area light sources: these include illumination from non-diffuse luminaires and reflections in non-diffuse surfaces. The method handles a wide range of emission and scattering distributions from ideal diffuse to sharply directional, which greatly extends the repertoire of effects that can be computed in closed form. As an example, the simulation shown in Figure 4.1 depicts a stained glass window design by Elsa Schmid [137] reflected in

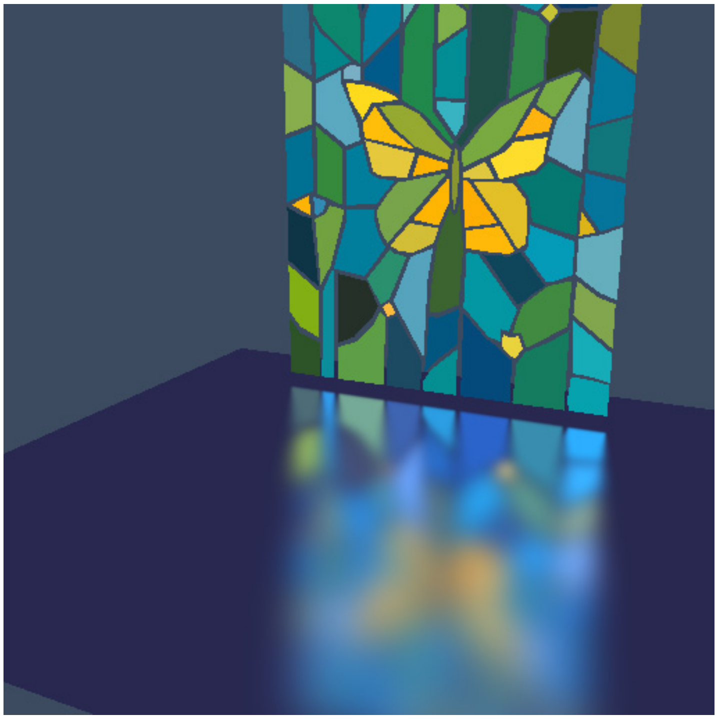

Figure 4.1: *A stained glass window reflected in a Phong-like glossy surface. The reflection was computed analytically at each pixel using a boundary integral for each luminaire. The expressions were derived using irradiance tensors.*

a nearly-specular surface with a Phong-like reflectance function. Reflections of this type are challenging to simulate using previous methods, but are straightforward using the techniques developed in this chapter.

The fundamental building block of this chapter is the *irradiance tensor*, a tensor representation of irradiance whose elements are *angular moments*, that is, weighted integrals of radiance with respect to direction [118]. Methods based on angular moments have a long history in the field of radiative transfer [84,145], but are applied here in a fundamentally different way. In classical radiative transfer problems only the low-order moments are relevant since detailed reflections that occur at surfaces can generally be neglected [145]. For image synthesis, however, where the light reflected from surfaces is paramount, high-order moments can be used to capture the appearance of a non-diffuse surface or the distribution of light emitted from a

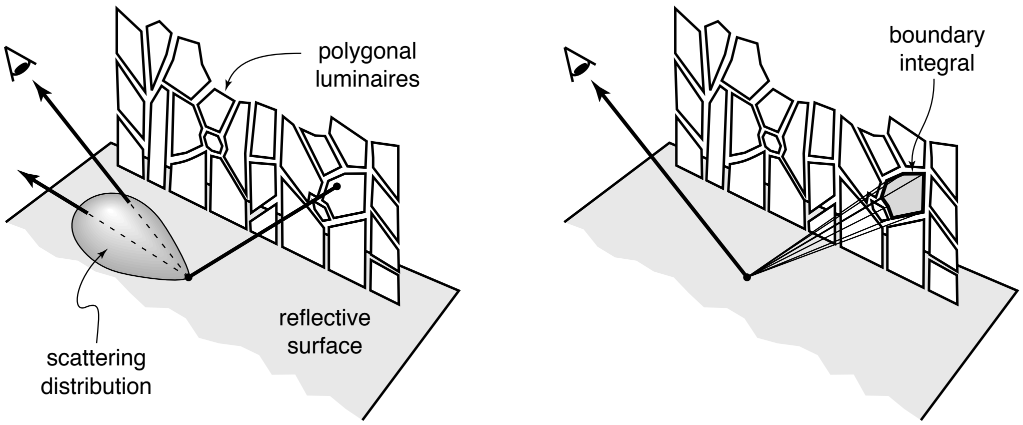

Figure 4.2: *Reflected radiance is determined by the scattering distribution integrated over each polygonal luminaire. Each surface integral can be replaced by a boundary integral.*

directional luminaire.

Angular moments are an extremely general tool for representing radiance distribution functions; for instance, they apply to all emission and reflectance functions that are defined in terms of polynomials over the sphere. However, the specific algorithms presented in this chapter address only a small class of these polynomials corresponding to Phong distributions [113]. These polynomials have a fundamental connection with irradiance tensors and admit practical closed-form expressions in polyhedral environments. The resulting closed-form expressions were used in computing the reflection in Figure 4.1, where the contribution of each polygonal window element was computed analytically at each point of the reflecting floor by first converting it to a boundary integral. See Figure 4.2.

The new expressions are computed in much the same way as Lamabert's formula for irradiance [108,140], which requires that we first compute the visible contours of each luminaire when occlusions are present. When applicable, the new expressions offer a number of advantages over previous methods, including efficient evaluation, exact answers (in the absence of roundoff error), relative ease of implementation, and symbolic evaluation of related expressions, such as derivatives. Finally, the statistical error, or *noise*, that accompanies Monte Carlo methods is eliminated.

The remainder of the chapter is organized as follows. Section 4.2 introduces basic concepts that motivate the definition of an irradiance tensor, which is further developed in section 4.3. These tensors are quite general, with applications in image synthesis well beyond those explored in this thesis. Section 4.4 describes an extension of Lambert's formula, based on irradiance tensors, that makes it applicable to non-diffuse environments. In section 4.5 we use the tensor version of Lambert's formula to derive expressions for *axial moments*, a convenient form of moment with immediate applications. In section 4.6 we focus on polygonal luminaires and derive the closed-form expressions and algorithms that are subsequently applied to three different non-diffuse simulations in section 4.7. Finally, section 4.8 describes two extensions: non-planar luminaires, and spatially varying luminaires.

4.2 Preliminaries

This section introduces the physical and mathematical concepts that motivate the definition of irradiance tensors and summarizes some of the tools from differential geometry that are needed to derive expressions for the new tensor quantities. We begin by describing a collection of tensors with immediate physical interpretations and then extend this collection to an infinite family. The members of the family are of interest because they embody higher-order moments of radiance distribution functions, which are useful in characterizing the directional variation of such functions.

4.2.1 Three Related Quantities

Most radiometric quantities can be defined in terms of weighted integrals of radiance. We shall examine three such quantities that lead naturally to irradiance tensors. First, the *monochromatic radiation energy density* [97] at the point $\mathbf{r} \in \mathrm{I\!R}^3$, denoted $u(\mathbf{r})$ and defined by

$$u(\mathbf{r}) \equiv \frac{1}{c} \int_{\mathcal{S}^2} f(\mathbf{r}, \mathbf{u}) \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{joules}}{\text{m}^3} \right], \tag{4.1}$$

is the radiant energy per unit volume at $\mathbf{r}$. Here c is the speed of light in the medium. A similar quantity known as the *scalar irradiance* [122, p. 39] is given by $c\, u(\mathbf{r})$, which has the units [watts/m^2] because of the constant. The vector irradiance at a point $\mathbf{r}$, which appeared in chapter 3, is defined by the closely related vector integral

$$\Phi(\mathbf{r}) \equiv \int_{\mathcal{S}^2} \mathbf{u} \, f(\mathbf{r}, \mathbf{u}) \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{watts}}{\text{m}^2} \right]. \tag{4.2}$$

The scalar quantity $\Phi(\mathbf{r}) \cdot \mathbf{v}$ is the net flow of radiant energy through a surface at $\mathbf{r}$ with normal $\mathbf{v}$ [122]. Finally, the *radiation pressure tensor* [118] at $\mathbf{r}$, denoted by $\Psi(\mathbf{r})$, is a symmetric 3×3 matrix found by integrating the outer product $\mathbf{u}\mathbf{u}^{\mathrm{T}}$:

$$\Psi(\mathbf{r}) \equiv \frac{1}{c} \int_{\mathcal{S}^2} \mathbf{u}\mathbf{u}^{\mathrm{T}} \, f(\mathbf{r}, \mathbf{u}) \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{joules}}{\text{m}^3} \right]. \tag{4.3}$$

The physical meaning of the bilinear form $\mathbf{w}^{\mathrm{T}} \Psi(\mathbf{r}) \mathbf{v}$ is the rate at which photon momentum in the direction $\mathbf{w}$ flows across a surface at $\mathbf{r}$ with normal $\mathbf{v}$. This tensor is exactly analogous to the stress tensor encountered in the theory of elasticity [98]. Thus, each of the above integrals has a different meaning and provides distinct information about the radiance distribution function at the point $\mathbf{r}$.

Note that in equations (4.2) and (4.3), the integral is in effect distributed across the elements of the vector or matrix. In equation (4.2) each element of the vector is a weighted integral of $f(\mathbf{r}, \cdot)$ with a weighting function that is a first-order monomial on the sphere; that is, one of the coordinate functions x, y, or z, where

$$\begin{aligned} x(\mathbf{u}) &\equiv \mathbf{e}_1 \cdot \mathbf{u}, \\ y(\mathbf{u}) &\equiv \mathbf{e}_2 \cdot \mathbf{u}, \\ z(\mathbf{u}) &\equiv \mathbf{e}_3 \cdot \mathbf{u}, \end{aligned}$$

for all $\mathbf{u} \in \mathcal{S}^2$, and $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the canonical basis for $\mathrm{I\!R}^3$. Alternatively, the functions x, y, and z can be viewed as the *direction cosines* of the vector $\mathbf{u}$. In equation (4.3) the weighting functions are the second-order monomials, x^2, y^2, z^2, xy, xz, and yz. Thus, the scalar-valued integrals embodied in equations (4.2) and (4.3) respectively correspond to first- and second-order angular moments of the radiance distribution function.

Although the three quantities defined above are distinct, they are also interrelated. For example, since $x^2 + y^2 + z^2 = 1$, it follows that

$$\text{trace } \Psi(\mathbf{r}) = u(\mathbf{r}) \tag{4.4}$$

for all $\mathbf{r} \in \mathrm{I\!R}^3$, which demonstrates a connection between the 2nd and 0th-order moments. Relationships of this nature also exist among higher-order analogues and will be an essential element in their construction.

4.2.2 Polynomials over the Sphere

While the integrals defining vector irradiance and the radiation presure tensor embody first- and second-order weighting functions respectively, we now motivate the extension of this idea to higher order monomials. First, we observe that polynomials over the sphere based on the coordinate functions x, y, and z can approximate a very large class of functions, in exact analogy with polynomials over the real line. Secondly, we show that the polynomials are well suited to approximating a class of functions that arise in simulating non-diffuse phenomena.

The general approximation property can be shown very easily using a non-constructive argument. Observe that any collection of functions over a common domain defines an associated vector consisting of the algebraic closure of the set under pointwise addition and scalar multiplication, that is

$$\begin{aligned} (\lambda f)(\mathbf{u}) &\equiv \lambda f(\mathbf{u}), \\ (f + g)(\mathbf{u}) &\equiv f(\mathbf{u}) + g(\mathbf{u}), \end{aligned}$$

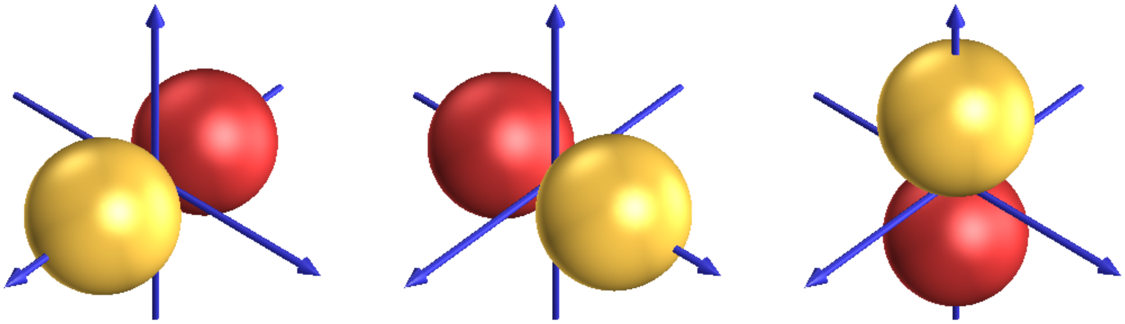

Figure 4.3: *The three direction cosines x, y, and z have graphs consisting of two spherical lobes; one positive and one negative. These functions generate an algebra of functions over the sphere that is dense in L_p.*

where λ is a scalar. The resulting collection may be further extended by introducing a third operation, pointwise multiplication, defined by

$$(f \cdot g)(\mathbf{u}) \;\; \equiv \;\; f(\mathbf{u})\, g(\mathbf{u}).$$

The collection of all functions obtained from the initial set by means of finite products and linear combinations forms an *algebra*. Let $\mathcal{A}$ denote the algebra generated by the real-valued functions x, y, and z defined on $\mathcal{S}^2$, which are depicted in Figure 4.3. The set $\mathcal{A}$, which is algebraically closed under the usual vector space operations augmented by pointwise multiplication, defines the polynomials over the sphere.

We now show that the elements of this algebra can approximate any L_p-function on the sphere with $p < \infty$; that is, any function in the space $L_p(\mathcal{S}^2, \sigma)$ consisting of all measurable functions $f : \mathcal{S}^2 \to \mathrm{I\!R}$ with respect to σ such that

$$\int_{\mathcal{S}^2} |f(\mathbf{u})|^p \; d\sigma(\mathbf{u}) \; < \; \infty. \tag{4.5}$$

That polynomials over $\mathcal{S}^2$ can uniformly approximate this large class of functions follows easily from the Stone-Weierstrass theorem [128, p. 174], stated below.

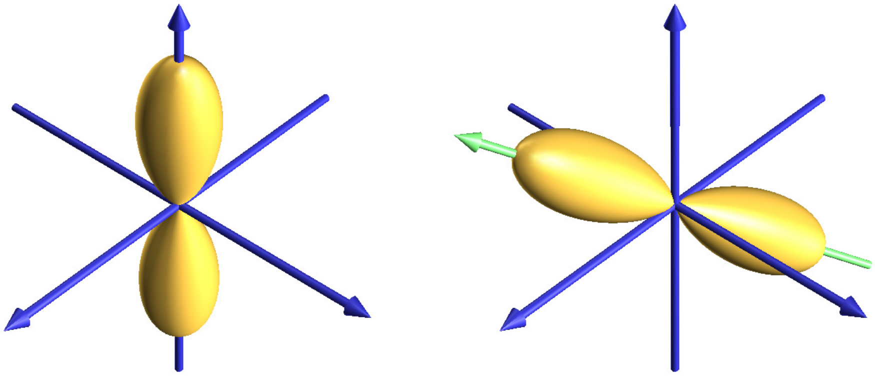

Figure 4.4: *(a) A cosine lobe defined by a 6th-order monomial, and (b) the same lobe about an arbitrary axis, which represents a 6th-order polynomial over the sphere.*

Theorem 5 (Stone-Weierstrass): *If X is a compact set and $\mathcal{N}$ is an algebra of continuous functions over X that separates the points of X and contains the constant functions, then $\mathcal{N}$ is uniformly dense in the continuous functions over X.*

A collection of functions over a set X is said to *separate* its points if for any two points x and y in X there exists a function f in the collection such that $f(x) \neq f(y)$. To see that the theorem applies to the algebra $\mathcal{A}$ of polynomials over the sphere, we observe that 1) the unit sphere is a compact set, 2) the algebra generated by x, y, and z contains the constant functions since $x^2 + y^2 + z^2 = 1$, and 3) the functions x, y, and z separate the points of $\mathcal{S}^2$ since distinct points must differ in at least one coordinate.

Finally, the approximation property extends to all L_p functions with $p < \infty$ because the set $C(\mathcal{S}^2)$ of continuous functions on $\mathcal{S}^2$ is dense in $L_p(\mathcal{S}^2, \sigma)$ with respect to the L_∞-norm [132, p. 71]. That is, any L_p function on the sphere may be approximated to within $\epsilon > 0$ at almost every point by a continuous function. It follows that $\mathcal{A}$ is dense in $L_p(\mathcal{S}^2, \sigma)$.

Of greater practical significance is the ease with which these polynomials can approximate certain functions that arise in the simulation of non-diffuse emission and scattering. As a simple example consider the monomial z^6, whose graph consists of two elongated lobes, as shown in Figure 4.4a. The same shape can be defined along any axis by a 6th-order polynomial; for this reason, the representation is sometimes referred to as *steerable* [106]. See Figure 4.4b. Scattering distributions can be approximated by superposing lobes along different axes and of different orders. Figure 4.5 shows a hypothetical BRDF constructed from three lobes. Directed lobes have other properties that make them a convenient tool for representing radiance distributions in general, as we show in section 4.5.

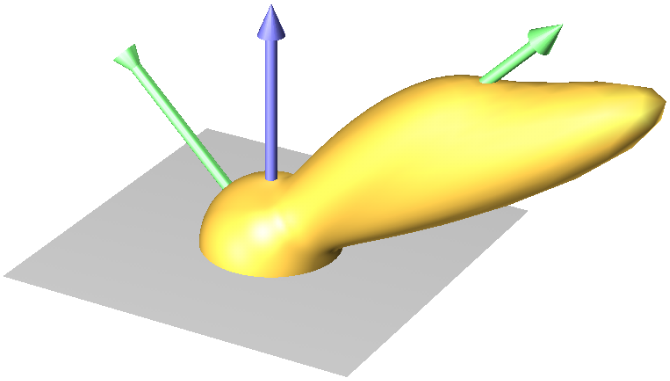

Figure 4.5: *A hypothetical bidirectional reflectance distribution function defined as a polynomial over the sphere; the polynomial is formed from the superposition of three directed lobes.*

In the remainder of this section we establish several basic facts about these integrals that will be useful in integrating polynomials over subsets of the sphere. In particular, we consider the *normalization function* for monomials over the sphere, which we define by

$$\eta(i,j,k) \equiv \int_{\mathcal{S}^2} x^i y^j z^k \, d\sigma(\mathbf{u}). \tag{4.6}$$

This function plays an important role in the development of irradiance tensors. For example, it is the key to the "odd-even" behavior of polynomials over the sphere; that is, a fundamental difference between monomials whose exponents are all even and those with at least one odd exponent. We now state the two central theorems concerning the function $\eta(i,j,k)$.

Theorem 6 *Let i, j, and k be non-negative integers. Then $\eta(i,j,k) = 0$ if and only if at least one of i, j, and k is odd.*

Proof: See Appendix A.1.

Where it is nonzero, the normalization function can be succinctly expressed in two very different forms. The first is in terms of the *double factorial*, defined by

$$n!! \equiv \begin{cases} 2 \cdot 4 \cdot 6 \cdots (n-2) \cdot n & \text{if } n \text{ is even} \\ 1 \cdot 3 \cdot 5 \cdots (n-2) \cdot n & \text{if } n \text{ is odd.} \end{cases}$$

The second representation of $\eta(i,j,k)$ involves the *multinomial coefficient*

$$\begin{pmatrix} & n & \\ i & j & k \end{pmatrix} \equiv \frac{n!}{i!\, j!\, k!}, \tag{4.7}$$

where $n = i + j + k$. This coefficient gives the number of distinct permutations among n objects of three distinguishable types; i of the first type, j of the second type, and k of the third type [44]. Using these definitions, we now state the second theorem.

Theorem 7 *Let i, j, and k be non-negative even integers. Then*

$$\eta(i,j,k) = \frac{(i-1)!!\,(j-1)!!\,(k-1)!!}{(n+1)!!} \tag{4.8}$$

$$= \frac{1}{n+1} \frac{\begin{pmatrix} & n' & \\ i' & j' & k' \end{pmatrix}}{\begin{pmatrix} & n & \\ i & j & k \end{pmatrix}}, \tag{4.9}$$

where $n = i + j + k$ and the primes denote division by 2.

Proof: See Appendix A.2

4.2.3 Elements of Differential Geometry

To study higher-order generalizations of irradiance we shall employ some basic machinery from differential geometry. In particular, we shall use the language of tensors and differential forms. This section summarizes the relevant facts that will be needed in the remainder of the chapter.

Differential forms are a formalism for representing integrands of multiple integrals. The algebraic properties of differential forms subsume the classical differential operators of divergence, gradient, and curl, and extend immediately to arbitrary dimensions. To emphasize the special behavior of multi-dimensional integrands, the classical notation $dy\, dx$ is replaced with the wedge product $dy \wedge dx$, which is referred to as a *2-form*; higher-order forms are constructed from multiple wedge products. The collection of algebraic rules associated with differential forms is known as the *exterior calculus*; it describes the interaction of the wedge product $\wedge$ and the *exterior derivative* operator d in terms of classical derivatives. The identities from the exterior calculus that we require are these:

$$dx \wedge dy = -dy \wedge dx \tag{4.10}$$

$$d(f\, d\beta) = df \wedge d\beta \tag{4.11}$$

$$df = \frac{\partial f}{\partial x}\, dx \;+\; \frac{\partial f}{\partial y}\, dy \;+\; \frac{\partial f}{\partial z}\, dz, \tag{4.12}$$

where f is a real-valued function defined on $\mathbb{R}^3$. Several intrinsic properties of differential forms are also important. In particular, a p-form β is said to be *closed* if $d\beta = 0$ and *exact* if there exists a $(p-1)$-form α such that $\beta = d\alpha$. The principle motivation behind the development of the exterior calculus was to determine when differential forms are exact. The concept of an exact differential form is crucial to the following development, and will further explain the “odd-even” behavior of polynomials over the sphere.

Tensors are another important tool in the study of multi-dimensional integration, and they will play a large role in what follows. The concept of a tensor may be defined in a number of ways; in the present work an nth-order tensor (or n-tensor) shall mean a multilinear functional defined on the n-fold cartesian product $V \times \cdots \times V$ of a vector space V [57]. In the following discussion, we shall only consider the case where V is the Euclidean space $\mathbb{R}^3$.

Tensors of any order can be formed using the tensor product operator, denoted by $\otimes$, which constructs bilinear mappings on the cartesian product of two vector spaces. If $\mathbf{A}$ and $\mathbf{B}$ are both linear functionals on $\mathbb{R}^3$, then $\mathbf{A} \otimes \mathbf{B}$ is a bilinear functional on $\mathbb{R}^3 \times \mathbb{R}^3$, or $\mathbb{R}^6$, defined by

$$(\mathbf{A} \otimes \mathbf{B})(x, y) \quad \equiv \quad \mathbf{A}(x) \cdot \mathbf{B}(y) \tag{4.13}$$

for all $x, y \in \mathbb{R}^3$. The tensor product operation is associative, so products of the form $\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}$ are unambiguous.

Every linear functional $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}$ is uniquely represented as an inner product with some vector, which in turn may be specified by its coordinates with respect to some basis. Thus, the 1-tensor $\mathbf{A}$ may be identified with its *cartesian components*, denoted by $\mathbf{A}_i$ for $i = 1, 2, 3$. The components of an n-tensor will be denoted using either n subscripts or a single multi-index consisting of n subscripts. The tensor product can therefore be written in component form as $(\mathbf{A} \otimes \mathbf{B})_{ij} \equiv \mathbf{A}_i \mathbf{B}_j$ for $1 \leq i, j \leq 3$, where $\mathbf{A}$ and $\mathbf{B}$ are both 1-tensors.

For notational convenience we shall adopt the summation convention, so that summation is implicit over every index occurring twice in a term. For example,

$$\mathbf{A}_{ij} \mathbf{B}_{jk} \quad \equiv \quad \sum_{j=1}^{3} \mathbf{A}_{ij} \mathbf{B}_{jk}$$

is a 2-tensor. This convention will not apply to sub-indices, that is, to indices of other indices. The tensors that we shall derive will depend largely on two fundamental tensors: the *Kronecker delta* and the *Levi-Civita symbol* (also called the *permutation* symbol), respectively defined by

$$\delta_{ij} \quad \equiv \quad \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

$$\varepsilon_{ijk} \equiv \begin{cases} 1 & \text{if } (i,j,k) \text{ is an even permutation} \\ -1 & \text{if } (i,j,k) \text{ is an odd permutation} \\ 0 & \text{if } (i,j,k) \text{ is not a permutation.} \end{cases}$$

In the derivations that follow we employ a number of useful relationships that exist among these tensors. The basic identities that we shall require are summarized below without proof.

$$\delta_{ij}\delta_{ij} = 3 \tag{4.14}$$

$$\varepsilon_{ijk}\,\varepsilon_{ljk} = 2\delta_{il} \tag{4.15}$$

$$\varepsilon_{pjl}\,\varepsilon_{kml} = \delta_{pk}\delta_{jm} - \delta_{pm}\delta_{jk} \tag{4.16}$$

$$\delta_{ij}\,\mathbf{A}_i\mathbf{B}_j = \mathbf{A}\cdot\mathbf{B} \tag{4.17}$$

$$\varepsilon_{ijk}\,\mathbf{A}_j\mathbf{B}_k = (\mathbf{A}\times\mathbf{B})_i \tag{4.18}$$

Here · and × denote the standard dot and cross products, respectively. Finally, we require the generalized Stokes' theorem, which may be written

$$\int_A d\alpha = \int_{\partial A} \alpha, \tag{4.19}$$

where A is a subset of $\mathbb{R}^p$, or more generally, a p-manifold in $\mathbb{R}^{p+k}$, and α is a $(p-1)$-form. Here ∂A denotes the boundary of A. Equation (4.19) is an extension of the fundamental theorem of calculus, which subsumes both the classical Stokes' theorem and Gauss's theorem (see, for example, Bishop and Goldberg [21] or Spivak [156]). The generalized Stokes' theorem is the tool by which we shall reduce the area integrals corresponding to irradiance tensors into boundary integrals, and in some instances obtain closed-form expressions.

4.3 Generalizing Irradiance

The physical quantities defined in the previous section can be extended to higher-order forms using the formalism of tensors. Doing so will provide a means of characterizing radiance distributions in terms of high-order moments.

4.3.1 Definition of Irradiance Tensors

Aside from the constant $1/c$, which has the units [sec/m], all of the integrals described in section 4.2.1 are multilinear functionals of the form

$$\int_{\mathcal{S}^2} \mathbf{u} \otimes \cdots \otimes \mathbf{u} \; f(\mathbf{r}, \mathbf{u}) \; d\sigma(\mathbf{u}) \qquad \left[\frac{\text{watts}}{\text{m}^2} \right], \tag{4.20}$$

where $\otimes$ denotes a tensor product; that is, each of the integrals defines a mapping from $\mathbb{R}^2 \times \cdots \times \mathbb{R}^3$ to $\mathbb{R}$ that is linear in each of the vector arguments independently. Viewed in this way, radiation energy density, vector irradiance, and the radiation pressure tensor are tensors of order 0, 1, and 2 respectively, with equation (4.20) providing the natural extension to tensors of all higher orders. When the function f represents radiance, the family of integrals in equation (4.20) subsumes the notion of vector irradiance, and each member possesses the units of irradiance [watts/m^2]. We therefore refer to this family of multilinear functionals as *irradiance tensors*. Although high-order irradiance tensors do not have immediate physical interpretations [118, p. 5], they are nevertheless useful vehicles for integrating polynomial functions over the sphere, such as those representing emission or scattering distributions.

In this work we shall restrict $f(\mathbf{r}, \cdot)$ to be piecewise constant, or more generally, piecewise polynomial over the sphere. Angular moments of f then reduce to polynomials integrated over regions of the sphere. To concisely represent these integrals we introduce a simplified form of irradiance tensor defined by

$$\mathbf{T}^n(A) \equiv \int_A \underbrace{\mathbf{u} \otimes \cdots \otimes \mathbf{u}}_{n \text{ factors}} \; d\sigma(\mathbf{u}), \tag{4.21}$$

where $A \subset \mathcal{S}^2$ and $n \geq 0$ is an integer. The integrand of such a tensor contains all monomials of the form $x^i y^j z^k$ where $(x, y, z) \in \mathcal{S}^2$, and $i + j + k = n$; thus, $\mathbf{T}^n(A)$ consists of nth-order monomials integrated over A.

In what follows, the region $A \subset \mathcal{S}^2$ will represent the spherical projection of a luminaire $P \subset \mathbb{R}^3$. Using the notation of chapter 1, we shall write $A = \mathbf{\Pi}(P)$

where again, we assume without loss of generality that the origin is the center of projection. The tensors $\mathbf{T}^n(A)$ allow us to perform several useful computations; for example, we may compute angular moments of the illumination due to uniform Lambertian luminaires, or the irradiance due to directionally varying luminaires.

Irradiance tensors of all orders are defined by surface integrals, and in some instances may be reduced to one-dimensional integrals by means of the generalized Stokes' theorem. The resulting boundary integrals can be evaluated analytically for certain patch geometries, such as polygons. The foregoing approach extends the classical derivation of Lambert's formula for irradiance [100,155], yielding more complex boundary integrals in the case of higher-order tensors.

4.3.2 The Form Differential Equation

In this section we investigate some of the properties of the tensor $\mathbf{T}^n(A)$ and show that it may be expressed in terms of a boundary integral and a term involving the solid angle of A, which is $\sigma(A)$.

One of the fundamental tools of this chapter is a representation of solid angle in terms of a surface integral. If each ray through the origin meets a surface $P \subset \mathbb{R}^3$ at most once, then the solid angle subtended by P is

$$\sigma(\mathbf{\Pi}(P)) = \left| \int_P d\omega \right|, \tag{4.22}$$

where the 2-form $d\omega$ is given by

$$d\omega \equiv -\frac{x(dy \wedge dz) + y(dz \wedge dx) + z(dx \wedge dy)}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \tag{4.23}$$

over the domain $\mathbb{R}^3 - \{0\}$. Here x, y, and z denote coordinate charts over the manifold P. For the following development it is convenient to omit the absolute value signs in equation (4.22), thus allowing the solid angle to be signed according to the orientation of the surface. The negative sign in equation (4.23) is included so that solid angle will be positive for surfaces whose orientation is positive or

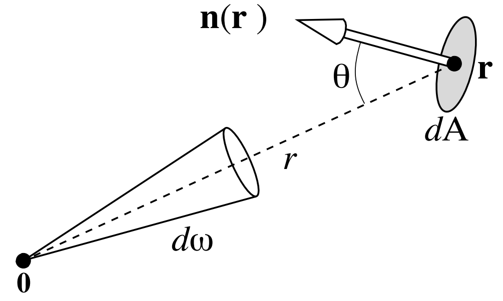

Figure 4.6: *The geometrical relation between the 2-forms dω and dA, interpreting them as infinitesimal quantities.*

counterclockwise with respect to the origin. Note that the definition of the 2-form $d\omega$ corresponds to the familiar change of variable

$$d\omega = \frac{\cos\theta}{r^2}\, dA, \tag{4.24}$$

where $d\omega$ and dA are interpreted as infinitesimal solid angle and infinitesimal area, respectively, and r and θ are as shown in Figure 4.6. More formally, equation (4.24) is the pullback of the solid angle 2-form to the volume element dA of the surface.

The connection between the measure σ and the 2-form $d\omega$ is simply

$$\sigma(A) \;=\; \int_A d\sigma(\mathbf{u}) \;=\; \int_A d\omega, \tag{4.25}$$

where $A \subset \mathcal{S}^2$ and A is positively oriented. We shall always assume that A is sufficiently well-behaved that integrals of both forms are defined; that is, A is both measurable and rectifiable [130]. Which representation we use will depend on the context. In general, we adopt the language of differential forms $d\omega$ for computations that involve the integrand, and the measure-theoretic notation when the emphasis is on the integral as a whole, either as a function or as a transformation.

Our strategy for obtaining closed-form expressions for each of the irradiance tensors is to first convert the surface integrals into integrals over the boundaries. The resulting boundary integrals can then be expressed in closed-form for simple geometries, such as polygons. This approach is analogous to classical derivations of formulas for irradiance [43,182,155]. To extend the approach to more general settings, we seek a 1-form $\boldsymbol{\alpha}$, which is itself an nth-order tensor, such that

$$\int_{\partial A} \boldsymbol{\alpha} = \int_A \mathbf{u}^n \; d\omega, \tag{4.26}$$

where $\mathbf{u} \equiv \mathbf{r} / \,||\, \mathbf{r} \,||$, and $\mathbf{u}^n$ denotes the n-fold tensor product of $\mathbf{u}$ [54]; that is

$$\mathbf{u}^n \equiv \underbrace{\mathbf{u} \otimes \cdots \otimes \mathbf{u}}_{n \text{ terms}}.$$

Thus, each element of $\mathbf{u}^n$ is a product of n direction cosines:

$$\mathbf{u}^n_{\mathrm{I}} \equiv \mathbf{u}_{i_1} \mathbf{u}_{i_2} \cdots \mathbf{u}_{i_n}.$$

The relationship between the differential forms associated with the two integrals in equation (4.26) is provided by the generalized Stokes' theorem, from which it follows that

$$d\boldsymbol{\alpha} = \; \mathbf{u}^n \; d\omega. \tag{4.27}$$

In terms of individual components, equation (4.27) is equivalent to the 3^n equations

$$d\boldsymbol{\alpha}_{i_1 \cdots i_n} = \mathbf{u}_{i_1} \cdots \mathbf{u}_{i_n} d\omega, \tag{4.28}$$

where $1 \leq i_k \leq 3$ for $k = 1, 2, \cdots n$. As we shall see, in some instances no 1-form $\boldsymbol{\alpha}$ satisfies the above equation; consequently, a slight modification of equation (4.28) will be required in order to integrate $\mathbf{u}^n \, d\omega$ over A. Because the product of the n scalars on the right of equation (4.28) commute, the tensor $\boldsymbol{\alpha}$ possesses a large number of symmetries. Of its 3^n elements, only

$$\begin{pmatrix} n+2 \\ 2 \end{pmatrix} = \frac{(n+2)(n+1)}{2} \tag{4.29}$$

are unique. Equation (4.29) gives the number of ways to distribute n indistinguishable objects (exponents) among three bins (x, y, and z).

Now we shall consider the question of whether there exists a 1-form $\boldsymbol{\alpha}$ that satisfies equation (4.28), which is equivalent to determining when the 2-form

$\mathbf{u}_{i_1} \cdots \mathbf{u}_{i_n}\, d\omega$ is exact. We first observe that if β is an exact 2-form defined on $\mathcal{S}^2$, then

$$\int_{\mathcal{S}^2} \beta = 0.$$

This is a direct consequence of Stokes' theorem, for if β is exact, then $\beta = d\gamma$ for some 1-form γ defined on $\mathcal{S}^2$. It follows that

$$\int_{\mathcal{S}^2} \beta \;=\; \int_{\partial\mathcal{S}^2} \gamma \;=\; \int_{\emptyset} \gamma \;=\; 0,$$

since $\mathcal{S}^2$ has no boundary. An important implication of this fact is that the 2-form $d\omega$ is not exact, since

$$\int_{\mathcal{S}^2} d\omega = 4\pi, \tag{4.30}$$

which is the surface area of the unit sphere. Hence, no 1-form $\boldsymbol{\alpha}$ can satisfy equation (4.28) when $n = 0$, which implies that the solid angle subtended by A cannot be expressed as an integral over its outer contour as seen from the origin. Nevertheless, it is common practice to denote the 2-form for solid angle by $d\omega$, with the apparent implication that it is the differential of some 1-form ω [156].

Because $d\omega$ is not exact we require a direct means of computing $\mathbf{T}^0(A)$, which corresponds to solid angle, since it cannot be reduced to a boundary integral. When A is a polygon, this is easily accomplished using an elementary result of spherical trigonometry [19], as shown below.

Theorem 6 implies that the 2-form $x^i y^j z^k\, d\omega$ cannot be exact if i, j, and k are all even, which indicates that the complication noted above for computing solid angle must also arise with all the other even-order tensors. Fortunately, the general case is no more difficult to handle than the 0th-order case. To see this, we add a constant to each monomial that forces equation (4.30) to hold, which is a necessary condition for exactness. That is, we introduce constants c_{ijk} such that

$$\int_{\mathcal{S}^2} \left[x^i y^j z^k - c_{ijk} \right] \, d\omega = 0 \tag{4.31}$$

for all i, j, and k. Clearly $c_{ijk} = \eta(i,j,k)$, so equation (4.28) may be replaced by

$$d\boldsymbol{\alpha}_{i_1 \cdots i_n} = \left[x^i y^j z^k - \eta(i,j,k)\right] d\omega, \tag{4.32}$$

where i, j, and k are the number of occurrences of the indices 1, 2, and 3 respectively in $i_1, i_2, \ldots, i_n$. To show that the new 2-form on the right of equation (4.32) is now exact, by De Rham's theorem [42, p. 67] it suffices in this instance to show that it is closed. To show that the right hand side of equation (4.32) is closed, we show that all expressions of the form $x^i y^j z^k \, d\omega$ are closed. See Appendix A.4. Therefore, we are guaranteed that a 1-form $\boldsymbol{\alpha}$ satisfying equation (4.32) exists in all cases. To express equation (4.32) in tensor form, we define

$$\Xi^n_{\mathrm{I}} \;\equiv\; \eta(\kappa^1_{\mathrm{I}}, \kappa^2_{\mathrm{I}}, \kappa^3_{\mathrm{I}}), \tag{4.33}$$

where I is the multi-index $(i_1, \ldots, i_n)$ and κ^i_{I} denotes the number of occurrences of the index i in I. Equation (4.32) can now be written as

$$d\boldsymbol{\alpha} = [\mathbf{u}^n - \Xi^n]\, d\omega, \tag{4.34}$$

which is the fundamental equation for $\boldsymbol{\alpha}$. This equation is an example of a *form differential equation*, since the unknown is the differential form $\boldsymbol{\alpha}$ [139, p. 95]. Given a 1-form $\boldsymbol{\alpha}$ satisfying equation (4.34), the n-tensor $\mathbf{T}^n(A)$ is given by

$$\mathbf{T}^n(A) \;=\; \Xi^n\, \sigma(A) \;+\; \int_{\partial A} \boldsymbol{\alpha}. \tag{4.35}$$

Thus, the tensor Ξ^n is a component of the final expression for $\mathbf{T}^n(A)$, which makes explicit the role of solid angle in all even-order tensors. Equation (4.35) is the generalization of Lambert's formula that we develop further and apply in the remainder of this chapter.

To make equation (4.35) more concrete, the following theorem provides an expression for Ξ^n in terms of elementary tensors.

Theorem 8 *When n is odd, $\Xi^n = \mathbf{0}$. When n is even*

$$\Xi^n_{\mathrm{I}} = \frac{1}{(n+1)!} \sum_{\mathrm{J} \in S(\mathrm{I})} \delta_{j_1 j_2} \cdots \delta_{j_{n-1} j_n}, \tag{4.36}$$

where $S(\mathrm{I})$ is the set of all permutations of the multi-index $\mathrm{I} = (i_1, \ldots, i_n)$.

Proof: See Appendix A.3.

Listed below are the first few cases of equation (4.36), which have been simplified by combining equivalent terms.

$$\begin{aligned}
\Xi^0 &= 1 \\
\Xi^2 &= \frac{\delta_{ij}}{3} \\
\Xi^4 &= \frac{\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{15} \\
\Xi^6 &= (\ \delta_{ij}\delta_{kl}\delta_{pq} + \delta_{ij}\delta_{kp}\delta_{lq} + \delta_{ij}\delta_{kq}\delta_{pl} + \delta_{ik}\delta_{jl}\delta_{pq} + \delta_{ik}\delta_{jp}\delta_{lq} \\
&\quad + \delta_{ik}\delta_{jq}\delta_{pl} + \delta_{il}\delta_{kj}\delta_{pq} + \delta_{il}\delta_{kp}\delta_{jq} + \delta_{il}\delta_{kq}\delta_{pj} + \delta_{ip}\delta_{kl}\delta_{jq} \\
&\quad + \delta_{ip}\delta_{kj}\delta_{lq} + \delta_{ip}\delta_{kq}\delta_{jl} + \delta_{iq}\delta_{kl}\delta_{pj} + \delta_{iq}\delta_{kp}\delta_{lj} + \delta_{iq}\delta_{kj}\delta_{pl}\ )\ /\ 105.
\end{aligned}$$

While many of the terms of the expansion in equation (4.36) can be combined, the number of distinct terms nevertheless grows rapidly, as is already evident for $n = 6$. The number of terms remaining after simplification is the number of distinct groupings of the n indices into pairs. To count them, consider the number of ways to complete a sequence of $n/2$ pairs, removing the indices one at a time without replacement; this number is $(n-1)!!$. Consequently, equation (4.36) is useful for symbolic purposes, but quickly becomes impractical for computation as n increases. On the other hand, as n increases, Ξ^n becomes increasingly sparse, vanishing completely whenever n is odd. When n is even, only

$$\binom{2 + n/2}{2} = \frac{(n+4)(n+2)}{8} = \mathrm{O}(n^2) \tag{4.37}$$

of the 3^n elements are non-zero. This is the number of ways to distribute the $n/2$ powers of 2 among x, y, and z, or equivalently, the number of distinct arrangements of $n/2$ objects and two partitions.

4.4 Extending Lambert's Formula

In this section we address the problem of finding the 1-form $\boldsymbol{\alpha}$ that satisfies equation (4.34). The derivation will be completely general, applying to irradiance tensors of all orders. We shall show that the higher-order tensors can be conveniently expressed by means of a recurrence relation of the form

$$\mathbf{T}^n = G_n(\mathbf{T}^{n-2}) + H_n(\partial A), \tag{4.38}$$

with $\mathbf{T}^{-1} \equiv 0$, and $\mathbf{T}^0(A) \equiv \sigma(A)$. Thus, each tensor of order 1 or higher can be constructed from two components: a function of the boundary and a function of a lower-order tensor. We shall start by examining the second-order tensor, as it will illustrate some of the steps that also apply in the general case.

4.4.1 The Radiation Pressure Tensor

We first demonstrate the approach by deriving an expression for the second-order tensor $\mathbf{T}^2(A)$, which is proportional to the radiation pressure tensor defined on page 62. We then derive a closed-form expression for this tensor when A is a spherical polygon.

First, let $A \subset \mathcal{S}^2$ be an arbitrary region. It follows from equation (4.34) and theorem 8 that the 1-form $\boldsymbol{\alpha}$ corresponding to $\mathbf{T}^2(A)$ is a 3×3 matrix such that

$$d\boldsymbol{\alpha} = (\mathbf{u}\mathbf{u}^{\mathrm{T}} - \tfrac{1}{3}\mathbf{I})\, d\omega. \tag{4.39}$$

This is the form differential equation associated with the radiation pressure tensor. Letting $\mathbf{r}$ denote an arbitrary point on the surface $P \subset \mathrm{I\!R}^3$, we may rewrite 2-form

$d\omega$, defined in equation (4.23), as

$$d\omega \equiv -\frac{\varepsilon_{qlm}\,\mathbf{r}_q\,d\mathbf{r}_l \wedge d\mathbf{r}_m}{2\,||\,\mathbf{r}\,||^3}, \tag{4.40}$$

where $||\cdot||$ is the Euclidean norm. Henceforth, we denote the scalar $||\,\mathbf{r}\,||$ by r. From equations (4.39) and (4.40) it follows that $d\boldsymbol{\alpha}$ may be written

$$d\boldsymbol{\alpha}_{ij} = \mathbf{Q}_{ijlm}\,d\mathbf{r}_l \wedge d\mathbf{r}_m, \tag{4.41}$$

where we have introduced the 4-tensor $\mathbf{Q}_{ijlm}$ defined by

$$\mathbf{Q}_{ijlm} \equiv \frac{\left(\delta_{ij}\,r^2 - 3\mathbf{r}_i\mathbf{r}_j\right)\varepsilon_{qlm}\mathbf{r}_q}{6r^5}. \tag{4.42}$$

For each i and j, the 1-form $\boldsymbol{\alpha}_{ij}$ can be written as a linear combination of basic 1-forms, which implies that there exists a 3-tensor $\mathbf{A}$ such that

$$\boldsymbol{\alpha}_{ij} = \mathbf{A}_{ijl}\,d\mathbf{r}_l. \tag{4.43}$$

Differentiating both sides of equation (4.43) and simplifying, we have

$$d\boldsymbol{\alpha}_{ij} = \mathbf{A}_{ijl,m}\,d\mathbf{r}_m \wedge d\mathbf{r}_l, \tag{4.44}$$

where the index following the comma indicates a partial derivative. Therefore, the problem of finding $\boldsymbol{\alpha}$ reduces to finding a 3-tensor $\mathbf{A}_{ijl}$ that satisfies

$$\mathbf{A}_{ijl,m}\,d\mathbf{r}_m \wedge d\mathbf{r}_l = \mathbf{Q}_{ijlm}\,d\mathbf{r}_l \wedge d\mathbf{r}_m \tag{4.45}$$

for all i and j in $\{1, 2, 3\}$. Since $\mathbf{r}_i \wedge \mathbf{r}_j = -\mathbf{r}_j \wedge \mathbf{r}_i$, equation (4.45) is equivalent to

$$\varepsilon_{kml}\,\mathbf{A}_{ijl,m} = \varepsilon_{klm}\,\mathbf{Q}_{ijlm}, \tag{4.46}$$

for all i, j, and k in $\{1, 2, 3\}$. Note that the transposition of $d\mathbf{r}_l$ and $d\mathbf{r}_m$ is accounted for by transposing the indices of one of the ε factors. Substituting equation (4.42) into the above equation and using identity (4.15), equation (4.46) becomes

$$\varepsilon_{kml}\mathbf{A}_{ijl,m} = \frac{\left(\delta_{ij}\,r^2 - 3\mathbf{r}_i\mathbf{r}_j\right)\mathbf{r}_k}{3r^5} \tag{4.47}$$

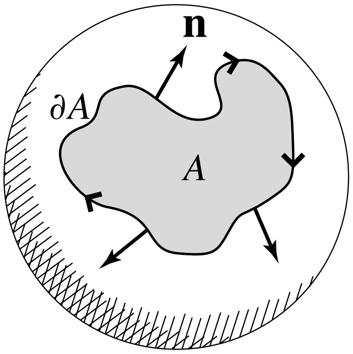

Figure 4.7: *For any region $A \subset \mathcal{S}^2$, the unit vector $\mathbf{n}$ is normal to the boundary ∂A, directed outward, and tangent to the sphere.*

for all i, j, and k in $\{1, 2, 3\}$. This is the fundamental equation associated with the tensor $\mathbf{T}^2$; given any 3-tensor $\mathbf{A}$ satisfying equation (4.47), we can then express the tensor $\mathbf{T}^2(A)$ in terms of a boundary integral and a term involving surface area. In particular, it follows from equation (4.35) that

$$\mathbf{T}^2_{ij}(A) = \frac{1}{3}\,\delta_{ij}\,\sigma(A) \;+\; \int_{\partial A} \mathbf{A}_{ijl}\, d\mathbf{r}_l. \tag{4.48}$$

The following theorem provides a formula for $\mathbf{T}^2(A)$, in which the 3-tensor $\mathbf{A}$ appears as an expression involving the outward normal to the boundary curve ∂A, as shown in Figure 4.7. We then find a closed-form expression for the boundary integral appearing on the right of equation (4.48) when A is a spherical polygon.

Theorem 9 *For any $A \subset \mathcal{S}^2$ the 2-tensor $\mathbf{T}^2(A)$ is given by*

$$\mathbf{T}^2(A) \;=\; \mathbf{I}\,\frac{\sigma(A)}{3} \;-\; \frac{1}{3}\int_{\partial A} \mathbf{u}\,\mathbf{n}^{\mathrm{T}}\, ds, \tag{4.49}$$

where ds denotes integration with respect to arclength and $\mathbf{n}$ is the outward normal to the curve ∂A.

<u>**Proof:**</u> We need only show that if $-\mathbf{u}\,\mathbf{n}^{\mathrm{T}}\, ds = \mathbf{A}_{ijl}\, d\mathbf{r}_l$, then $\mathbf{A}_{ijl}$ satisfies equation (4.47). We proceed by expressing the integral in equation (4.49) in terms of

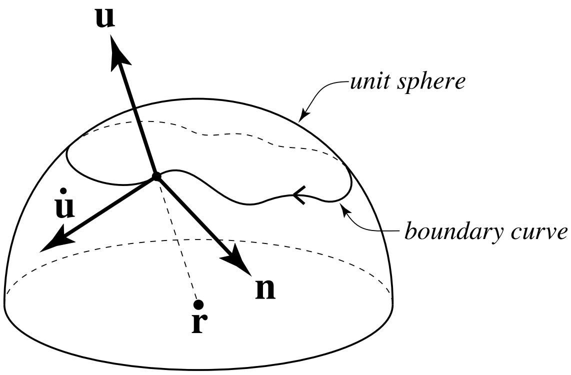

Figure 4.8: *The outward normal to the boundary curve is obtained from* **u** *and its derivative with respect to arclength, denoted by* $\dot{\mathbf{u}}$*.*

the position vector **r** and its derivative. From Figure 4.8 we see that $\mathbf{n} = \mathbf{u} \times \dot{\mathbf{u}}$, where $\dot{\mathbf{u}}$ denotes the derivative with respect to arclength $d\mathbf{u}/ds$. Therefore,

$$\begin{aligned} \mathbf{n}\, ds &= \mathbf{u} \times d\mathbf{u} \\ &= \frac{\mathbf{r}}{r} \times \frac{(\mathbf{I} - \mathbf{u}\mathbf{u}^{\mathrm{T}})}{r}\, d\mathbf{r} \\ &= \frac{\mathbf{r} \times d\mathbf{r}}{r^2}. \end{aligned} \tag{4.50}$$

It follows that we may replace $-\mathbf{u}\,\mathbf{n}^{\mathrm{T}}\, ds$ with the expression $\mathbf{A}_{ijl}\, d\mathbf{r}_l$ where

$$\mathbf{A}_{ijl} \equiv -\frac{\varepsilon_{jpl}\,\mathbf{r}_p\,\mathbf{r}_i}{3r^3}. \tag{4.51}$$

To show that the above expression satisfies equation (4.47), we compute its partial derivative and multiply by ε_{kml}. First, note that

$$\frac{\partial}{\partial\, \mathbf{r}_m}\left(\frac{1}{||\,\mathbf{r}\,||^k}\right) = -k\frac{\mathbf{r}_m}{||\,\mathbf{r}\,||^{k+2}}, \tag{4.52}$$

for any $n \geq 1$. Thus, the partial of $\mathbf{A}_{ijl}$ with respect to $\mathbf{r}_m$ is

$$\mathbf{A}_{ijl,m} = -\varepsilon_{pjl}\left(\frac{\mathbf{r}_i\,\mathbf{r}_p\,\mathbf{r}_m}{r^5} - \frac{\delta_{im}\,\mathbf{r}_p + \delta_{pm}\,\mathbf{r}_i}{3r^3}\right). \tag{4.53}$$

Multiplying both sides of equation (4.53) by $3r^5\varepsilon_{kml}$, identity (4.16) yields

$$3r^5\varepsilon_{kml}\mathbf{A}_{ijl,m} = \left(\delta_{pm}\delta_{jk} - \delta_{pk}\delta_{jm}\right)\left(3\mathbf{r}_i\mathbf{r}_p\mathbf{r}_m - \delta_{im}\mathbf{r}_p r^2 - \delta_{pm}\mathbf{r}_i r^2\right).$$

Expanding and simplifying using identities (4.14) and (4.17), we have

$$\begin{aligned} 3r^5 \varepsilon_{kml} \mathbf{A}_{ijl,m} &= r^2 \left(3\delta_{jk}\mathbf{r}_i - \delta_{jk}\mathbf{r}_i - 3\delta_{jk}\mathbf{r}_i \right) - \left(3\mathbf{r}_i\mathbf{r}_j\mathbf{r}_k - \delta_{ij}\mathbf{r}_k r^2 - \delta_{jk}\mathbf{r}_i r^2 \right) \\ &= \left(\delta_{ij} r^2 - 3\mathbf{r}_i\mathbf{r}_j \right) \mathbf{r}_k. \end{aligned}$$

Dividing by $3r^5$ yields equation (4.47), which proves the theorem. □□

Corollary 1 *Let $A \subset \mathcal{S}^2$ be a spherical polygon with vertices $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. Then the second-order irradiance tensor $\mathbf{T}^2(A)$ is given by*

$$\mathbf{T}^2(A) = \mathbf{I}\,\frac{\sigma(A)}{3} - \frac{2}{3} \sum_{i=1}^{k} \sin\frac{\Theta_i}{2}\, \mathbf{b}_i\, \mathbf{n}_i^{\mathrm{T}}, \tag{4.54}$$

where Θ_i is the angle between $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$, and

$$\mathbf{n}_i \equiv \frac{\mathbf{u}_i \times \mathbf{u}_{i+1}}{||\, \mathbf{u}_i \times \mathbf{u}_{i+1}\, ||}, \tag{4.55}$$

$$\mathbf{b}_i \equiv \frac{\mathbf{u}_i + \mathbf{u}_{i+1}}{||\, \mathbf{u}_i + \mathbf{u}_{i+1}\, ||}, \tag{4.56}$$

are the unit normals and unit bisectors, respectively,

<u>**Proof:**</u> Let ζ_i be the ith edge of A. That is, ζ_i is the great arc connecting vertices $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$. We compute the boundary integral along ζ_i using two facts: 1) the normal $\mathbf{n}$ is constant along each edge, and 2) the integral of $\mathbf{u}$ along the arc results in a vector that bisects the angle. See Figure 4.9. Thus,

$$\begin{aligned} \int_{\zeta_i} \mathbf{u}\, \mathbf{n}^{\mathrm{T}}\, ds &= \left[\int_{\zeta_i} \mathbf{u}\, ds \right] \mathbf{n}_i^{\mathrm{T}} \\ &= \left[\int_{-\Theta_i/2}^{\Theta_i/2} \cos\theta\, d\theta \right] \mathbf{b}_i \mathbf{n}_i^{\mathrm{T}} \\ &= 2\, \sin\frac{\Theta_i}{2}\, \mathbf{b}_i\, \mathbf{n}_i^{\mathrm{T}}. \end{aligned} \tag{4.57}$$

Substituting the expression above into equation (4.49) gives the result. □□

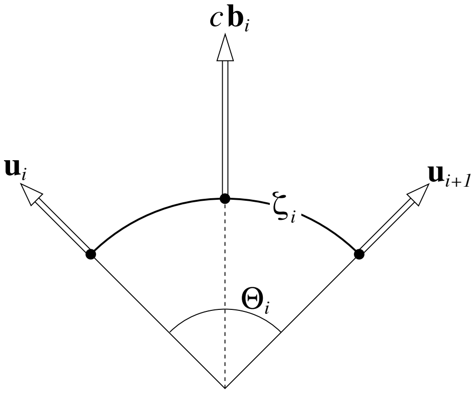

Figure 4.9: *The integral of* $\mathbf{u}$ *over a great arc is a vector proportional to the bisector of the arc. This follows from symmetry.*

Corollary 1 extends Lambert's formula to the second-order form. In itself, this generalization is of little utility other than illustrating some of the machinery needed for the general form, which is the topic of the next section.

4.4.2 A General Recurrence Relation

In this section we show that equation (4.49) can be extended to all higher orders by means of a recurrence relation expressing each irradiance tensor in terms of lower-order tensors.

We introduce some special notation before giving the general theorem for irradiance tensors. If I is the multi-index $(i_1, i_2, \ldots, i_n)$, we define I_k to be the kth index of I, and for any integer $1 \leq k \leq n$ we define I/k to be the $(n-1)$-element multi-index obtained by deleting the kth index of I. That is,

$$\mathrm{I}/k \equiv (i_1, i_2, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n). \tag{4.58}$$

Using the new notation, we now state and prove the central theorem of the chapter.

Theorem 10 *Let $n \geq 0$ be an integer, and let $A \subset \mathcal{S}^2$ be a measurable set. Then the tensor $\mathbf{T}^n(A)$ satisfies the recurrence relation*

$$\mathbf{T}^n_{\mathrm{I}j}(A) = \frac{1}{n+1} \left(\sum_{k=1}^{n-1} \delta_{j\,\mathrm{I}_k} \mathbf{T}^{n-2}_{\mathrm{I}/k}(A) \; - \; \int_{\partial A} \mathbf{u}^{n-1}_{\mathrm{I}} \, \mathbf{n}_j \, ds \right), \tag{4.59}$$

with $\mathbf{T}^0(A) \equiv \sigma(A)$ and $\mathbf{T}^{-1}(A) \equiv 0$, where I is an $(n-1)$-index, ds denotes integration with respect to arclength, and $\mathbf{n}$ is the outward normal to the curve ∂A.

Proof: The strategy will be to transform the boundary integral in equation (4.59) into a sum of irradiance tensors, obtaining the other two terms in the recurrence relation. The proof will proceed in four steps. Step one is to express the boundary integral in terms of $\mathbf{r}$ and $d\mathbf{r}$, which will make it easy to manipulate. Step two is to convert this boundary integral into a surface integral by means of Stokes' theorem. Step three, which is the most laborious, is to express the surface integral in terms of solid angle. The fourth and final step is to show that the resulting surface integral is equivalent to the remaining terms in equation (4.59). For the purpose of the derivation, we shall let P denote any polygon whose projection is the spherical polygon A. That is, $A = \mathbf{\Pi}(P)$.

Step 1: (Rewrite the boundary integral)

We begin by expressing $\mathbf{n}\, ds$ in terms of the position vector $\mathbf{r}$ and its derivative. Note that $\mathbf{n}$ is a function of $\mathbf{u}$ since it is an element of the tangent space of $\mathcal{S}^2$ at $\mathbf{u}$. From equation (4.50), we have $\mathbf{n}\, ds = \mathbf{r} \times d\mathbf{r} \,/\, r^2$. Therefore, we may write

$$\begin{aligned} \frac{1}{n+1} \int_{\partial A} \mathbf{u}^{n-1}_{\mathrm{I}} \, \mathbf{n}_j \, ds &= \frac{1}{n+1} \int_{\partial P} \left[\frac{\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n-1}} \right] \left[\frac{\varepsilon_{jpl} \, \mathbf{r}_p \, d\mathbf{r}_l}{r^2} \right] \\ &= \int_{\partial P} \mathbf{A}^{n+1}_{\mathrm{I}jl} \, d\mathbf{r}_l, \end{aligned} \tag{4.60}$$

where we have introduced the $(n+1)$-order tensor $\mathbf{A}^{n+1}$, which is defined by

$$\mathbf{A}^{n+1}_{\mathrm{I}jl} \equiv - \frac{\varepsilon_{jpl} \, \mathbf{r}_p \, \mathbf{r}^{n-1}_{\mathrm{I}}}{(n+1)\, r^{n+1}}. \tag{4.61}$$

Step 2: (Convert the boundary integral to a surface integral)

This step follows immediately from Stokes' theorem, and the fact that

$$d\left[f_i(\mathbf{r})\, d\mathbf{r}_i\right] \;=\; f_{i,m}(\mathbf{r})\, d\mathbf{r}_m \wedge d\mathbf{r}_i, \tag{4.62}$$

which is a consequence of equations (4.11) and (4.12). Applying Stokes' theorem to equation (4.60), followed by equation (4.62), we have

$$\int_{\partial P} \mathbf{A}^{n+1}_{\mathrm{I}jl}\, d\mathbf{r}_l \;=\; \int_P d\left(\mathbf{A}^{n+1}_{\mathrm{I}jl}\, d\mathbf{r}_l\right) \;=\; \int_P \mathbf{A}^{n+1}_{\mathrm{I}jl,m}\, d\mathbf{r}_m \wedge d\mathbf{r}_l. \tag{4.63}$$

Computing the partial derivative of $\mathbf{A}^{n+1}_{\mathrm{I}jl}$ with respect to $\mathbf{r}_m$, yields

$$\mathbf{A}^{n+1}_{\mathrm{I}jl,m} = \varepsilon_{pjl}\left(\frac{\mathbf{r}_m\,\mathbf{r}_p\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n+3}} \;-\; \frac{\delta_{pm}\,\mathbf{r}^{n-1}_{\mathrm{I}} \;+\; \mathbf{r}_p\,\mathbf{r}^{n-1}_{\mathrm{I},m}}{(n+1)\, r^{n+1}}\right). \tag{4.64}$$

By substituting equation (4.64) into equation (4.63) we obtain a surface integral, although it is not yet in the desired form. This is remedied in the next step.

Step 3: (Express the surface integral in terms of solid angle)

Rewriting the surface integral in terms solid angle will be done in several steps. We first obtain a factor of ε_{kst} by exploiting the anti-commutativity of the wedge product and applying identity (4.16); in doing so we introduce additional dummy indices k, s, and t. Thus,

$$\begin{aligned}
\int_P \mathbf{A}^{n+1}_{\mathrm{I}jl,m}\, d\mathbf{r}_m \wedge d\mathbf{r}_l \;&=\; \int_P \mathbf{A}^{n+1}_{\mathrm{I}jl,m}\left[\frac{d\mathbf{r}_m \wedge d\mathbf{r}_l - d\mathbf{r}_l \wedge d\mathbf{r}_m}{2}\right] \\
&=\; \int_P \mathbf{A}^{n+1}_{\mathrm{I}jl,m}\left[\frac{\delta_{tl}\,\delta_{ms} \;-\; \delta_{tm}\,\delta_{sl}}{2}\right] d\mathbf{r}_s \wedge d\mathbf{r}_t \\
&=\; \int_P \left[\varepsilon_{kml}\,\mathbf{A}^{n+1}_{\mathrm{I}jl,m}\right]\left[\frac{\varepsilon_{kst}\, d\mathbf{r}_s \wedge d\mathbf{r}_t}{2}\right].
\end{aligned} \tag{4.65}$$

Note that the presence of ε_{kst} makes the final factor on the right very similar to $d\omega$; it differs only by a factor of $\mathbf{r}_k/r^3$. To complete it, we obtain the missing factor after expanding the other bracketed expression. Multiplying equation (4.64) by ε_{kml} and simplifying, we have

$$\begin{aligned}
\varepsilon_{kml}\,\mathbf{A}^{n+1}_{\mathrm{I}jl,m} &= \left(\delta_{pm}\,\delta_{jk} - \delta_{pk}\,\delta_{jm}\right)\left[\frac{\mathbf{r}_m\,\mathbf{r}_p\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n+3}} - \frac{\delta_{pm}\,\mathbf{r}^{n-1}_{\mathrm{I}} + \mathbf{r}_p\,\mathbf{r}^{n-1}_{\mathrm{I},m}}{(n+1)\,r^{n+1}}\right] \\
&= \left[\frac{\mathbf{r}_k\,\mathbf{r}^{n-1}_{\mathrm{I},j}}{(n+1)\,r^{n+1}} - \frac{\mathbf{r}_j\,\mathbf{r}_k\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n+3}}\right] + \delta_{jk}\left[\frac{\mathbf{r}_m\,\mathbf{r}^{n-1}_{\mathrm{I},m} - (n-1)\,\mathbf{r}^{n-1}_{\mathrm{I}}}{(n+1)\,r^{n+1}}\right].
\end{aligned}$$

The above expression simplifies further since the final term vanishes; this follows from the observation that

$$\mathbf{r}_m\,\mathbf{r}^{n-1}_{\mathrm{I},m} \;=\; \mathbf{r}_m \sum_{k=1}^{n-1} \delta_{m\,\mathrm{I}_k}\,\mathbf{r}^{n-2}_{\mathrm{I}/k} \;=\; \sum_{k=1}^{n-1} \mathbf{r}_{\mathrm{I}_k}\,\mathbf{r}^{n-2}_{\mathrm{I}/k} \;=\; (n-1)\,\mathbf{r}^{n-1}_{\mathrm{I}}. \tag{4.66}$$

Incorporating this simplification, equation (4.65) may now be written as

$$\begin{aligned}
\int_P \mathbf{A}^{n+1}_{\mathrm{I}jl,m}\,d\mathbf{r}_m \wedge d\mathbf{r}_l &= \int_P \left[\frac{\mathbf{r}_k\,\mathbf{r}^{n-1}_{\mathrm{I},j}}{(n+1)\,r^{n+1}} - \frac{\mathbf{r}_j\,\mathbf{r}_k\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n+3}}\right]\left[\frac{\varepsilon_{kst}\,d\mathbf{r}_s \wedge d\mathbf{r}_t}{2}\right] \\
&= \int_P \left[\frac{\mathbf{r}^{n-1}_{\mathrm{I},j}}{(n+1)\,r^{n-2}} - \frac{\mathbf{r}_j\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^{n}}\right] d\omega,
\end{aligned} \tag{4.67}$$

where the factor of $\mathbf{r}_k/r^3$ has been used to complete the solid angle 2-form $d\omega$. The result is an integral over solid angle, as desired.

Step 4: (Convert to a sum of irradiance tensors)

It is now straightforward to express equation (4.67) in terms of irradiance tensors. Breaking equation (4.67) into two terms and expanding $\mathbf{r}^{n-1}_{\mathrm{I},j}$, as we did in equation (4.66), we have

$$\int_{\partial P} \mathbf{A}^{n+1}_{\mathrm{I}jl}\,d\mathbf{r}_l \;=\; \frac{1}{n+1}\int_P \left[\sum_{k=1}^{n-1} \frac{\delta_{j\,\mathrm{I}_k}\,\mathbf{r}^{n-2}_{\mathrm{I}/k}}{r^{n-2}}\right] d\omega \;-\; \int_P \frac{\mathbf{r}_j\,\mathbf{r}^{n-1}_{\mathrm{I}}}{r^n}\,d\omega. \tag{4.68}$$

Finally, using equation (4.60) to replace the expression of the left, and identifying the two integrals on the right of equation (4.68) as irradiance tensors of orders n and $n-2$ respectively, we have

$$\frac{1}{n+1}\int_{\partial A} \mathbf{u}^{n-1}_{\mathrm{I}}\,\mathbf{n}_j\,ds \;=\; \frac{1}{n+1}\sum_{k=1}^{n-1} \delta_{j\,\mathrm{I}_k}\,\mathbf{T}^{n-2}_{\mathrm{I}/k}(A) \;-\; \mathbf{T}^{n}_{\mathrm{I}j}(A) \tag{4.69}$$

which proves the theorem. □□

From theorem 10 it can be seen that each tensor of the form shown in equation (4.21) can be reduced to a boundary integral and a term constructed from the tensor of two orders lower; the latter being added along generalized "rows" and "columns" of $\mathbf{T}^n$. Note that the recursive formulation subsumes the expression for Ξ^n given in theorem 8.

Another consequence of the theorem is that $\mathbf{T}^n(A)$ can be computed analytically whenever the corresponding boundary integrals and base case can be. In particular, when A is the spherical projection of a k-sided polygon and $n = 1$, equation (4.59) yields

$$\mathbf{T}^1_j(A) \;=\; -\frac{1}{2}\int_{\partial A} \mathbf{n}_j \, ds \;=\; -\frac{1}{2}\sum_{i=1}^{k} \Theta_i \, \mathbf{n}^i_j, \tag{4.70}$$

where Θ_i is the angle subtended by the ith edge of the polygon, and $\mathbf{n}^i$ is its outward normal. This is the vector form of Lambert's well-known formula.

Theorem 10 defines a family of closed-form expressions that provides a natural extension of Lambert's formula. Although equation (4.59) is impractical computationally for moments of order three and higher, it succinctly expresses the relationship among all the tensors. For instance, it is apparent that all even-order tensors incorporate solid angle $\mathbf{T}^0(A)$, while the odd-order tensors do not. We now derive several useful formulas from this equation and apply them to problems of image synthesis.

4.5 Angular Moments

With equation (4.59) as a starting point, we may obtain expressions for individual moments, or sums of moments, without explicitly constructing the tensors. This is of great practical importance since the size of $\mathbf{T}^n(A)$ grows exponentially with n, yet only $\mathrm{O}(n^2)$ of its elements are distinct. In particular, we shall find that the polynomials corresponding to the steerable lobes described in section 4.2.2 can be integrated over polygonal domains with the aid of irradiance tensors. These lobes define a special class of weighting functions over the sphere that are fundamentally related to irradiance tensors, and lead to a number of useful formulas by means of equation (4.59).

4.5.1 Axial Moments

We begin by considering the special case of moments about an axis, which defines a simple class of polynomials over the sphere. Given an arbitrary subset $A \subset \mathcal{S}^2$ and a unit vector $\mathbf{w}$, we define the nth *axial moment* of A about $\mathbf{w}$ by

$$\bar{\tau}^n(A, \mathbf{w}) \equiv \int_A (\mathbf{w} \cdot \mathbf{u})^n \; d\sigma(\mathbf{u}). \tag{4.71}$$

More precisely, equation (4.71) is a moment of the *characteristic function* of A; that is, the function defined on $\mathcal{S}^2$ that is one on A and zero elsewhere. As a cosine to a power, the polynomial weighting function within $\bar{\tau}^n$ is essentially a Phong distribution centered around the direction $\mathbf{w}$, as shown in Figure 4.10a. These functions are *steerable* in the sense that they to be re-oriented without raising their order [106]. The ease of controlling the shape and direction of the lobe makes this polynomial function useful in approximating reflectance functions, as Ward did with Gaussians [173], or an exact representation of simple Phong-like functions. To obtain a closed-form expression for $\bar{\tau}^n$, we begin by expressing the integrand of equation (4.71) as a composition of tensors:

$$(\mathbf{w} \cdot \mathbf{u})^n = (\mathbf{u} \otimes \cdots \otimes \mathbf{u})_{\mathrm{I}} \, (\mathbf{w} \otimes \cdots \otimes \mathbf{w})_{\mathrm{I}}. \tag{4.72}$$

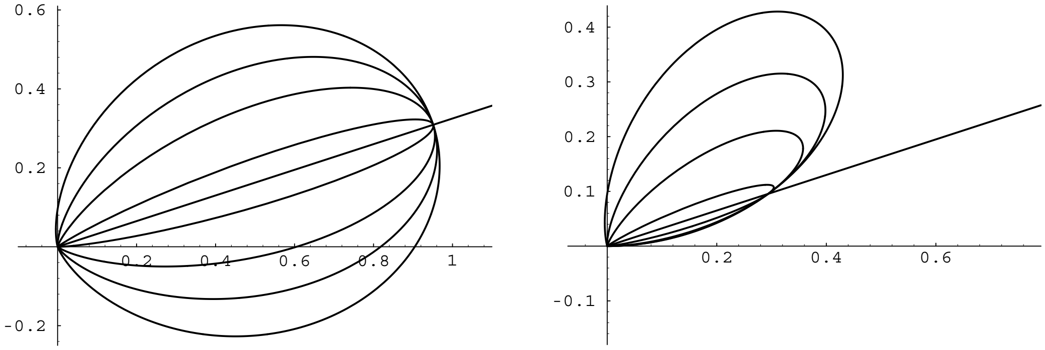

Figure 4.10: *Cross-sections of the weighting functions* $(\mathbf{w}\cdot\mathbf{u})^n$ *and* $(\mathbf{w}\cdot\mathbf{u})^n(\mathbf{v}\cdot\mathbf{u})$ *where* $\mathbf{v}$ *is the vertical axis; moment orders are 2, 4, 10, and 100, starting from the outer curves.*

The summation convention applies to all repeated pairs of indices, which includes all sub-indices of I in equation (4.72). It follows that

$$\bar{\tau}^n(A, \mathbf{w}) = \mathbf{T}^n_{\mathrm{I}}(A)(\mathbf{w} \otimes \cdots \otimes \mathbf{w})_{\mathrm{I}}, \tag{4.73}$$

which associates the nth axial moment with the nth-order tensor. Using equation (4.59) to expand equation (4.73), and simplifying by means of equation (4.17), we obtain a recurrence relation for the axial moment $\bar{\tau}^n(A, \mathbf{w})$:

$$(n+1)\;\bar{\tau}^n \;=\; (n-1)\;(\mathbf{w}\cdot\mathbf{w})\;\bar{\tau}^{n-2} \;-\; \int_{\partial A} (\mathbf{w}\cdot\mathbf{u})^{n-1}\;\mathbf{w}\cdot\mathbf{n}\,ds, \tag{4.74}$$

where the function arguments have been omitted for brevity. Equation (4.74) is a recurrence relation for $\bar{\tau}^n$ with base cases $\bar{\tau}^{-1}(A) = 0$ and $\bar{\tau}^0(A) = \sigma(A)$. When $n > 0$ the recurrence relation reduces to a single boundary integral involving a polynomial in $\mathbf{w}\cdot\mathbf{u}$. Since $\mathbf{w}$ is a unit vector, we have

$$\begin{aligned}(n+1)\;\bar{\tau}^n \;=\; \bar{\tau}^q \;-\; \int_{\partial A} \Big[&(\mathbf{w}\cdot\mathbf{u})^{n-1} + (\mathbf{w}\cdot\mathbf{u})^{n-3} + \cdots \\ &+\;(\mathbf{w}\cdot\mathbf{u})^{q+1}\Big]\;\mathbf{w}\cdot\mathbf{n}\;ds,\end{aligned} \tag{4.75}$$

where $q = 0$ if n is even, and $q = -1$ if n is odd. This expression is useful as a component of more general expressions, such as double-axis moments.

4.5.2 Double-Axis Moments

An important generalization of equation (4.71) is to allow for moments with respect to multiple axes simultaneously; this will prove useful for handling radiant exchanges involving pairs of surfaces. We define the *double-axis moment* of A with respect to $\mathbf{w}$ and $\mathbf{v}$ by

$$\bar{\bar{\tau}}^{n,m}(A, \mathbf{w}, \mathbf{v}) \equiv \int_A (\mathbf{w} \cdot \mathbf{u})^n \; (\mathbf{v} \cdot \mathbf{u})^m \; d\sigma(\mathbf{u}). \tag{4.76}$$

A recurrence relation for $\bar{\bar{\tau}}^{n,m}$ can also be obtained from equation (4.59) by expressing the integrand as a tensor composition with n copies of $\mathbf{w}$ and m copies of the vector $\mathbf{v}$. We shall only consider the case where $m = 1$, which corresponds to $\mathbf{T}^{n+1}(A)(\mathbf{w} \otimes \cdots \otimes \mathbf{w} \otimes \mathbf{v})$ and yields the formula

$$(n+2)\; \bar{\bar{\tau}}^{n,1}(A, \mathbf{w}, \mathbf{v}) \;=\; n\,(\mathbf{w} \cdot \mathbf{v})\; \bar{\tau}^{n-1}(A, \mathbf{w}) \;-\; \int_{\partial A} (\mathbf{w} \cdot \mathbf{u})^n \; \mathbf{v} \cdot \mathbf{n}\; ds. \tag{4.77}$$

Figure 4.10a shows how an additional axis can change the shape of the weighting function. Note that when $\mathbf{v} = \mathbf{w}$, equation (4.77) reduces to the $(n+1)$-order axial moment given by equation (4.74). Recurrence relations for $\bar{\bar{\tau}}^{n,m}$ with $m > 1$ can be obtained in a similar manner, although the resulting boundary integrals are more difficult to evaluate.

Evaluating equations (4.75) and (4.77) in closed form is the topic of the next section. In section 4.7 we show how these moments can be applied to the simulation of non-diffuse phenomena.

4.6 Exact Evaluation of Moments

Equations (4.59) and (4.75) reduce tensors and moments to one-dimensional integrals and (in the case of even orders) solid angle. This section describes how both of these components can be evaluated in closed form. Thus far no restrictions have been placed on the region $A \subset \mathcal{S}^2$; however, we shall now assume that A is the spherical projection of a polygon $P \subset \mathrm{I\!R}^3$, which may be non-convex. The

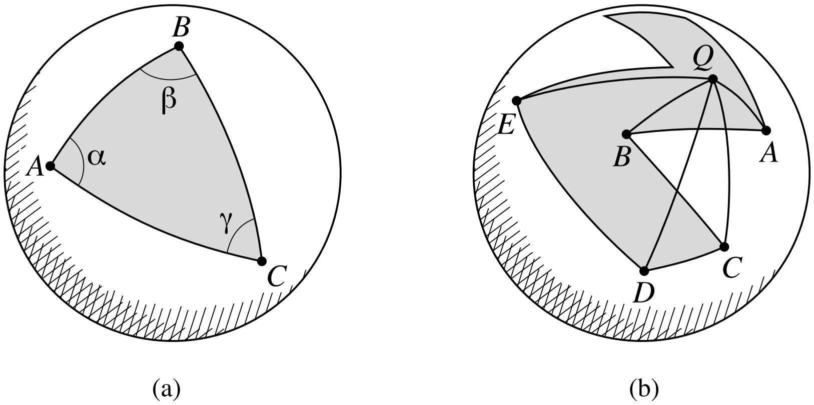

Figure 4.11: *(a) The solid angle of a spherical triangle is easily obtained from the internal angles. (b) Non-convex spherical polygons can be handled by spoking into triangles from an arbitrary point Q and summing the signed areas.*

resulting projection is a *geodesic* or *spherical* polygon, whose edges are great arcs; that is, segments of great circles.

When P is a polygon, the computation of solid angles and boundary integrals are both greatly simplified. First, and most importantly, the outward normal $\mathbf{n}$ is constant along each edge of a spherical polygon which allows the factors of $\mathbf{w}\cdot\mathbf{n}$ and $\mathbf{v}\cdot\mathbf{n}$ to be moved outside the integrals in equations (4.75) and (4.77) respectively. A second simplification emerges in the parametrization of the boundary integrals, as we shall see below.

4.6.1 Solid Angle

The solid angle subtended by A is given by a simple surface integral. Because the corresponding differential 2-form is not exact, however, it cannot be simplified further to a boundary integral [156, p. 131]. Fortunately, the solid angle subtended by a polygon can be computed directly in another way. If P is a triangle in $\mathbb{R}^3$ its projection $\mathbf{\Pi}(P)$ is a spherical triangle $\triangle ABC$. Girard's formula for spherical triangles [19, p. 278] then states that

$$\sigma(\triangle ABC) = \alpha + \beta + \gamma - \pi, \tag{4.78}$$

where α, β, and γ are the three internal angles, as shown in Figure 4.11a. The internal angles are the dihedral angles between the planes containing the edges. For instance, the angle α in Figure 4.11a is given by

$$\alpha = \cos^{-1} \frac{(B \times A) \cdot (A \times C)}{||\, B \times A \,||\; ||\, A \times C \,||}. \tag{4.79}$$

Equation (4.78) generalizes immediately to arbitrary convex polygons [19, p. 279]; however, non-convex polygons require a slightly different approach. The solid angle subtended by a non-convex polygon can be computed by covering its spherical projection with n triangles, one per edge, all sharing an arbitrary vertex $Q \in \mathcal{S}^2$. See Figure 4.11b. The solid angle is then the sum of the triangle areas signed according to orientation. In the figure, $\triangle QAB$, $\triangle QCD$, and $\triangle QDE$ are all positive, while $\triangle QBC$ is negative, according to the clock-sense of the vertices. This method avoids the complication of decomposing $\mathbf{\Pi}(P)$ into triangles.

4.6.2 Boundary Integrals

The boundary integrals in equations (4.75) and (4.77) can be approximated by numerical quadrature, or evaluated analytically in terms of $\mathrm{O}(n)$ elementary functions per edge. We shall only describe analytic evaluation and a related approximation, both based upon a sum of integrals of the form

$$F(x, n) \equiv \int_0^x \cos^n \theta \; d\theta. \tag{4.80}$$

Integrals of this form may be evaluated exactly using a recurrence relation given below. To express the integral in equation (4.75) in terms of $F(x, n)$ when A is a spherical polygon, we proceed by parametrizing the great arc ζ defined by each edge as

$$\mathbf{u}(\theta) = \mathbf{s} \cos\theta \; + \; \mathbf{t} \sin\theta,$$

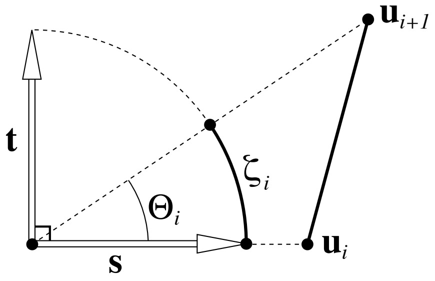

Figure 4.12: *The vectors used to parametrize the arc defined by a polygon edge.*

where $\mathbf{s}$ and $\mathbf{t}$ are orthonormal vectors in the plane containing the edge and the origin, with $\mathbf{s}$ directed toward the first vertex. See Figure 4.12.

To simplify the line integral over the great arc $\zeta \in \mathcal{S}^2$, let Θ be the angle subtended by the arc, which is also the length of ζ, and let

$$\begin{aligned} a &\equiv \mathbf{w} \cdot \mathbf{s} \\ b &\equiv \mathbf{w} \cdot \mathbf{t} \\ c &\equiv \sqrt{a^2 + b^2}. \end{aligned}$$

From the parametrization given above, it follows that

$$\begin{aligned} \int_\zeta (\mathbf{w} \cdot \mathbf{u})^n \, ds &= \int_0^\Theta \left[a \cos\theta + b \sin\theta \right]^n \, d\theta \\ &= c^n \int_0^\Theta \cos^n(\theta - \phi) \, d\theta, \\ &= c^n \left[F(\Theta - \phi, n) - F(-\phi, n) \right], \end{aligned} \tag{4.81}$$

where ϕ is the angle satisfying $\cos\phi = a/c$ and $\sin\phi = b/c$. The function $F(x, n)$ may then be evaluated in $\lfloor (n+1)/2 \rfloor$ steps by means of the recurrence relation

$$F(x, n) = \frac{1}{n} \left[\cos^{n-1} x \, \sin x + (n-1) F(x, n-2) \right], \tag{4.82}$$

where $F(x, 1) = \sin x$ and $F(x, 0) = x$. Recurrence relation (4.82) follows easily from equation (4.80) after integrating by parts. Finally, the complete integral in equation (4.75) is a weighted sum of the integrals shown in equation (4.81) with a sequence of different exponents, which are all even or all odd. By computing this sequence of integrals incrementally, the complete sum can be expressed in terms of O(n) elementary functions, as demonstrated in the following section.

4.6.3 Algorithms for Efficient Evaluation

We now show that nth-order axial moments of a k-sided polygon may be computed exactly (in the absence of roundoff error) in O(nk) time, and provide the complete algorithm in pseudo-code. The algorithm evaluates equation (4.75) in O(n) time for each of the k edges, using recurrence relation (4.82). The steps are conveniently broken into three procedures that build upon one another: *CosSumIntegral*, *LineIntegral*, and *BoundaryIntegral*. The key to efficient evaluation is the procedure *CosSumIntegral*, which computes the sum

$$c^k F(x, k) + c^{k+2} F(x, k+2) + \cdots + c^n F(x, n), \tag{4.83}$$

where $k = m$ if $m + n$ is even, and $k = m + 1$ otherwise, for a given integer $m \geq 0$. The reason for the parameter m will become apparent later; it is included so that the procedure *CosSumIntegral* can accommodate both single- and double-axis moments.

Because recurrence relation (4.82) generates integrals of cosine with increasing powers, all integrals in expression (4.83) may be generated as easily as the last term $c^n F(x, n)$. This strategy is embodied in the procedure *CosSumIntegral*, which is shown below.

```
real CosSumIntegral( real $x, c$; integer $m, n$ )
    integer $i \leftarrow$ if even($n$) then $0$ else $1$;    Loop counter.
    real $F \leftarrow$ if even($n$) then $x$ else $\sin x$;    Cosine integrals.
    real $S \leftarrow 0$;    Accumulates the final sum.
    while $i \leq n$ do
        if $i \geq m$ then $S \leftarrow S + c^i * F$;
        $F \leftarrow [\cos^{i+1} x \ \sin x + (i+1)F] \ / \ (i+2)$;
        $i \leftarrow i + 2$;
        endwhile
    return $S$;
    end
```

The next procedure, *LineIntegral*, reduces the line integral corresponding to a polygon edge into a sum of cosine integrals; the steps correspond to equation (4.81), summed over a sequence of exponents from m to n.

```
real LineIntegral( vector $\mathbf{A}, \mathbf{B}, \mathbf{w}$; integer $m, n$ )
    if $(n < 0)$ or $(\mathbf{w} \perp \mathbf{A}$ and $\mathbf{w} \perp \mathbf{B})$ then return $0$;
    vector $\mathbf{s} \leftarrow$ Normalize $[\mathbf{A}]$;
    vector $\mathbf{t} \leftarrow$ Normalize $[(\mathbf{I} - \mathbf{s}\mathbf{s}^{\mathrm{T}})\,\mathbf{B}]$;    Component orthogonal to $\mathbf{A}$.
    real $a \leftarrow \mathbf{w} \cdot \mathbf{s}$;
    real $b \leftarrow \mathbf{w} \cdot \mathbf{t}$;
    real $c \leftarrow \sqrt{a^2 + b^2}$;
    real $\Theta \leftarrow$ angle between $\mathbf{A}$ and $\mathbf{B}$;
    real $\phi \leftarrow \mathrm{sign}(\ b\ ) * \cos^{-1}(a/c)$;
    return $CosSumIntegral(\Theta - \phi, c, m, n) - CosSumIntegral(-\phi, c, m, n)$;
    end
```

The next procedure, *BoundaryIntegral*, computes the complete boundary integral for a given k-sided polygon P by forming a weighted sum of k line integrals. The weight associated with each edge is the cosine of the angle between its outward normal and the second vector $\mathbf{v}$, which may coincide with $\mathbf{w}$.

```
**real** *BoundaryIntegral*( **polygon** $P$; **vector** $\mathbf{w}$, $\mathbf{v}$; **integer** $m, n$ )
    **real** $b \leftarrow 0$;    *Accumulates the contribution from each edge.*
    **for each** edge **AB in** $P$ **do**
        **vector** $\mathbf{n} \leftarrow$ *Normalize* $[\mathbf{A} \times \mathbf{B}]$;
        $b \leftarrow b + (\mathbf{n} \cdot \mathbf{v}) * \mathit{LineIntegral}(\mathbf{A}, \mathbf{B}, \mathbf{w}, m, n)$;
        **endfor**
    **return** $b$;
    **end**
```

With these three basic procedures we may now define the procedure *AxialMoment*, which computes the nth axial moment of a polygon P with respect to the axis $\mathbf{w}$. Even-order moments also require the computation of a "signed" solid angle, which is handled by this procedure. Equation (4.75) then corresponds to the simple procedure *AxialMoment* which follows.

```
**real** *AxialMoment*( **polygon** $P$; **vector** $\mathbf{w}$; **integer** $n$ )
    **real** $a \leftarrow -\mathit{BoundaryIntegral}(P, \mathbf{w}, \mathbf{w}, 0, n-1)$;
    **if even**$(n)$ **then** $a \leftarrow a + \mathit{SolidAngle}(P)$;
    **return** $a/(n+1)$;
    **end**
```

The function *SolidAngle* returns the solid angle subtended by the polygon P using the method described in section 4.6.1. Because the sign of the boundary integral depends on the orientation of the polygon, the solid angle must be similarly signed. Thus, *SolidAngle* is positive if the vertices of P are counter-clockwise, as seen from the origin, and negative otherwise.

Finally, the procedure *DoubleAxisMoment* computes the nth-order moment of a polygon P with respect to the $\mathbf{w}$ axis and the 1st-order moment with respect to the $\mathbf{v}$ axis. Equation (4.77) then corresponds to the procedure *DoubleAxisMoment*, which is given next.

```
real DoubleAxisMoment( polygon $P$; vector $\mathbf{w}, \mathbf{v}$; integer $n$ )
    if $n = 0$ then return AxialMoment$(P, \mathbf{v}, n)$;
    real $a \leftarrow$ AxialMoment$(P, \mathbf{w}, n-1)$;
    real $b \leftarrow$ BoundaryIntegral$(P, \mathbf{w}, \mathbf{v}, n, n)$;
    return $(n * a * \mathbf{w} \cdot \mathbf{v} - b) \,/\, (n+2)$;
    end
```

It is easy to see that both *AxialMoment* and *DoubleAxisMoment* require $\mathrm{O}(nk)$ time, assuming that trigonometric and other elementary functions are evaluated in constant time. Note that the redundancy in calling both *AxialMoment* and *BoundaryIntegral* can easily be removed; however, this does not affect the order of the algorithm.

4.6.4 Normalization

In applying the above methods, it is frequently useful to normalize the resulting distributions while ignoring negative lobes. We now show how this can be done for distributions defined in terms of double-axis moments.

The two planes orthogonal to the axes of a double-axis moment partition the sphere into four spherical *lunes* (also known as spherical *digons*); let L_1 and L_2 be the two lunes in the positive half-space defined by $\mathbf{v}$, as shown in Figure 4.13. To normalize the weighting function associated with a double-axis moment, for example, we must compute the integral

$$\begin{aligned} N(\mathbf{w}, \mathbf{v}, n) &\equiv \int_{L_2} (\mathbf{w} \cdot \mathbf{u})^n \; (\mathbf{v} \cdot \mathbf{u}) \; d\sigma(\mathbf{u}) \\ &= \overline{\overline{\overline{\tau}}}^{n,1}(L_2, \mathbf{w}, \mathbf{v}), \end{aligned} \tag{4.84}$$

where $\mathbf{v}$ is the surface normal; thus, the integrand of equation (4.84) is positive for all exponents n on the lune L_2. All luminaires must be clipped by the planes defining this lune when computing the moments.

Integral (4.84) can be evaluated analytically using equation (4.77), which results in two boundary integrals corresponding to the arcs C_{w} and C_{v} in Figure 4.13. The special nature of the boundaries greatly simplifies the computation. In particular, we have $\mathbf{w} \cdot \mathbf{u} = 0$ and $\mathbf{n} = \mathbf{w}$ on the curve C_{w}, while $\mathbf{v} \cdot \mathbf{n} = 1$ on the curve C_{v}. Substituting equation (4.75) into equation (4.77) with $\partial A = C_{\mathrm{v}} \cup C_{\mathrm{w}}$, and applying the aforementioned simplifications, we have

$$\begin{aligned}
(n+2)\ N(\mathbf{w}, \mathbf{v}, n) \ &= \ \int_{C_{\mathrm{v}}} (\mathbf{w} \cdot \mathbf{u})^n \ ds \\
&+ \ (\mathbf{w} \cdot \mathbf{v}) \left[\, \bar{\tau}^q(L_2, \mathbf{w}) \ + \ \int_{C_{\mathrm{w}}} |q| \ ds \, \right] \\
&+ \ (\mathbf{w} \cdot \mathbf{v})^2 \int_{C_{\mathrm{v}}} \left[(\mathbf{w} \cdot \mathbf{u})^{n-2} + \cdots + (\mathbf{w} \cdot \mathbf{u})^{q+1} \right] ds, \qquad (4.85)
\end{aligned}$$

where $q = 0$ if $n - 1$ is even, and $q = -1$ if $n - 1$ is odd.

To apply the procedure *CosSumIntegral* to the computation of the integrals above, we split the lune L_2 into two identical spherical triangles. Letting $\triangle$ denote one of the halves, by symmetry it follows that

$$N(\mathbf{w}, \mathbf{v}, n) \ = \ 2\, \bar{\bar{\tau}}^{n,1}(\triangle, \mathbf{w}, \mathbf{v}).$$

The computation within the loop of procedure *CosSumIntegral* is greatly simplified by the fact that $x = \pi/2$, which eliminates the trigonometric terms. Consequently, the terms of the sum (4.83) can be generated using the recurrence formula

$$t_{i+2} \ = \ c^2 \left(\frac{i+1}{i+2} \right) t_i,$$

starting with $t_0 = \pi/2$ if n is even, and with $t_1 = c$ if n is odd, where

$$c = \sqrt{1 - (\mathbf{w} \cdot \mathbf{v})^2}.$$

To compute $\bar{\tau}^0(\triangle, \mathbf{w}) = \sigma(\triangle)$ we note that the area of the lune L_2 is twice its internal angle θ_2, and $\cos\theta_1 = \mathbf{w} \cdot \mathbf{v}$. Thus,

$$\sigma(\triangle) \ = \ \theta_2 \ = \ \pi \ - \ \cos^{-1} \mathbf{w} \cdot \mathbf{v}.$$

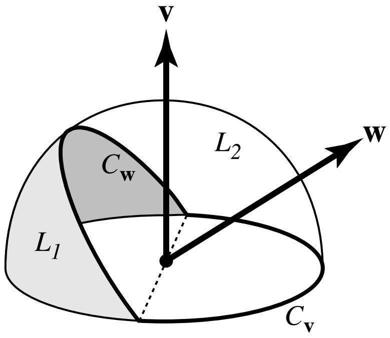

Figure 4.13: *Normalizing the weighting function of a double-axis moment reduces to computing a boundary integral along the two half circles C_v and C_w. This boundary defines the spherical lune L_2 on which the moment is positive for all n.*

Combining these facts and the minor differences due to parity, we obtain a procedure for evaluating integral (4.84) given arbitrary unit vectors **w** and **v** and any integer $n \geq 0$. The optimized algorithm, which requires $O(n)$ time, is shown below.

```
real $N$( vector $\mathbf{w}, \mathbf{v}$; integer $n$ )
    real $S \leftarrow 0$;   Accumulates boundary integral along $C_v$.
    real $d \leftarrow \mathbf{w} \cdot \mathbf{v}$;
    real $c \leftarrow \sqrt{1 - d^2}$;
    real $t \leftarrow$ if even($n$) then $\pi/2$ else $c$;
    real $A \leftarrow$ if even($n$) then $\pi/2$ else $\pi - \cos^{-1}(d)$;
    integer $i \leftarrow$ if even($n$) then $0$ else $1$;
    while $i \leq n - 2$ do
        $S \leftarrow S + t$;
        $t \leftarrow t * c^2 * (i + 1) \,/\, (i + 2)$;
        $i \leftarrow i + 2$;
        endwhile
    return $2 * (t \, + \, d * A \, + \, d^2 * S) \,/\, (n + 2)$;
    end
```

The three terms t, $d * A$, and $d^2 * S$ in the final expression of the above procedure correspond to the first, second, and third terms of equation (4.85) respectively.

4.6.5 Optimizations

We conclude this section on exact evaluation by describing some simple optimizations to the procedures for computing moments, and by showing how the exact formulas can lead to efficient approximations with error bounds.

For clarity, the pseudo-code in the previous section does not depict a number of simple optimizations. For instance, the powers in procedure *CosSumIntegral* may be computed incrementally by repeated multiplication, and no trigonometric functions need be evaluated in the inner loop. Also, in computing double-axis moments, a great deal of redundant computation may be avoided by allowing procedure *CosSumIntegral* to return one additional term in the series, as well as by computing both endpoints simultaneously [7]. These optimizations do not change the time complexity of the algorithms, but can significantly reduce the constant.

While there are generally benefits to obtaining closed-form expressions, they are not always practical computationally. An important means of speeding the computation is to settle for an approximation. To arrive at one such approximation, note that the terms in equation (4.75) decrease in magnitude monotonically since $|F_{k+2}| < |F_k|$ for all k, and $0 \leq c \leq 1$. When the terms approach zero rapidly we may therefore obtain an accurate approximation with little work. Moreover, by bounding the tail of the series it is possible to guarantee a given tolerance. For example, to compute a double-axis moment to a relative accuracy of ϵ, the loop in *CosSumIntegral* may be terminated immediately upon updating S if the condition

$$\left| \frac{(\mathbf{v} \cdot \mathbf{n})\, c^n\, F}{(\mathbf{u} \cdot \mathbf{v})(\mathbf{u} \cdot \mathbf{n})} \right| + \left| \frac{c^k\, F}{1 - c^2} \right| \leq \epsilon\, |S| \tag{4.86}$$

is met. In this case the tail of the series and the final integral in equation (4.77) may be dropped. Early termination of the loop is most useful with high orders. Because the test is costly, it should not be performed at every iteration of the loop.

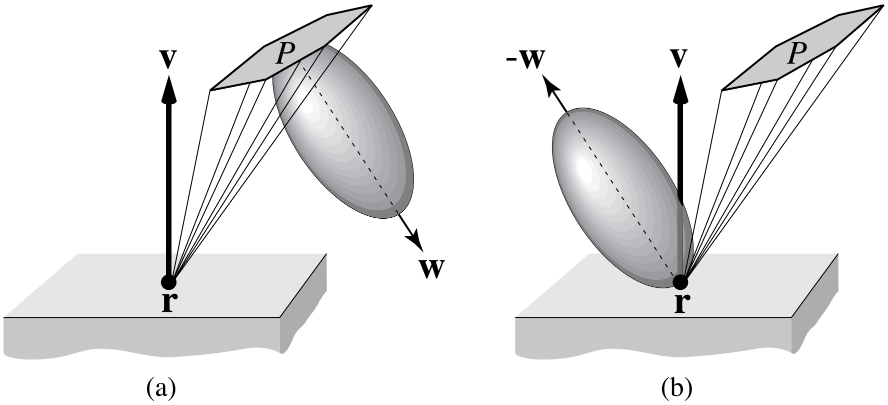

Figure 4.14: *(a) Computing the irradiance at the point* **r** *due to a directionally-varying area light source* P *is equivalent to (b) computing a double-axis moment of a uniform Lambertian source of the same shape.*

4.7 Applications to Image Synthesis

For ideal diffuse surfaces irradiance is sufficient to compute reflected radiance. The situation is dramatically different for non-diffuse surfaces, however. In the extreme case of ideal specular surfaces, all features of the incident illumination appear again in the reflection. For surfaces that are neither ideal diffuse nor ideal specular, high-order moments can be used to quantify additional features of the incident illumination, much as a power series expansion. For polygonal environments with emission and reflection distributions defined in terms of simple polynomials, the procedures given in the previous section may be applied to the simulation of directional luminaires, glossy reflections, and glossy transmissions. These applications are described in the following sections.

4.7.1 Directional Luminaires

Methods for simulating the illumination due to diffuse area sources [108] and directional point sources [168] are well known; however, directional area sources are problematic for deterministic methods. In this section we shall see how a class of

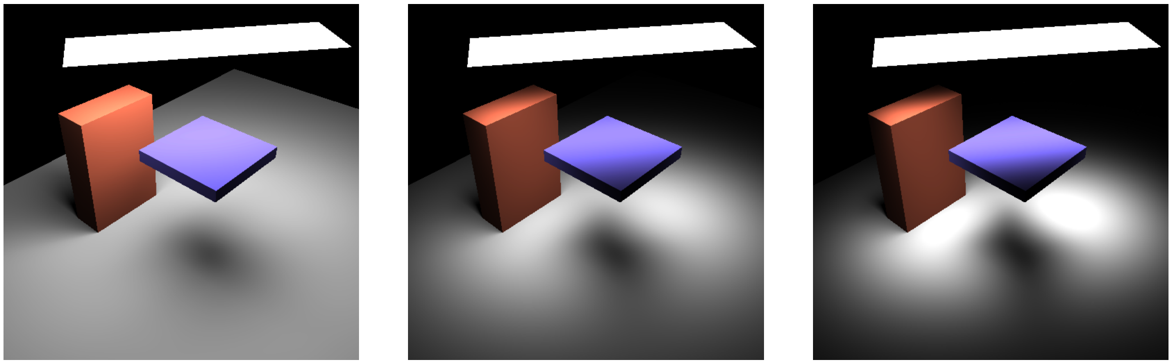

Figure 4.15: *Examples of directional luminaires. At each point the irradiance is computed analytically from the area light source. The moment orders are 0, 10, and 20 respectively.*

directional luminaires can be handled using double-axis moments.

Let P be a polygonal luminaire whose emission distribution is spatially uniform but varies directionally according to a Phong distribution; that is, as a cosine to a power [113]. For instance, the direction of maximum radiance may be normal to the plane of the luminaire and fall off rapidly in other directions, as shown by the distribution in Figure 4.14a. The irradiance at the point $\mathbf{r}$ is then given by

$$\int_{\mathbf{\Pi}(P')} |\mathbf{u} \cdot \mathbf{w}|^n \cos\theta \; d\sigma(\mathbf{u}), \tag{4.87}$$

where P' is the luminaire translated by $-\mathbf{r}$, and θ is the angle of incidence of $\mathbf{u}$; that is, $\cos\theta = \mathbf{v} \cdot \mathbf{u}$, where $\mathbf{v}$ is the surface normal. Observe that this computation is equivalent to a double-axis moment of a Lambertian source P, where the nth moment is taken with respect to $-\mathbf{w}$ and the factor of $\cos\theta$ is accounted for by the second axis $\mathbf{v}$. See Figure 4.14b. Therefore, procedure *DoubleAxisMoment* can be used to compute the irradiance due to directional luminaires of this form exactly.

Figure 4.15 shows a simple scene illuminated by an area source with three different directional distributions. Note that the areas directly beneath the luminaire get brighter with higher orders, increasing the contrast with the surrounding areas, which get darker. Polygonal occlusions are handled by clipping the luminaire

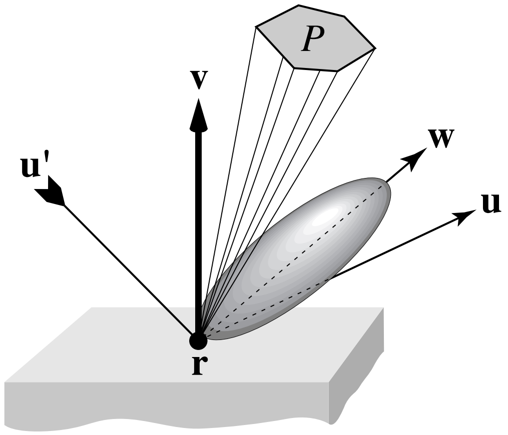

Figure 4.16: *A simple BRDF defined by an axial moment around the mirror reflection* $\mathbf{w}$ *of* $\mathbf{u}'$*. By reciprocity, the radiance in the direction* $-\mathbf{u}'$ *due to* P *reduces to a double-axis moment of* P *with respect to* $\mathbf{w}$ *and* $\mathbf{v}$*.*

against all blockers and computing the contribution from each remaining portion, precisely as Nishita and Nakamae [108] handled Lambertian sources.

4.7.2 Glossy Reflection

A similar strategy can be used for computing glossy reflections of polygonal Lambertian luminaires. Let $\mathbf{r}$ be a point on a reflective surface. Then the reflected radiance at $\mathbf{r}$ in the direction $\mathbf{u}$ due to luminaire P is given by

$$f(\mathbf{r}, \mathbf{u}) = \int_{\mathbf{\Pi}(P)} \rho(\mathbf{u}' \to \mathbf{u})\, f(\mathbf{r}, \mathbf{u}') \cos\theta \; d\sigma(\mathbf{u}'), \tag{4.88}$$

where ρ is the BRDF and θ is the angle of incidence of $\mathbf{u}'$. Now consider a simple BRDF defined in terms a Phong exponent. Let

$$\rho(\mathbf{u}' \to \mathbf{u}) \equiv c\; [\, \mathbf{u}^{\mathrm{T}} \, (\mathbf{I} - 2\mathbf{v}\mathbf{v}^{\mathrm{T}})\, \mathbf{u}' \,]^n \qquad \left[\frac{1}{\mathrm{sr}}\right], \tag{4.89}$$

where c is a constant and $\mathbf{v}$ is the surface normal. Note that the Householder matrix $\mathbf{I} - 2\mathbf{v}\mathbf{v}^{\mathrm{T}}$ performs a reflection through the tangent plane at $\mathbf{r}$. This BRDF defines a cosine lobe about an axis in the direction of mirror reflection, as shown in Figure 4.16. Because ρ obeys the reciprocity relation $\rho(\mathbf{u}' \to \mathbf{u}) = \rho(\mathbf{u} \to \mathbf{u}')$, the radiance reflected in the direction $-\mathbf{u}'$ can be found by integrating over the distribution shown in the figure. To obey energy conservation the constant c must be bounded by $2\pi/(n+2)$.

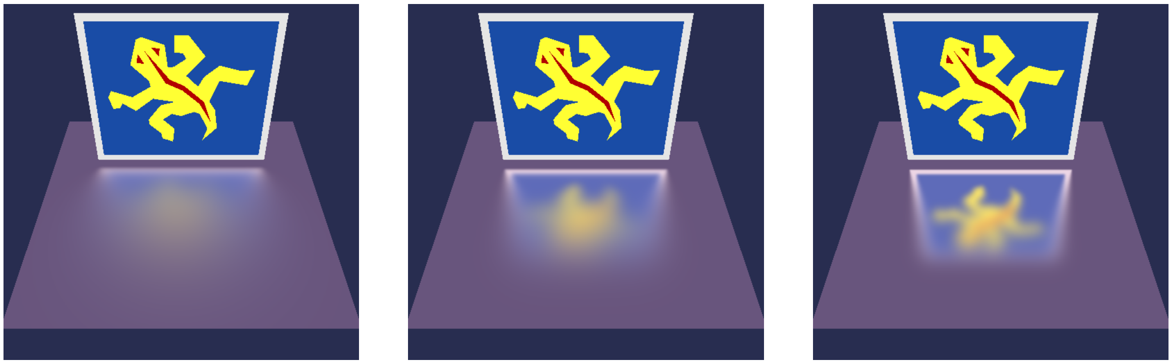

Figure 4.17: *Analytically computed glossy reflection of a convex polygon. From left to right, the moment orders are 10, 65, and 300.*

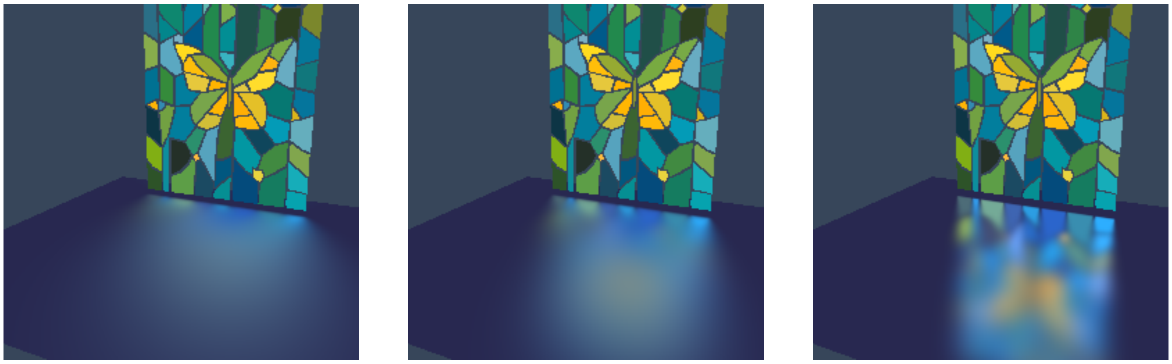

Figure 4.18: *Analytically computed glossy reflection of a stained glass window. From left to right, moment orders are 10, 45, and 400.*

When the luminaire P is Lambertian, the function $f(\mathbf{r}, \cdot)$ is constant. Therefore, the integral in equation (4.88) reduces to a double-axis moment of P with respect to the surface normal $\mathbf{v}$ and the vector

$$\mathbf{w} \equiv \left(\mathbf{I} - 2\mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\mathbf{u}'.$$

Procedure *DoubleAxisMoment* may therefore be used in this context as well, to compute the glossy reflection of a diffuse area light source. The technique is demonstrated in Figure 4.1, where the order of the moment is 500, and in Figures 4.17 and 4.18, which depict a sequence of moment orders to demonstrate surfaces with varying finishes.

More complex BRDFs may be formed by superposing lobes of different orders and/or different axes. Other effects such as anisotropic reflection and specular reflection near grazing can be simulated by allowing c and n to vary with the incident direction $\mathbf{u}'$; doing so does not alter the moment computations, however reciprocity is generally violated.

4.7.3 Glossy Transmission

As a final example, we note that glossy transmission can be handled in much the same way as glossy reflection; the difference is in the choice of the axes $\mathbf{w}$ and $\mathbf{v}$, which must now exit from the far side of the transparent material. Figure 4.17 shows a sequence of images depicting "frosted glass", with different finishes corresponding to axial moments of different orders. This effect was first demonstrated by Wallace et al. [170] using a form of stochastic sampling; in Figure 4.19 a similar effect has been computed analytically using procedure *DoubleAxisMoment*.

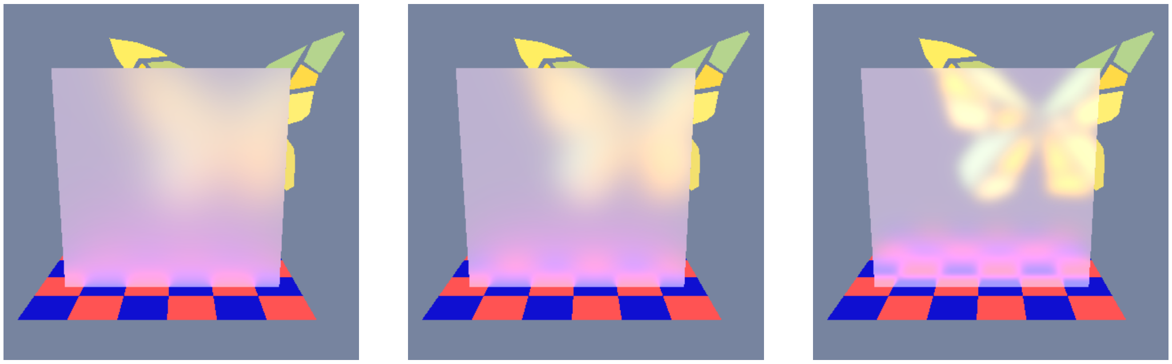

Figure 4.19: *A frosted glass simulation demonstrating glossy transmission. From left to right, the moment orders are 4, 10, and 65.*

4.8 Further Generalizations

We conclude the chapter by describing two further generalizations of the methods considered thus far. We shall see that the two most fundamental constraints can in fact be relaxed and, in some instances, still yield analytic solutions. These generalizations, which are taken up in the following two sections, are 1) luminaires with non-polygonal geometry, and 2) luminaires with spatially varying brightness.

4.8.1 Spherical Luminaires

Formulas (4.75) and (4.77) apply to arbitrary regions $A \subset \mathcal{S}^2$, and therefore to all luminaire geometries. However, when the luminaire is non-polygonal the spherical projection will generally not be bounded by great arcs, so we can no longer take advantage of the simplifications exploited earlier. In particular, the normal to the boundary ∂A will generally vary continuously, and therefore cannot be removed from the boundary integral. Nevertheless, there are instances in which closed-form solutions can still be obtained; the simplest example being spherical luminaires.

We shall consider the problem of computing the axial moment $\bar{\tau}^n(A, \mathbf{w})$ when $A \subset \mathcal{S}^2$ is the unoccluded projection of a sphere. As with polygonal luminaires, the computation of $\bar{\tau}^n(A, \mathbf{w})$ entails a sequence of integrals of the form

$$\int_{\partial A} (\mathbf{w} \cdot \mathbf{u})^n \ \mathbf{w} \cdot \mathbf{n} \, ds,$$

which follows from equation (4.75). Since the boundary ∂A is a circle, it is easily parametrized. The radius of ∂A is $\sin\alpha$, where α is the half-angle of the circular cone defined by the luminaire. To conveniently represent the path corresponding to ∂A, we introduce a set of orthogonal vectors $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ where $\mathbf{a}$ points toward the center of the spherical luminaire. Then the dot products $\mathbf{w} \cdot \mathbf{u}$ and $-\mathbf{w} \cdot \mathbf{n}$ can be expressed parametrically as

$$\begin{aligned} u(\theta) &= \mathbf{w} \cdot [\mathbf{a} \cos\alpha \ + \ (\mathbf{b} \cos\theta + \mathbf{c} \sin\theta) \sin\alpha] \,, \\ v(\theta) &= \mathbf{w} \cdot [\mathbf{a} \sin\alpha \ - \ (\mathbf{b} \cos\theta + \mathbf{c} \sin\theta) \cos\alpha] \,, \end{aligned}$$

where $0 \leq \theta \leq 2\pi$. The function v can also be expressed in terms of u as

$$v(\theta) \;=\; \frac{\mathbf{w}\cdot\mathbf{a}}{\sin\alpha} \;-\; u(\theta)\,\cos\alpha.$$

It follows that

$$\begin{aligned} -\int_{\partial A} (\mathbf{w}\cdot\mathbf{u})^n \;\mathbf{w}\cdot\mathbf{n}\; d\theta \;&=\; \sin\alpha \int_0^{2\pi} u^n(\theta)\, v(\theta)\; d\theta \\ &=\; \mathbf{w}\cdot\mathbf{a} \int_0^{2\pi} u^n \; d\theta \;-\; \cos\alpha \int_0^{2\pi} u^{n+1}\, d\theta, \qquad (4.90) \end{aligned}$$

where the additional factor of $\sin\alpha$ on the right results from re-parametrizing in terms of arclength. Thus, the normal vector $\mathbf{n}$ is incorporated into the integral rather than factored out, as in the case of polygonal luminaires. We next concentrate on evaluating the integrals on the right of equation (4.90). To simplify notation we define the following scalar quantities:

$$\begin{aligned} a \;&\equiv\; \mathbf{w}\cdot\mathbf{a}\,\cos\alpha, \\ b \;&\equiv\; \mathbf{w}\cdot\mathbf{b}\,\sin\alpha, \\ c \;&\equiv\; \mathbf{w}\cdot\mathbf{c}\,\sin\alpha. \end{aligned}$$

With these definitions, we may write

$$\begin{aligned} \int u^n\, d\theta \;&=\; \int (a + b\,\cos\theta + c\,\sin\theta)^n\, d\theta \\ &=\; \int [\; a + h\,\cos(\theta - \beta)\;]^n\, d\theta, \qquad (4.91) \end{aligned}$$

where $h \equiv \sqrt{b^2 + c^2}$, and β is the angle such that $\cos\beta = b/h$ and $\sin\beta = c/h$. When the integral is taken over the range $[0, 2\pi]$, which is true for an unoccluded sphere, we may set $\beta = 0$. Integrals of this form may be evaluated using the following lemma, which generalizes recurrence relation (4.82).

Lemma 1 *Let a and h be real numbers, let n be an integer, and define*

$$F_n \;\equiv\; \int (a + h\,\cos\theta)^n \, d\theta.$$

Then F_n satisfies the recurrence relation

$$nF_n \;=\; h(a + h\,\cos\theta)^{n-1}\sin\theta \;+\; a(2n-1)\,F_{n-1} \;+\; (n-1)(h^2 - a^2)\,F_{n-2},$$

where $n \geq 1$, $F_0 = \theta$, and $F_{-1} = 0$.

Proof: The result follows from integration by parts. See Appendix A.5.

The complete algorithm for evaluating equation (4.75) for a spherical luminaire is quite similar to the corresponding algorithm for polygons. The integral

$$-\int_{\partial A} \Big[\, (\mathbf{w}\cdot\mathbf{u})^{n-1} + (\mathbf{w}\cdot\mathbf{u})^{n-3} + \cdots + (\mathbf{w}\cdot\mathbf{u})^{q} \,\Big]\; \mathbf{w}\cdot\mathbf{n}\; ds$$

may be reduced by means of equations (4.90) and (4.91) to the expression

$$\mathbf{w}\cdot\mathbf{a}\,[\,F_{n-1} \;+\; F_{n-3} \;+\cdots+\; F_q\,] \;-\; \cos\alpha\,[\,F_n \;+\; F_{n-2} \;+\cdots+\; F_{q+1}\,],$$

where F_n is defined as in lemma 1, and $q = 1$ if n is even, and $q = 0$ if n is odd. The above expression is analogous to expression (4.83). Note that this expression incorporates all of the integrals $F_1, F_2, \ldots, F_n$, which are generated naturally by the recurrence relation in lemma 1, and that F_0 also appears in the left-hand expression when n is odd. The following pseudo-code summarizes the algorithm.

```
real AxialMoment( sphere S; vector w; integer n )
    real $\alpha \leftarrow$ half-angle of cone defined by S;
    if $n = 0$ then return $2\pi(1 - \cos\alpha)$;    Solid angle.
    real a $\leftarrow$ normalized vector toward center of S;
    real $a \leftarrow (\mathbf{w} \cdot \mathbf{a}) * \cos\alpha$;
    real $h \leftarrow \sqrt{1 - \mathbf{w} \cdot \mathbf{a}^2}\ \sin\alpha$;
    real $F_0 \leftarrow 2\pi$;
    real $F_1 \leftarrow 2\pi a$;
    real $S \leftarrow$ if even($n$) then 0 else $2\pi$;    Sum terms to $n - 1$
    real $T \leftarrow 0$;    Sum terms to $n$
    for $i \leftarrow 1, 2, 3, \ldots, n$ do
        if even($i + n$)
            then $T \leftarrow T + F_i$;
            else $S \leftarrow S + F_i$;
        $F_{i+1} \leftarrow [\, a\,(2i + 1) * F_i + i\,(h^2 - a^2) * F_{i-1} \,] \;/\; (i + 1)$;
        endfor
    real $A \leftarrow (\mathbf{w} \cdot \mathbf{a}) * S - \cos\alpha * T$;
    if even($n$) then $A \leftarrow A + 2\pi(1 - \cos\alpha)$;    Add solid angle.
    return $A \;/\; (n + 1)$;
    end
```

This procedure easily generalizes to spheres that are partially occluded by polyhedra or other spheres. In such a case, the outer contour of the visible portion of a given sphere is a collection of lines and circular arcs.

4.8.2 Spatially Varying Luminaires

This section describes several fundamental observations about the problem of computing moments of luminaires whose directional distributions are a function of *position*; we refer to such luminaires as *spatially varying*. Handling this type of luminaire is an important generalization with immediate applications to high-order finite element methods for simulating global illumination. We shall demonstrate

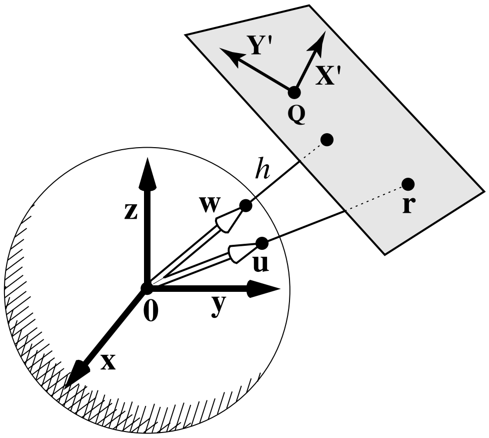

Figure 4.20: *A polynomial over the plane may be represented as a rational polynomial over the sphere.*

that 1) the problem is equivalent to integrating a class of rational functions over the sphere, 2) moments of spatially varying luminaires are interrelated via a generalization of recurrence relation (4.59), and 3) even for polygonal luminaires with polynomially varying brightness, cases arise in which no solution in terms of elementary functions is possible.

We shall only consider luminaires with polynomially varying radiant exitance; this type of variation can be accommodated by a simple generalization of irradiance tensors. In particular, we introduce tensors whose elements are a restricted form of rational polynomials integrated over regions of the sphere.

To see how rational polynomials arise in this context, consider a polygonal luminaire $P \subset \mathbb{R}^3$ in a plane not containing the origin. Suppose that the radiant exitance of P varies according to a polynomial $\phi(u, v)$ with respect to some orthogonal coordinate system $(\mathbf{X}', \mathbf{Y}')$ and origin $\mathbf{Q}$ in the plane. To compute angular moments of $f(\mathbf{0}, \cdot)$, the radiance distribution function at the origin, we first express this function in terms of direction cosines. Let $\mathbf{w} \in \mathcal{S}^2$ denote the vector orthogonal to the plane and let h denote the distance between the plane and the origin, as shown in Figure 4.20.

Let $\mathbf{r} \in \mathrm{I\!R}^3$ be an arbitrary point on the plane, and let (x', y') denote its local coordinates with respect to the imposed coordinate frame. That is,

$$\begin{aligned} x' &= (\mathbf{r} - \mathbf{Q}) \cdot \mathbf{X}', \\ y' &= (\mathbf{r} - \mathbf{Q}) \cdot \mathbf{Y}'. \end{aligned}$$

To express these coordinates in terms of direction cosines, we employ the usual notation $\mathbf{u} \equiv \mathbf{r} / \,||\, \mathbf{r} \,||$ and note that $h \;=\; ||\, \mathbf{r} \,||\;\; \mathbf{u} \cdot \mathbf{w}$. Then

$$\mathbf{r} = \frac{h}{\mathbf{u} \cdot \mathbf{w}} \, \mathbf{u}.$$

It follows that the polynomial function defined on the surface of the luminaire may be expressed as a rational polynomial over the sphere. Specifically

$$\phi(x', y') \;=\; \phi\left(h \frac{\mathbf{u} \cdot \mathbf{X}'}{\mathbf{u} \cdot \mathbf{w}} + t_x,\; h \frac{\mathbf{u} \cdot \mathbf{Y}'}{\mathbf{u} \cdot \mathbf{w}} + t_y \right), \tag{4.92}$$

where $t_x \;=\; \mathbf{Q} \cdot \mathbf{X}'$, and $t_y \;=\; \mathbf{Q} \cdot \mathbf{Y}'$. By expanding the above expression and regrouping terms, it follows that the radiance distribution function at the origin may be expresses as

$$f(\mathbf{0}, -\mathbf{u}) = \widehat{\phi}(x, y, z), \tag{4.93}$$

where x, y, and z are the direction cosines of $\mathbf{r}$, and $\widehat{\phi}$ is the rational polynomial. The form of $\widehat{\phi}$ is constrained, however, as only powers of $\mathbf{u} \cdot \mathbf{w}$ appear in the denominator. To integrate polynomials of this form over regions $A \subset \mathcal{S}^2$, such as a spherical polygon, we introduce a tensor whose elements are integrals of appropriate rational functions. We define

$$\mathbf{T}^{n,q}(A, \mathbf{w}) \equiv \int_A \frac{\mathbf{u} \otimes \cdots \otimes \mathbf{u}}{(\mathbf{u} \cdot \mathbf{w})^q} \, d\sigma(\mathbf{u}). \tag{4.94}$$

Then the integral of $\widehat{\phi}(x, y, z)$ over A, and all of its moments, can be formed from elements of the new tensors. Note that $\mathbf{T}^{n,q}$ reduces to $\mathbf{T}^n$ when $q = 0$. We shall show that the new tensors with rational elements satisfy a recurrence relation of the form

$$\mathbf{T}^{n,q} = G_{n,q}(\mathbf{T}^{n-2,q}, \mathbf{T}^{n-1,q+1}) + H_{n,q}(\partial A),$$

where all components implicitly depend on the axis $\mathbf{w}$ and the region $A \subset \mathcal{S}^2$. Equation (4.95) is thus an extension of equation (4.35).

Theorem 11 *Let $n \geq 0$ and $q \geq 0$ be integers with $n \geq q$, and let $A \subset \mathcal{S}^2$ be a measurable set. Then the tensor $\mathbf{T}^{n,q}(A, \mathbf{w})$ satisfies the recurrence relation*

$$\mathbf{T}^{n,q}_{\mathrm{I}j} = \frac{1}{n-q+1} \left(\sum_{k=1}^{n-1} \delta_{j\,\mathrm{I}_k} \mathbf{T}^{n-2,q}_{\mathrm{I}/k} \;-\; q\,\mathbf{w}_j\, \mathbf{T}^{n-1,q+1} \;-\; \int_{\partial A} \frac{\mathbf{u}^{n-1}_{\mathrm{I}}\, \mathbf{n}_j}{(\mathbf{u}\cdot\mathbf{w})^q}\, ds \right), \quad (4.95)$$

where I *is an $(n-1)$-index, ds denotes integration with respect to arclength, and* $\mathbf{n}$ *is the outward normal to the curve ∂A.*

Proof: The proof parallels that of equation (4.59). See Appendix A.6.

By theorem 11 we see that all high-order tensors $\mathbf{T}^{n,q}$ with $n > q$ can be reduced to a sequence of rational boundary integrals and lower-order tensors, with $n \leq q$. While this provides a means of reducing the order of these tensors, two problems remain: 1) computing the boundary integrals, and 2) computing the base cases. We shall pursue the second point in the remainder of this section and expose one of the difficulties associated with spatially varying luminaires.

A complication that arises in dealing with spatially varying luminaires is that integrals of rational polynomials over the sphere cannot always be reduced to elementary functions, even when the domain of integration is a spherical polygon; in this respect the problem of spatially varying luminaires is intrinsically more difficult than directional luminaires. One form of integral that arises is

$$\Upsilon(\alpha, \beta) \equiv \int_0^\alpha \log\left(1 + \beta^2 \sec^2\theta\right)\, d\theta, \quad (4.96)$$

where $0 \leq \alpha \leq \pi/2$ and $\beta \geq 0$, which has no elementary solution in general. Figure 4.21 shows the graph of $\Upsilon(\alpha, \beta)$ for several values of β. For some parameters equation (4.96) reduces immediately to known definite integrals. For instance,

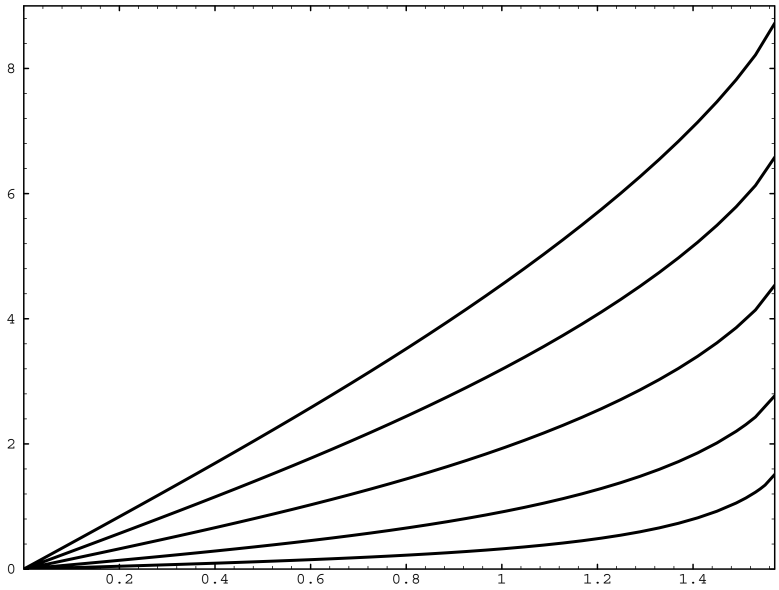

Figure 4.21: *The function* $\Upsilon(\alpha,\beta)$ *plotted as a function of* α *over the interval* $[0,\pi/2]$*. Here* $\beta = 0.5$*, 1, 2, 4, and 8, starting from the bottom curve.*

when $\alpha = \pi/2$, we may apply the formula

$$\int_0^{2\pi} \log\left(a^2\cos^2\theta \;+\; b^2\sin^2\theta\right) d\theta \;=\; \pi\,\log\left(\frac{a+b}{2}\right), \tag{4.97}$$

due to Gröbner and Hofreiter [56, vol. II, pp. 69–70], and derived in a different manner by Carlson [24], to show that

$$\Upsilon(\pi/2,\beta) = \pi\log\left(\beta + \sqrt{1+\beta^2}\right). \tag{4.98}$$

For other parameters, however, evaluation of $\Upsilon(\alpha,\beta)$ involves at least one special function. In particular, in Appendix A.7 we show that integral (4.96) can be expressed either in terms of the dilogarithm with a complex argument, or in terms of the Clausen integral [1,10,86]. It is interesting to note that the dilogarithm also appears in computing the form factor between two arbitrary polygons, as shown by Schröder and Hanrahan [140].

We now show that evaluating integral (4.96) is a prerequisite for solving the general problem of polynomially varying luminaires, and not the consequence of

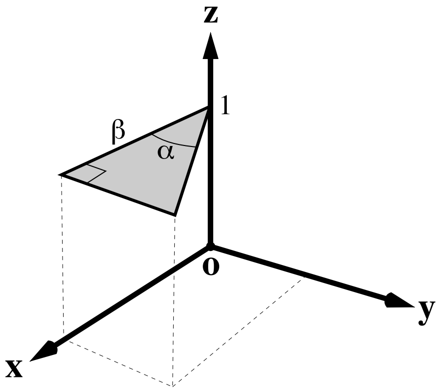

Figure 4.22: *The irradiance at the origin due to a triangular luminaire with quadratically varying radiant exitance is given by* $\Upsilon(\alpha, a)$.

a particular solution strategy. To show this, we reduce the problem of computing $\Upsilon(x, y)$ to that of computing the irradiance due to a two-parameter family of luminaires with spatially varying radiant exitance.

Let $\triangle$ denote the triangle in Figure 4.22, which has two parameters: the angle α and the edge length β. Let A denote the spherical projection of the triangle, $\mathbf{\Pi}(\triangle)$, and let $\phi(\alpha, \beta)$ denote the irradiance at the origin due to this luminaire with radiance distribution given by

$$f(\mathbf{r}, \mathbf{u}) \equiv 2\left(\mathbf{r}_x^2 + \mathbf{r}_y^2 + 1 \right).$$

Thus, the radiance varies quadratically with position, but is independent of direction. It follows that the irradiance at the origin due to this luminaire is

$$\begin{aligned} \phi(\alpha, \beta) &= \int_A f\left(\frac{\mathbf{u}}{z}, -\mathbf{u}\right) \cos\theta \, \sigma(\mathbf{u}) \\ &= 2\int_A \left[\left(\frac{x}{z}\right)^2 + \left(\frac{y}{z}\right)^2 + 1 \right] z \, \sigma(\mathbf{u}) \\ &= 2\int_A \frac{1}{z} \, \sigma(\mathbf{u}). \end{aligned} \tag{4.99}$$

The problem therefore reduces to integrating the rational function $1/z$ over the region $A \subset \mathcal{S}^2$, where A depends on the parameters α and β. To evaluate this integral, we first convert it from an integral over solid angle to an integral over the area of the luminaire. By equation (4.24) we have

$$\frac{d\omega}{\cos\theta} = \frac{1}{r^2}\, dA.$$

Using this relationship, and the fact that $z = \cos\theta$ in this configuration, the integral in equation (4.99) can be written in Cartesian form as

$$\int\!\!\int_{\triangle} \frac{2}{x^2 + y^2 + 1}\, dx\, dy. \tag{4.100}$$

Finally, changing variables once again, to polar coordinates this time, the integral reduces to $\Upsilon(\alpha, \beta)$ after simplification. We have

$$\begin{aligned} 2\int_A \frac{1}{z}\, \sigma(\mathbf{u}) &= \int_0^\alpha \int_0^{\beta \sec\theta} \frac{2}{1 + r^2}\, r\, dr\, d\theta \\ &= \int_0^\alpha \log(1 + r^2)\, \Big|_0^{\beta \sec\theta}\, d\theta \\ &= \Upsilon(\alpha, \beta). \qquad (4.101) \end{aligned}$$

Therefore, any method capable of computing irradiance due to polygonal luminaires with polynomially varying radiant exitance must also be capable of evaluating the function $\Upsilon(\alpha, \beta)$ over its entire range of parameters.

These results constitute a start toward obtaining closed-form expressions for moments of luminaires whose radiant exitance varies polynomially over the surface. However, the meaning of "closed-form" in this context must be expanded to include at least one special function, such as Clausen's integral, or the function $\Upsilon(\alpha, \beta)$.

Chapter 5

Comparison with Monte Carlo

Monte Carlo is a versatile simulation method that can be applied to virtually any problem of numerical approximation. In this chapter we describe several Monte Carlo integration methods for computing irradiance tensors and axial moments of all orders. The results are compared with those generated by the algorithms of chapter 4 to provide independent evidence of their correctness. This method of validation is appealing because the Monte Carlo algorithms are comparatively simple and rely on entirely different principles, making the chance of a systematic error remote.

A typical Monte Carlo strategy consists in converting a deterministic problem to a problem of statistical parameter estimation, such as estimating the mean value of a random variable. The effectiveness of the method hinges on the costs and statistical properties of the random variable so considerable effort is usually warranted in devising random variables with the correct mean and low variance. Monte Carlo methods are particularly well suited to problems of multi-dimensional integration, largely because of the close connection between integration and statistical expectation. Consequently, Monte Carlo integration appears throughout computer graphics; examples include estimating form factors [171], visibility [62], and illumination from complex or partially occluded luminaires [147].

We shall describe two Monte Carlo strategies for estimating irradiance tensors and double-axis moments: one strategy is based on rejection sampling and the other on stratified sampling. The stratified sampling is performed using a new algorithm for generating uniformly distributed random samples over spherical triangles, which has applications beyond those explored in this chapter.

5.1 Sampling by Rejection

The *rejection method* is the simplest and most general method for generating random samples according to a given distribution [74], and is particularly useful in dealing with irregular domains. The idea is to embed the desired distribution in a larger domain that can be sampled conveniently, and ignore samples that land outside the target domain. The remaining samples are then distributed appropriately. To obtain a Monte Carlo estimator of $\mathbf{T}^n(A)$ based on the rejection method, we observe that

$$\begin{aligned} \mathbf{T}^n(A) &= \int_A \mathbf{u}^n \, d\sigma(\mathbf{u}) \\ &= 4\pi \int_{\mathcal{S}^2} \mathcal{X}_A(\mathbf{u}) \; \mathbf{u} \otimes \cdots \otimes \mathbf{u} \; \frac{d\sigma(\mathbf{u})}{4\pi}, \end{aligned} \tag{5.1}$$

where $d\sigma/4\pi$ denotes integration with respect to the measure σ weighted by $1/4\pi$, and $\mathcal{X}_A$ is the *characteristic function* of the set A defined by

$$\mathcal{X}_A(x) \equiv \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

The measure $\sigma/4\pi$ is positive and normalized to one, since $\sigma(\mathcal{S}^2) = 4\pi$, which makes it a *probability measure*. Consequently, equation (5.1) may be viewed as the expected value of a random tensor variable. The essence of Monte Carlo lies in this shift of focus, by which equation (5.1) is given a probabilistic interpretation.

To make this interpretation precise, let $\boldsymbol{\xi}$ denote a random vector in $\mathcal{S}^2$ distributed in such a way that

$$\mathrm{Prob}[\,\boldsymbol{\xi} \in \Omega\,] = \frac{\sigma(\Omega)}{4\pi}, \tag{5.3}$$

for any subset $\Omega \subset \mathcal{S}^2$. Since the measure σ is invariant under rotations, the probability density of $\boldsymbol{\xi}$ is constant over the domain $\mathcal{S}^2$; we say that $\boldsymbol{\xi}$ is *uniformly distributed* over the sphere. The relationship between $\boldsymbol{\xi}$ and σ will be denoted by

$$\boldsymbol{\xi} \sim \frac{\sigma}{4\pi}. \tag{5.4}$$

When the random variable $\boldsymbol{\xi}$ is uniformly distributed, equation (5.1) implies that

$$\mathbf{T}^n(A) = 4\pi \left\langle \mathcal{X}_A(\boldsymbol{\xi})\, \boldsymbol{\xi} \otimes \cdots \otimes \boldsymbol{\xi} \right\rangle, \tag{5.5}$$

where $\langle\, x \,\rangle$ denotes the mean, or *expected value*, of the random variable x. The random variable $\mathcal{X}_A(\boldsymbol{\xi})\, \boldsymbol{\xi} \otimes \cdots \otimes \boldsymbol{\xi}$ is called a *primary estimator* of $\mathbf{T}^n(A)$ and may be used in constructing other estimators [60]. Obtaining a primary estimator by expanding the domain of integration, as we have done above, is the basis of the Monte Carlo rejection method. Since the estimator is zero whenever $\boldsymbol{\xi} \notin A$, it is only a crude approximation in itself, especially when $\sigma(A)$ is small.

To obtain a more desirable *secondary estimator*, one with lower variance, we define a new statistic $\boldsymbol{\beta}_1^n$ by

$$\boldsymbol{\beta}_1^n(A, m) \;\equiv\; \frac{4\pi}{m} \sum_{i=1}^{m} \mathcal{X}_A(\boldsymbol{\xi}_i)\, \boldsymbol{\xi}_i \otimes \cdots \otimes \boldsymbol{\xi}_i, \tag{5.6}$$

where $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$ are identically distributed random vectors chosen uniformly over the sphere $\mathcal{S}^2$. The new random variable is the sample mean for a sample of size m, which has the same expected value as the primary estimator. Therefore,

$$\mathbf{T}^n(A) = \langle\, \boldsymbol{\beta}_1^n(A, m)\,\rangle \tag{5.7}$$

for any $m > 0$. By the law of large numbers we have

$$\lim_{m\to\infty} \boldsymbol{\beta}_1^n(A, m) \;=\; \mathbf{T}^n(A), \tag{5.8}$$

with probability 1 whenever the variance is finite. Thus, the random vector $\boldsymbol{\beta}_1^n(A,m)$ is an unbiased estimator for $\mathbf{T}^n(A)$ the variance of which is controlled by the parameter m. In practice we use $\boldsymbol{\beta}_1^n(A,m)$ with a fixed m to estimate $\mathbf{T}^n(A)$:

$$\boldsymbol{\beta}_1^n(A,m) \approx \mathbf{T}^n(A). \tag{5.9}$$

Similarly, we may estimate the double-axis moment $\bar{\bar{\tau}}^{n,1}(A,\mathbf{w},\mathbf{v})$ directly by

$$\bar{\bar{\beta}}_1^n(A,\mathbf{w},\mathbf{v},m) \;\equiv\; \frac{4\pi}{m}\sum_{i=1}^{m} \mathcal{X}_A(\boldsymbol{\xi}_i)\,(\boldsymbol{\xi}_i\cdot\mathbf{w})^n\,(\boldsymbol{\xi}_i\cdot\mathbf{v})\,, \tag{5.10}$$

rather than estimating the corresponding tensor. The estimators $\boldsymbol{\beta}_1^n$ and $\bar{\bar{\beta}}_1^n$ are easily computed provided that we have some means of generating the random vectors $\boldsymbol{\xi}_i$.

A well-know method for simulating the random vector $\boldsymbol{\xi}$ is based on the following fact: a horizontal section through the unit sphere corresponding to $z \in [a,b] \subset [-1,1]$ has surface area $b-a$, which implies that sections of equal height have equal area. Consequently, given two random variables ξ_1 and ξ_2 that are uniformly distributed in the unit interval $[0,1]$, the three-tuple

$$\boldsymbol{\xi} \equiv \mathit{Normalize}\,(\,\cos 2\pi\xi_2,\ \sin 2\pi\xi_2,\ 1-2\xi_1\,)\,, \tag{5.11}$$

is uniformly distributed over the sphere. Here ξ_1 selects the z-coordinate uniformly from $[-1,1]$, and ξ_2 selects the rotation about the z-axis uniformly from $[0,2\pi]$. Equation (5.11) defines a transformation of the unit square $[0,1]^2$ onto the sphere that preserves uniform distributions. We construct a similar mapping for spherical triangles in the following section.

5.2 Stratified Sampling

One of the most common methods of improving the statistical efficiency of a Monte Carlo algorithm is *stratified sampling* [74]; within the computer graphics literature this technique is also known as *jitter sampling* [35]. The idea behind stratified sampling is to partition the region being sampled into disjoint subregions, or *strata*, which are then sampled independently. Used in conjunction with Monte Carlo integration, stratified sampling reduces the variance of the estimator when the integrand varies slowly over portions of the domain [61, p. 55]. Hence, Monte Carlo integration of continuous functions will generally benefit from stratification.

Let us first see how to reformulate the previous estimators when there is only one stratum, which corresponds to the entire region $A \subset \mathcal{S}^2$. In this case we have

$$\begin{aligned} \mathbf{T}^n(A) &= \sigma(A) \int_A \mathbf{u} \otimes \cdots \otimes \mathbf{u} \, \frac{d\sigma(\mathbf{u})}{\sigma(A)} \\ &= \sigma(A) \, \langle \boldsymbol{\xi}' \otimes \cdots \otimes \boldsymbol{\xi}' \rangle \, , \end{aligned} \tag{5.12}$$

with

$$\boldsymbol{\xi}' \sim \frac{\sigma|_A}{\sigma(A)}, \tag{5.13}$$

where $\sigma|_A$ is the measure σ restricted to A. Thus, $\boldsymbol{\xi}'$ is uniformly distributed over the region $A \subset \mathcal{S}^2$. This leads to the new secondary estimators

$$\begin{aligned} \boldsymbol{\beta}_2^n(A, m) &\equiv \frac{\sigma(A)}{m} \sum_{i=1}^{m} \boldsymbol{\xi}_i' \otimes \cdots \otimes \boldsymbol{\xi}_i', \\ \bar{\bar{\beta}}_2^n(A, \mathbf{w}, \mathbf{v}, m) &\equiv \frac{\sigma(A)}{m} \sum_{i=1}^{m} (\boldsymbol{\xi}_i' \cdot \mathbf{w})^n \, (\boldsymbol{\xi}_i' \cdot \mathbf{v}) \, , \end{aligned} \tag{5.14}$$

where $\boldsymbol{\xi}_1', \boldsymbol{\xi}_2', \ldots, \boldsymbol{\xi}_m'$ are independent random vectors distributed uniformly over the set A. The estimators $\boldsymbol{\beta}_3^n(A, m)$ and $\bar{\bar{\beta}}_3^n(A, \mathbf{w}, \mathbf{v}, m)$, which are based on m strata, are defined similarly:

$$\begin{aligned} \boldsymbol{\beta}_3^n(A, m) &\equiv \sum_{i=1}^{m} \sigma(A_i) \, \boldsymbol{\xi}_i'' \otimes \cdots \otimes \boldsymbol{\xi}_i'', \\ \bar{\bar{\beta}}_3^n(A, \mathbf{w}, \mathbf{v}, m) &\equiv \sum_{i=1}^{m} \sigma(A_i) \, (\boldsymbol{\xi}_i'' \cdot \mathbf{w})^n \, (\boldsymbol{\xi}_i'' \cdot \mathbf{v}) \, , \end{aligned} \tag{5.15}$$

where $A_1, A_2, \ldots, A_m$ is a partition of the domain A, and the random vector $\boldsymbol{\xi}_i''$ is uniformly distributed over the ith subregion of A. That is,

$$\boldsymbol{\xi}_i'' \sim \frac{\sigma|_{A_i}}{\sigma(A_i)}. \tag{5.16}$$

If the subregions are reasonably shaped, meaning convex with aspect ratios near one, then the samples will be more evenly spread over the domain. This property significantly reduces the clumping that inevitably occurs with uniform sampling and generally reduces the variance of estimators based on the samples. Of course, this method can only be applied if we have a means of generating samples from the given strata. In the case of $\boldsymbol{\beta}_3^n$ and $\bar{\bar{\beta}}_3^n$, this requires sampling regions of the sphere. We now address this problem for the special case of spherical triangles, which can be applied to arbitrary spherical polygons.

5.2.1 An Algorithm for Sampling Spherical Triangles

The general problem of generating uniformly distributed samples over a given n-dimensional domain may be solved by constructing a one-to-one mapping from a rectangular domain onto the given domain that preserves uniform distributions. Such a mapping $f : [0,1]^n \to X$ must have the following property: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two subsets of $[0,1]^n$ with equal volumes, then $f(\mathcal{S}_1)$ and $f(\mathcal{S}_2)$ are subsets of X with equal volumes. Obtaining a mapping with this property is generally difficult because it involves the inversion of one or more cumulative marginal density functions; a step that frequently entails numerical root-finding [74].

Given such a mapping, however, we may then generate the samples over the simple domain and then transform them to the complex domain X. In contrast to rejection, this method guarantees that all samples will fall within the desired region. Therefore, one advantage of this approach is that no samples are wasted. A far greater advantage is that it leads to a simple means of stratified sampling.

The mapping of the unit square onto an arbitrary spherical triangle may be constructed using elementary spherical trigonometry. Let T be the spherical triangle with area $\mathcal{A}$ and vertices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Let a, b, and c denote the edge lengths of T, and let α, β, and γ denote the three internal angles, which are the dihedral angles between the planes containing the edges. See Figure 5.1. To generate uniformly distributed samples over T we seek a bijection $f : [0,1]^2 \to \mathrm{T}$ with the property described above. The function f can be derived using standard Monte Carlo methods for sampling bivariate functions, as described by Spanier and Gelbard [153] and Rubinstein [129]. To apply these methods to sampling spherical triangles we require the following three identities:

$$\mathcal{A} \;=\; \alpha + \beta + \gamma - \pi, \tag{5.17}$$

$$\cos\gamma \;=\; -\cos\alpha\,\cos\beta \;+\; \sin\alpha\,\sin\beta\,\cos c, \tag{5.18}$$

$$\cos\beta \;=\; -\cos\gamma\,\cos\alpha \;+\; \sin\gamma\,\sin\alpha\,\cos b. \tag{5.19}$$

The first is Girard's formula, which we have already encountered in chapter 4, and the other two are spherical cosine laws for angles [19]. In all, there are six versions of the spherical law of cosines: three for angles and three for edges.

To generate a sample from T we proceed in two stages. In the first stage we randomly select a subtriangle $\widehat{\mathrm{T}} \subset \mathrm{T}$ with an area $\widehat{\mathcal{A}}$ that is uniformly distributed between 0 and the original area $\mathcal{A}$. In the second stage we randomly select a point along an edge of the new triangle. Both stages require the inversion of a probability distribution function.

The subtriangle $\widehat{\mathrm{T}}$ is formed by choosing a new vertex $\widehat{\mathbf{C}}$ on the edge between $\mathbf{A}$ and $\mathbf{C}$, and the sample point $\mathbf{P}$ is then chosen from the arc between $\mathbf{B}$ and $\widehat{\mathbf{C}}$, as shown in Figure 5.1. The point $\mathbf{P}$ is determined by finding its distance θ from $\mathbf{B}$ as well as the length of the new edge $\widehat{b}$; the values $\widehat{b}$ and θ are computed by inverting the distribution functions

$$F_1(\,\widehat{b}\,) \;\equiv\; \frac{\widehat{\mathcal{A}}}{\mathcal{A}}, \tag{5.20}$$

$$F_2(\,\theta\,|\,\widehat{b}\,) \;\equiv\; \frac{1-\cos\theta}{1-\cos\widehat{a}}, \tag{5.21}$$

where equation (5.21) is the conditional probability distribution of θ given $\widehat{b}$. Note that both $\widehat{\mathcal{A}}$ and $\widehat{a}$ are taken to be functions of $\widehat{b}$ in the equations above. Given two random variables ξ_1 and ξ_2 uniformly distributed in $[0,1]$, we first find $\widehat{b}$ by

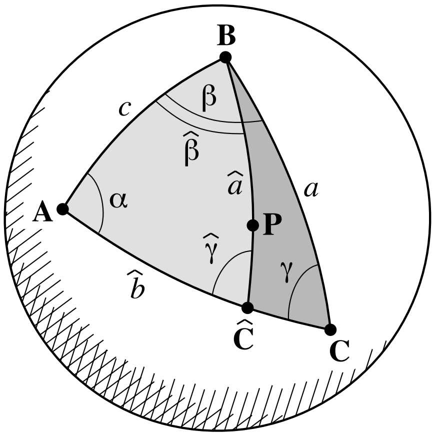

Figure 5.1: *The vertex $\widehat{\mathbf{C}}$ is determined by specifying the area of the subtriangle $\mathbf{A}\mathbf{B}\widehat{\mathbf{C}}$, and the point $\mathbf{P}$ is chosen from the arc between $\widehat{\mathbf{C}}$ and $\mathbf{B}$.*

solving $F_1(\,\widehat{b}\,) = \xi_1$, which is equivalent to $\widehat{\mathcal{A}}(\widehat{b}) = \xi_1\,\mathcal{A}$. Then $\widehat{b}$ will correspond to random subtriangles of T whose areas are uniformly distributed between 0 and $\mathcal{A}$. Having found $\widehat{b}$ in this way, the next step is to find θ such that $F_2(\,\theta\,|\,\widehat{b}\,) = \xi_2$. Then θ will be distributed along the edge $\mathbf{B}\widehat{\mathbf{C}}$ with a density that increases toward $\widehat{\mathbf{C}}$; more precisely, the density will be proportional to $(1 - \cos\theta)\,d\beta$, which is the area of the isosceles triangle with differential angle $d\beta$ and height θ.

We now carry out these two steps explicitly. To find the edge length $\widehat{b}$ that produces a subtriangle of area $\mathcal{A}\,\xi_1$, we use equations (5.17) and (5.18) to obtain

$$\cos\widehat{b} = \frac{\cos(\phi - \widehat{\beta})\,\cos\alpha \;-\; \cos\widehat{\beta}}{\sin(\phi - \widehat{\beta})\,\sin\alpha}, \tag{5.22}$$

where $\phi \equiv \widehat{\mathcal{A}} - \alpha$. Equation (5.22) completely determines $\widehat{b}$, since $0 < \widehat{b} < \pi$. From equations (5.17) and (5.19) it follows that

$$u\,\cos\widehat{\beta} \;+\; v\,\sin\widehat{\beta} = 0, \tag{5.23}$$

where

$$\begin{aligned} u &\equiv \cos(\phi) \;-\; \cos\alpha, \\ v &\equiv \sin(\phi) \;+\; \sin\alpha\,\cos c. \end{aligned}$$

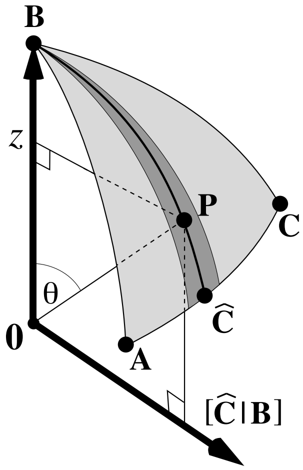

Figure 5.2: *The point* $\mathbf{P}$ *is chosen using a coordinate system with vertex* B *defining the vertical axis. The height* z *of* $\mathbf{P}$ *is uniformly distributed between* $\widehat{\mathbf{C}} \cdot \mathbf{B}$ *and 1.*

Solving equation (5.23) for the sine and cosine of $\widehat{\beta}$, we have

$$(\sin\widehat{\beta}, \cos\widehat{\beta}) = \pm\left(\frac{-u}{\sqrt{u^2+v^2}}, \frac{v}{\sqrt{u^2+v^2}}\right),$$

where the sign is determined by the constraint $0 < \widehat{\beta} < \pi$; however, the sign will be immaterial in what follows because only ratios of these quantities will be needed. Simplifying equation (5.22) using the above expressions, we obtain

$$\cos\widehat{b} = \frac{[\,v\cos\phi - u\sin\phi\,]\,\cos\alpha \,-\, v}{[\,v\sin\phi + u\cos\phi\,]\,\sin\alpha}. \tag{5.24}$$

Note that $\cos\widehat{b}$ determines $\widehat{b}$, which in turn determines the vertex $\widehat{\mathbf{C}}$.

As the final step, we select a point along the arc connecting $\widehat{\mathbf{C}}$ and $\mathbf{B}$. After choosing a convenient coordinate system, as shown in Figure 5.2, a random point is selected on the vertical ($\mathbf{B}$)-axis, with a height z uniformly distributed between $\widehat{\mathbf{C}} \cdot \mathbf{B}$ and 1. The point $\mathbf{P}$ is then constructed by projecting horizontally to the arc between $\widehat{\mathbf{C}}$ and $\mathbf{B}$, as shown in Figure 5.2. These steps correspond to solving for $z \equiv \cos\theta$ using equation (5.21) and the equation $\cos\widehat{a} = \widehat{\mathbf{C}} \cdot \mathbf{B}$. The complete algorithm is given in the next section.

5.2.2 Implementation and Results

In this section we present the complete sampling algorithm as well as the computation required for setup. First, the procedure *DefineTriangle* computes all the required quantities associated with a spherical triangle from the three vertices, which are assumed to be unit vectors.

```
DefineTriangle( vector A, B, C )

    Compute the internal angles.
        $\alpha \leftarrow \cos^{-1}\left[ \mathit{Normalize}(\mathbf{B} \times \mathbf{A}) \cdot \mathit{Normalize}(\mathbf{A} \times \mathbf{C}) \right];$
        $\beta \leftarrow \cos^{-1}\left[ \mathit{Normalize}(\mathbf{C} \times \mathbf{B}) \cdot \mathit{Normalize}(\mathbf{B} \times \mathbf{A}) \right];$
        $\gamma \leftarrow \cos^{-1}\left[ \mathit{Normalize}(\mathbf{A} \times \mathbf{C}) \cdot \mathit{Normalize}(\mathbf{C} \times \mathbf{B}) \right];$
    Compute the edge lengths.
        $a \leftarrow \cos^{-1}(\mathbf{B} \cdot \mathbf{C});$
        $b \leftarrow \cos^{-1}(\mathbf{A} \cdot \mathbf{C});$
        $c \leftarrow \cos^{-1}(\mathbf{A} \cdot \mathbf{B});$
    Compute the area of the triangle.
        $\mathcal{A} \leftarrow \alpha + \beta + \gamma - \pi;$
        end
```

We shall assume that the values computed by this procedure are available to the sampling algorithm, so *DefineTriangle* is a required pre-processing step. To succinctly express the sampling algorithm we shall let $[\,\mathbf{x} \,|\, \mathbf{y}\,]$ denote the normalized component of the vector $\mathbf{x}$ that is orthogonal to the vector $\mathbf{y}$. That is,

$$[\,\mathbf{x} \,|\, \mathbf{y}\,] \equiv \mathit{Normalize}\left(\mathbf{x} - (\mathbf{x} \cdot \mathbf{y})\mathbf{y}\right). \tag{5.25}$$

The algorithm for mapping the unit square onto the triangle T may then be summarized as shown below. The input to the procedure consists of two variables, ξ_1 and ξ_2, each in the unit interval, and the output is a point $\mathbf{P} \in \mathrm{T} \subset \mathcal{S}^2$.

```
**point** *SampleTriangle*( **real** $\xi_1$, **real** $\xi_2$ )
    *Use one random variable to select the new area.*
        $\widehat{\mathcal{A}} \leftarrow \xi_1 * \mathcal{A}$;
    *Save the sine and cosine of the angle $\phi$.*
        $s \leftarrow \sin(\widehat{\mathcal{A}} - \alpha)$;
        $t \leftarrow \cos(\widehat{\mathcal{A}} - \alpha)$;
    *Compute the pair $(u, v)$ that determines $\widehat{\beta}$.*
        $u \leftarrow t - \cos\alpha$;
        $v \leftarrow s + \sin\alpha * \cos c$;
    *Let $q$ be the cosine of the new edge length $\widehat{b}$.*
        $q \leftarrow \dfrac{[v * t \;-\; u * s] * \cos\alpha \;-\; v}{[v * s \;+\; u * t] * \sin\alpha}$;
    *Compute the third vertex of the subtriangle.*
        $\widehat{\mathbf{C}} \leftarrow q * \mathbf{A} + \sqrt{1 - q^2} * [\,\mathbf{C} \,|\, \mathbf{A}\,]$;
    *Use the other random variable to select $\cos\theta$.*
        $z \leftarrow 1 - \xi_2 * (1 - \widehat{\mathbf{C}} \cdot \mathbf{B})$;
    *Construct the corresponding point on the sphere.*
        $\mathbf{P} \leftarrow z * \mathbf{B} \;+\; \sqrt{1 - z^2} * [\,\widehat{\mathbf{C}} \,|\, \mathbf{B}\,]$;
        **return** $\mathbf{P}$;
        **end**
```

To generate uniformly distributed points over the triangle, we supply ξ_1 and ξ_2 that are independent random variables, uniformly distributed over $[0, 1]$. For example, these values might be supplied by a pseudo-random number generator, such as the algorithm proposed by Haas [58].

Note that $\cos\alpha$, $\sin\alpha$, $\cos c$, and $[\,\mathbf{C} \,|\, \mathbf{A}\,]$ in procedure *SampleTriangle* need only be computed once per triangle, not once per sample. Consequently the cost of computing these values, as well as invoking procedure *DefineTriangle*, can be amortized when many samples are to be generated from the same triangle. When few samples are required from a given triangle, the cost per sample is much higher.

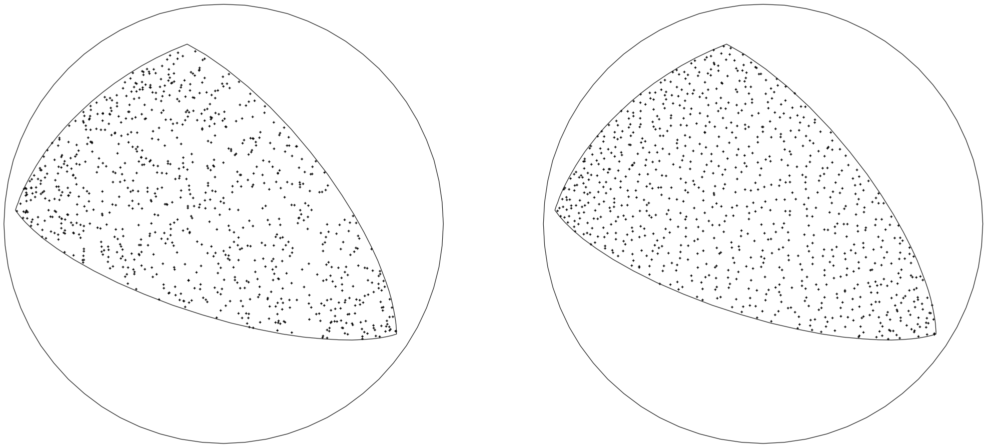

Figure 5.3: *(a) A spherical triangle sampled randomly with the new algorithm. (b) By applying the transformation to stratified samples in the unit square, we obtain stratified samples over the triangle, with much better statistical properties.*

Using procedure *SampleTriangle*, which provides a mapping of the unit square onto any given spherical triangle, it is straightforward to perform stratified sampling. This can be done simply by stratifying the input arguments $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ in the unit square. Because the mapping implemented by *SampleTriangle* is one-to-one, the stratification is carried over to the triangle T. The sampling algorithm may also be applied to spherical polygons by decomposing them into triangles and performing stratified sampling on each component independently. This strategy provides an effective means of sampling the solid angle subtended by a polygon and is analogous to the method proposed by Turk [165] for planar polygons.

Figure 5.3 shows the results of the new algorithm applied in two different ways. On the left, the samples are uniformly distributed, resulting in a pattern that is equivalent to that obtained by a rejection method. Each of the samples is guaranteed to land within the triangle, however, which is not the case with rejection. The pattern on the right was generated by first stratifying the input arguments using a regular grid, which dramatically improves the characteristics of the pattern.

5.3 Verification of Assorted Formulas

Results generated by the closed-form expressions described in chapter 4 were compared with the Monte Carlo estimators described in this chapter. In the first test, analytically computed irradiance tensors of orders 3 and 6 were compared with the corresponding Monte Carlo estimates. The tensors were computed at a sequence of 120 points along a line that was illuminated by two diffuse quadrilateral luminaires. The analytic results were compared with the Monte Carlo estimates in two ways.

Figure 5.4 shows the maximum deviation among all the tensor elements for the 6th-order tensor. The results of both uniform and stratified sampling are shown. The vertical axis is the error in the matrix element with the largest absolute deviation from the exact value, and the horizontal axis corresponds to position. The plot clearly shows the large reduction in variance from uniform sampling to stratified sampling, and also provides independent verification of the closed-form expression.

In Figures 5.5 and 5.6 a single element was chosen from each tensor and compared to the corresponding element of the Monte Carlo estimator. Figure 5.5 is the graph of xyz integrated over the solid angle subtended by the luminaires, and Figure 5.6 is a plot of $x^2y^2z^2$ over the same range of positions. Only stratified sampling is shown.

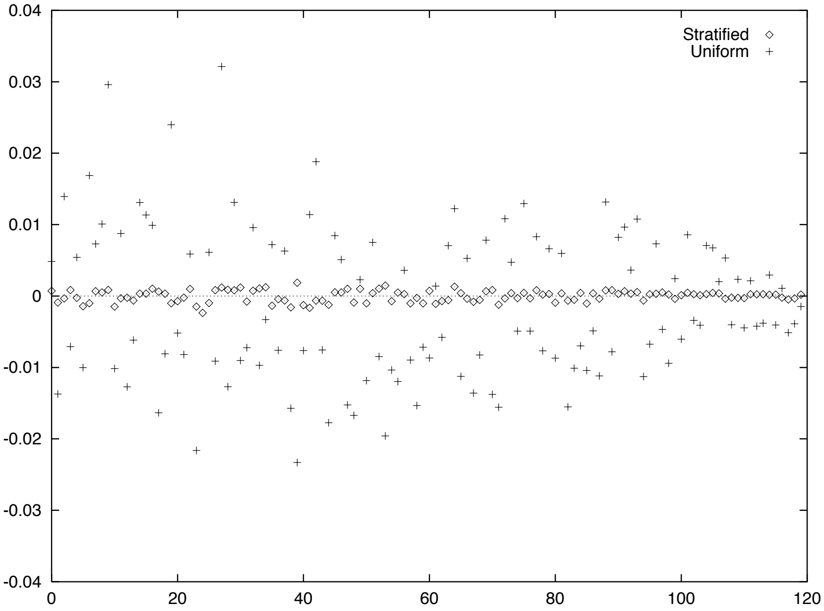

Figure 5.4: *The signed maximum deviation of the Monte Carlo estimate of* $\mathbf{T}^6(A)$ *from the analytic solution. Both stratified and uniform sampling with 400 samples are compared.*

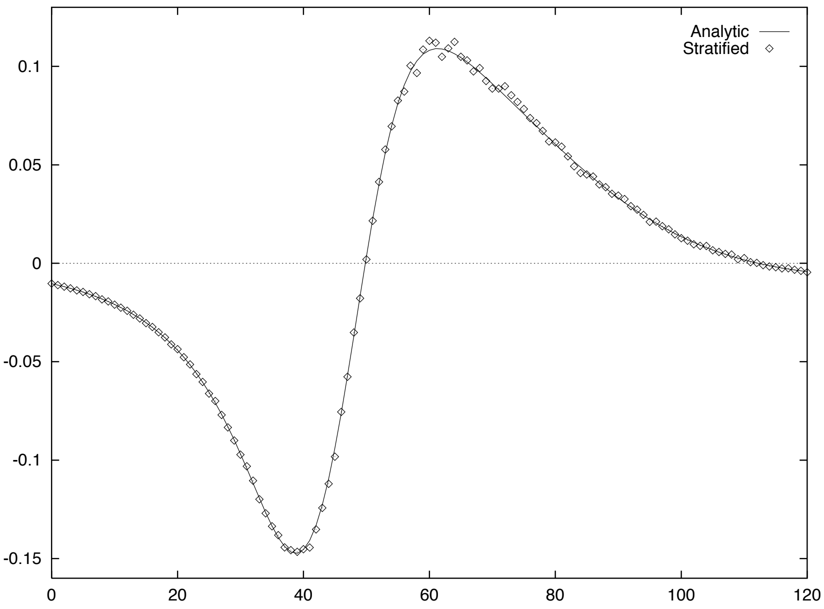

Figure 5.5: *The exact element* $\mathbf{T}^3_{123}$ *and a Monte Carlo estimate using 25 samples.*

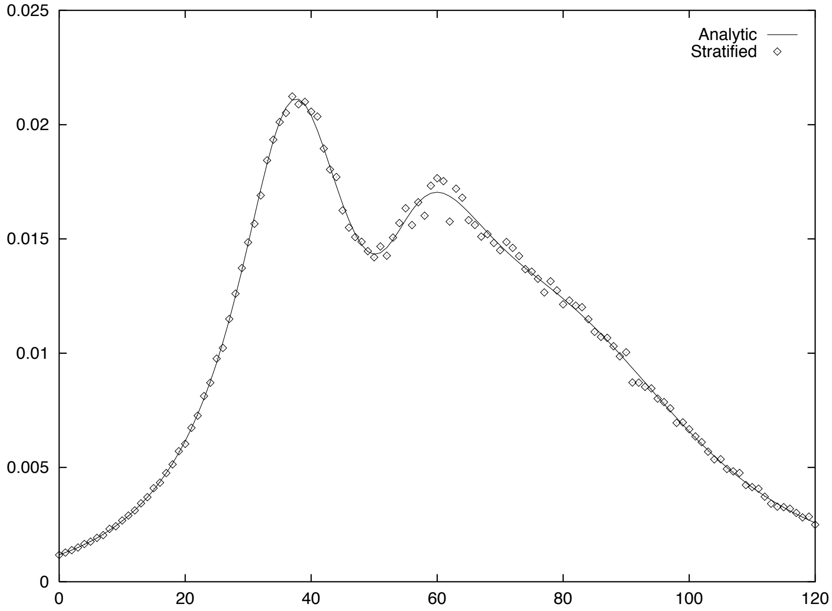

Figure 5.6: *The exact element* $\mathbf{T}^{6}_{112233}$ *and a Monte Carlo estimate using 25 samples.*

As a second test, the formulas for double-axis moments were verified using the corresponding Monte Carlo estimators. The solid line in Figures 5.7 and 5.8 is the graph of the 20th-order double-axis moment with respect to two polygons as the major axis is rotated across them.

Figure 5.7 also shows the result of an estimator based on uniform sampling at each of 200 positions of the axis, while Figure 5.8 shows the result of stratified sampling. The same scenario is shown in Figures 5.9 and 5.10 where the moment order has been increased to 401. The plots show agreement between the analytic solutions and the Monte Carlo estimators, and also demonstrate the effectiveness of stratified sampling in reducing variance.

An advantage of unbiased Monte Carlo estimates is that the independent tests taken together provide additional evidence of correctness. This follows by observing that the estimates are approximately evenly split among positive and negative deviations from the analytic solution.

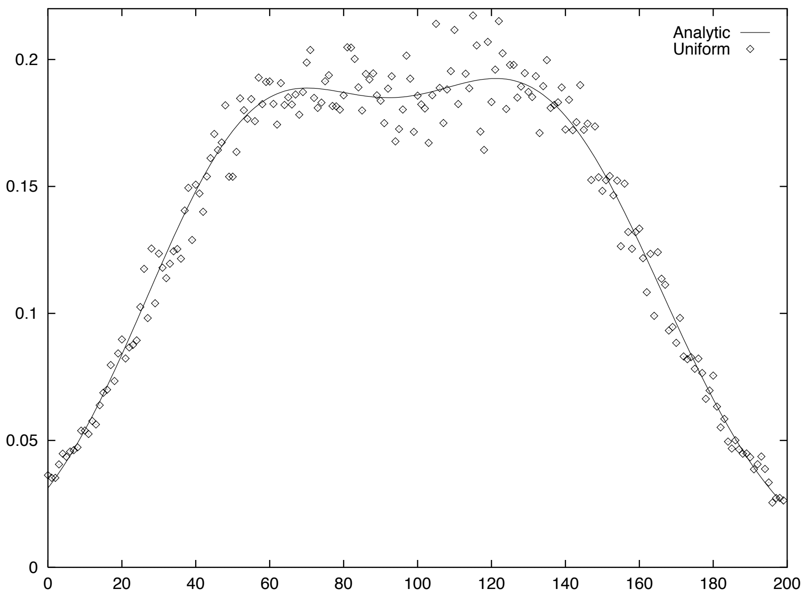

Figure 5.7: *Uniform sampling of a double-axis moment of order 20 with 100 samples.*

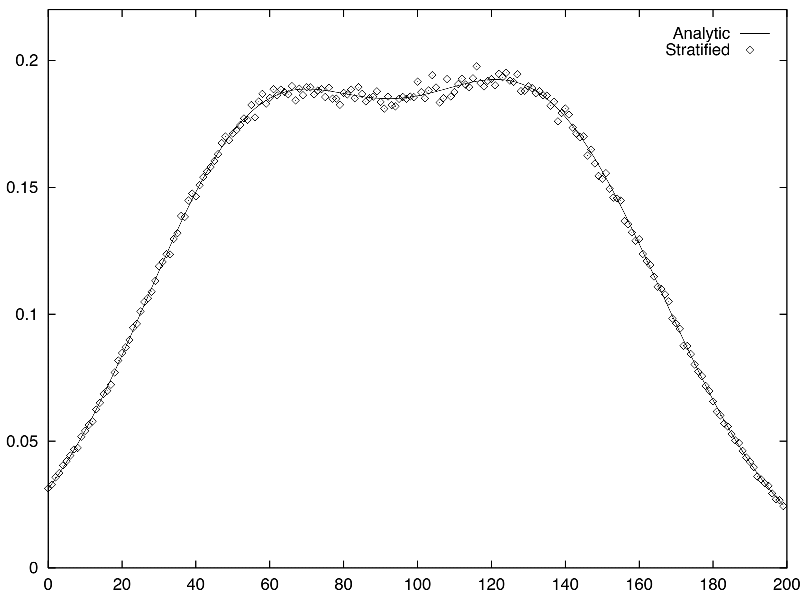

Figure 5.8: *Stratified sampling of a double-axis moment of order 20 with 100 samples.*

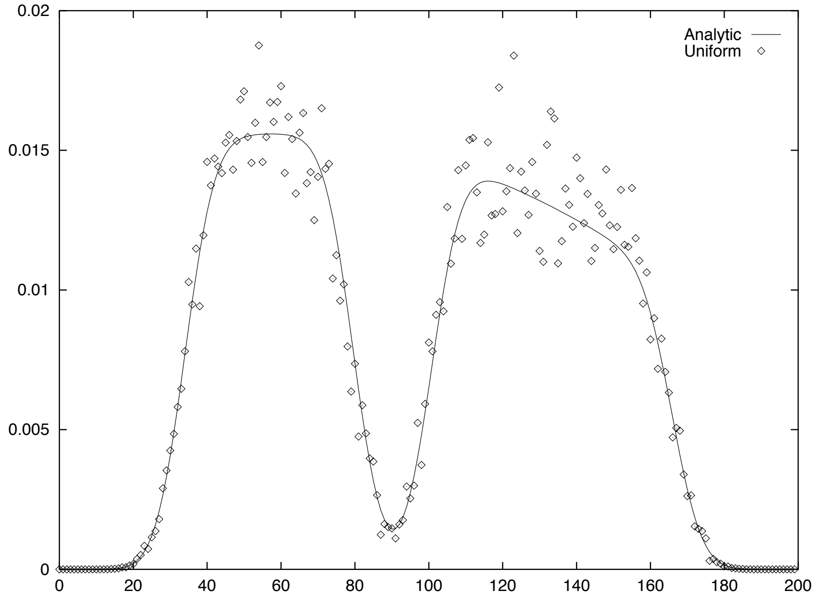

Figure 5.9: *Uniform sampling of a double-axis moment of order 401 with 900 samples.*

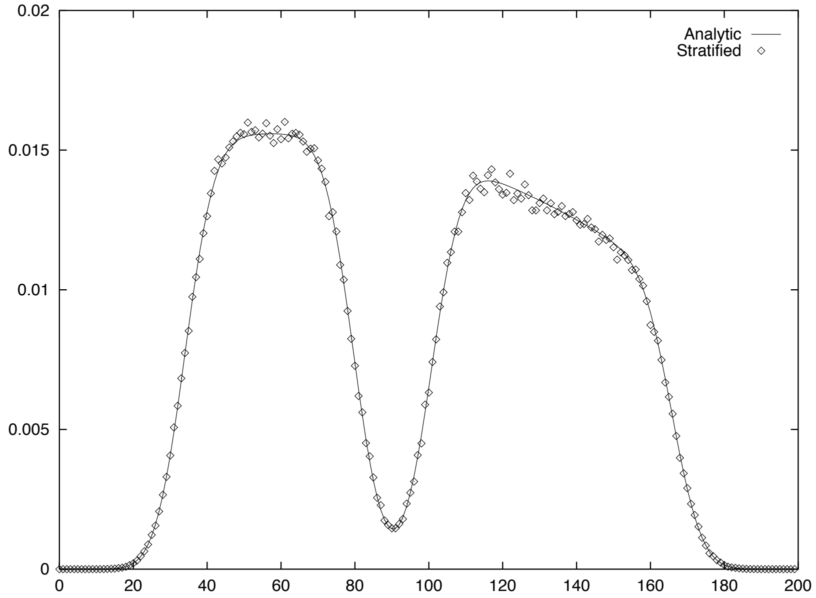

Figure 5.10: *Stratified sampling of a double-axis moment of order 401 with 900 samples.*

Chapter 6

Operators for Global Illumination

Thus far we have considered only direct illumination; that is, processes relating to light emission, shadowing, and first-order reflection. In this chapter we consider the *global illumination* problem, which includes the simulation of both direct and indirect components of illumination, and therefore accounts for interreflections among surfaces. This problem has received much attention in the fields of radiative transfer, illumination engineering, and computer graphics. Within the field of computer graphics global illumination encompasses all illumination problems that include indirect lighting, while the term *radiosity* refers to the special case of environments with only Lambertian surfaces. The latter terminology originated in engineering heat transfer, where problems of precisely the same nature arise.

The principles of geometrical optics are appropriate for simulating global illumination because the relevant interactions occur only at large scales and involve incoherent sources of light; thus, effects due to interference and diffraction are negligible. Physical or wave optics effects may enter as well, but only at the level of surface emission and scattering, which is incorporated into the boundary conditions of the global illumination problem.

Because light transport is a flow of energy, it is subject to thermodynamic constraints. In particular, the first and second laws of thermodynamics have a direct bearing on the process of light scattering. These laws require energy to be conserved and impose a reciprocity condition on surface reflectance. The reciprocity law for surface reflection can be demonstrated directly by means of a hypothetical configuration inside an isothermal enclosure [148, p. 64], or deduced as a special case of the general principle attributed to Helmholtz, which applies to all forms of electromagnetic energy propagation.

The physical principles governing global illumination can be embodied in a single equation, most commonly formulated as a linear integral equation. In this regard surface-based radiation problems differ fundamentally from those of conduction or convection, which are most naturally posed as differential equations [154]. The linearity of the equation is a consequence of the simplifying assumptions implicit in geometrical optics, and discussed in chapter 2. For instance, we neglect energy transfer between wavelength bands due to absorption and re-emission at surfaces. This assumption is valid to extremely high accuracy for illumination problems; only at very high temperatures does re-emission become significant in the visible part of the spectrum.

Equivalent formulations of the governing integral equation have appeared in numerous contexts. In illumination engineering, Moon [100, p. 328] gave a restricted version of the equation in terms of *luminosity*, the photometric counterpart of radiant exitance, which holds for diffuse environments. In thermal engineering, Polyak [117] posed the integral equation in terms of radiance and in a form that holds for arbitrary surface reflectances. Within computer graphics, Immel et al. [70] presented an equation equivalent to Polyak's, and Kajiya [71] offered a new formulation, known as the *rendering equation*, expressed in terms of *multi-point transport quantities* but otherwise equivalent to Polyak's version. Kajiya's formulation has provided a theoretical foundation for all global illumination algorithms since its introduction in 1986.

In this chapter we reformulate the governing equation for global illumination in

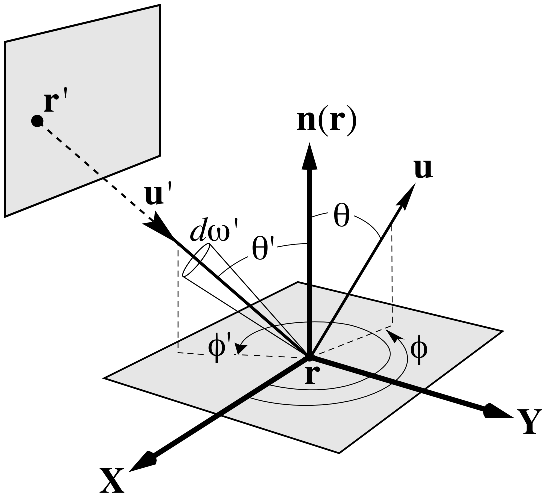

Figure 6.1: *The local coordinates of a bidirectional reflectance function. The incident vector* $\mathbf{u}'$ *and the reflected vector* $\mathbf{u}$ *are in world coordinates. The angles define a local coordinate system.*

terms of linear operators, which we then study in depth using standard techniques of functional analysis. The analysis naturally subsumes direct illumination as a special case. The new formulation clarifies the roles of various physical constraints; in particular, those of the first and second laws of thermodynamics. We identify appropriate function spaces for surface radiance functions, and show that the global solution satisfying the operator equation must lie within the same space as the surface emission function. Finally, we demonstrate that the new operator equation is equivalent to Kajiya's rendering equation and give an alternate interpretation of transport intensity, which is the quantity he used to express the equation. For simplicity, we assume monochromatic radiation throughout.

6.1 Classical Formulation

Let $\mathcal{M}$ denote a piecewise-smooth 2-manifold in $\mathbb{R}^3$ corresponding to the surfaces of an environment, and assume for now that $\mathcal{M}$ forms an enclosure. Let X denote a vector space of real-valued surface radiance functions defined on $\mathcal{M} \times \mathcal{S}^2$; that is, functions defined over all surface points and outward directions in the unit sphere $\mathcal{S}^2$. Given a *surface emission* function $f_0 \in X$, which specifies the origin and directional distribution of emitted light, the global illumination problem consists in determining the surface radiance function $f \in X$ that satisfies the linear integral equation

$$f(\mathbf{r}, \mathbf{u}) = f_0(\mathbf{r}, \mathbf{u}) + \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, f(\mathbf{r}', \mathbf{u}') \cos\theta' \, d\sigma(\mathbf{u}'), \tag{6.1}$$

where Ω_i is the hemisphere of incoming directions with respect to $\mathbf{r} \in \mathcal{M}$, k is a directional reflectivity function, θ' is the polar angle of the incoming direction vector $\mathbf{u}'$, and $\mathbf{r}'$ is a point on a distant surface determined by $\mathbf{r}$ and $\mathbf{u}'$. See Figure 6.1. This equation is essentially the formulation posed by Polyak [117], which embodies the same physical principles as Kajiya's rendering equation [71].

Equation (6.1) is essentially a Fredholm integral equation of the second kind; however it does not precisely conform to the standard definition because of the implicit function $\mathbf{r}'$, which depends on the argument $\mathbf{r}$ and the dummy variable $\mathbf{u}'$. This seemingly minor difference is responsible for the most important distinguishing feature of radiative transfer problems; the non-localness of the interactions. This feature is crucial to retain, although it is difficult to work with as it appears in equation (6.1). One solution is to rephrase the integral in equation (6.1) in terms of direct transfers among surfaces, which eliminates the implicit function. This was the approach taken by Kajiya. An alternative is to represent the influence of distant surfaces using an explicit linear operator. The latter approach has several advantages, as demonstrated below.

To precisely define the function $\mathbf{r}'$ appearing in equation (6.1), we introduce two intermediate functions similar to those employed by Glassner [47]. First, the *boundary distance function* $\nu(\mathbf{r}, \mathbf{u})$ is defined by

$$\nu(\mathbf{r}, \mathbf{u}) \;\equiv\; \inf\left\{x > 0 : \mathbf{r} + x\mathbf{u} \in \mathcal{M}\right\},$$

which is the distance from $\mathbf{r}$ to the nearest point on the surface $\mathcal{M}$ in the direction of $\mathbf{u}$. When $\mathcal{M}$ does not form an enclosure, it is possible that no such point exists; in such a case $\nu(\mathbf{r},\mathbf{u}) = \infty$ by definition. Next, we define the closely related *ray casting function* $\mathbf{p}(\mathbf{r},\mathbf{u})$ by

$$\mathbf{p}(\mathbf{r},\mathbf{u}) \;\equiv\; \mathbf{r} + \nu(\mathbf{r},\mathbf{u})\mathbf{u},$$

which is the point of intersection with the surface $\mathcal{M}$ when one exists; the function is undefined if no such point exists. Thus, we have

$$\mathbf{r}'(\mathbf{r},\mathbf{u}') \;=\; \mathbf{p}(\mathbf{r},-\mathbf{u}').$$

The kernel function k appearing in equation (6.1) is the bidirectional reflectance distribution function (BRDF) at each surface point phrased in terms of world-space direction vectors rather than angles, as it more commonly appears. That is,

$$k(\mathbf{r};\mathbf{u}' \to \mathbf{u}) \equiv \rho_{\mathbf{r}}(\theta',\phi',\theta,\phi), \tag{6.2}$$

where $\mathbf{u}'$ and $\mathbf{u}$ denote incident and reflected directions respectively, and ρ is the conventional radiometric definition of a BRDF [148,99], which is expressed in terms of four angles relative to a local coordinate system at the point $\mathbf{r}$, as shown in Figure 6.1. Here (θ',ϕ') and (θ,ϕ) denote the *polar* and *azimuth* angles of the incident and reflected directions respectively. By definition k is insensitive to the sign of its vector arguments; the angles formed with respect to the local coordinate system are the same whether the vectors are directed toward or away from the surface. This convention is convenient for opaque surfaces, where there is no need to disambiguate two distinct incident hemispheres.

Note that the BRDF may be an arbitrary function of position on $\mathcal{M}$, corresponding to changing materials or surface properties, and that all scattering is assumed to take place at the surface. The latter assumption is an idealization since the physical process of reflection may involve some amount of sub-surface scattering [115]. Consequently, the points of incidence and reflection need not coincide for real materials. We shall ignore this effect here.

The function k has two crucial properties that follow from the thermodynamic principles of energy conservation and reciprocity. By conservation we have

$$\int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u}) \cos\theta \, d\sigma(\mathbf{u}) \leq 1, \tag{6.3}$$

where $\mathbf{u}' \in \Omega_i$, and Ω_o is the outgoing hemisphere. Both Ω_i and Ω_o implicitly depend on the surface point $\mathbf{r}$. Equation (6.3) states that the energy reflected from a surface cannot exceed that of the incident beam. Reciprocity states that

$$k(\mathbf{r}; \mathbf{u}' \to \mathbf{u}) = k(\mathbf{r}; \mathbf{u} \to \mathbf{u}') \tag{6.4}$$

for all $\mathbf{u}' \in \Omega_i$ and $\mathbf{u} \in \Omega_o$. These facts play a major role in the following analysis.

Note that the implicit function $\mathbf{r}'$ in equation (6.1) is the means of constructing the distribution of energy impinging on a surface from the distribution of energy leaving distant surfaces; that is, it constructs local field radiance from distant surface radiance. The connection afforded by $\mathbf{r}'(\mathbf{r}, \mathbf{u}')$ corresponds to the intuitive notion of tracing a ray from $\mathbf{r}$ in the direction $-\mathbf{u}'$. This simple coupling is a consequence of steady-state radiance being invariant along rays in free space; in the presence of participating media the coupling is replaced by the equation of transfer along the ray [26], which turns equation (6.1) into an integro-differential equation [6].

Equation (6.1), which first appeared in this general form in thermal engineering, is a continuous balance equation governing the direct exchange of monochromatic radiant energy among surfaces with arbitrary reflectance characteristics. In current radiative heat transfer literature, this governing equation is used only when participating media are ignored [110,99]. In current computer graphics literature, this equation and its equivalent formulations are the foundation for most physically-based rendering algorithms.

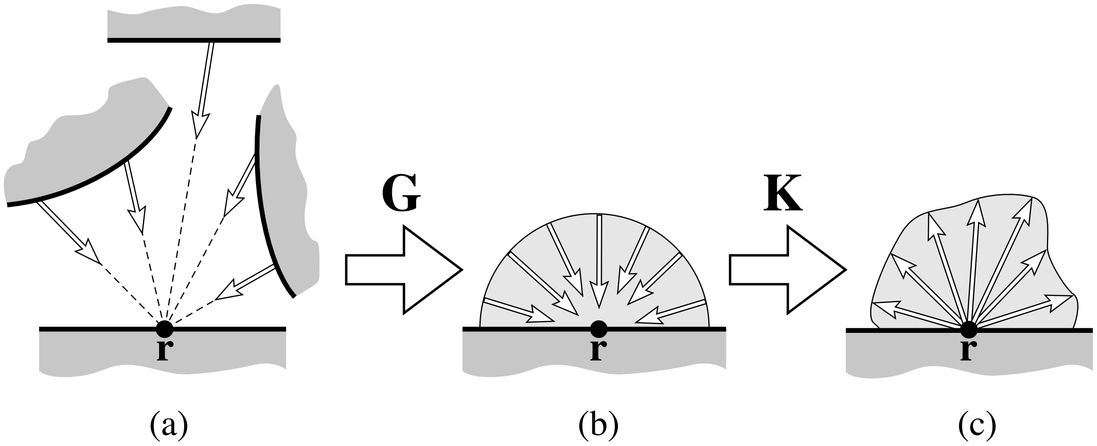

Figure 6.2: *The actions of* **G** *and* **K** *at a single point* **r**. *The operator* **G** *converts (a) distant surface radiance directed toward* **r** *into (b) local field radiance, where (c) it is again mapped into surface radiance by* **K**.

6.2 Linear Operator Formulation

Integral equations such as equation (6.1) can be expressed more abstractly as *operator equations.* Operator equations tend to be more concise than their integral counterparts while capturing essential algebraic properties such as linearity and associativity of composition, as well as topological properties such as boundedness or compactness [76,159]. The abstraction afforded by operators is appropriate when the emphasis is on integrals as transformations rather than on the numerical aspects of integration. Operator equations were first applied to global illumination by Kajiya [71] although their connection with integral equations in general has been studied for nearly a century [17,15].

Equation (6.1) can be expressed as an operator equation in numerous ways. Here we shall construct the central operator from two simpler ones, each motivated by fundamental radiometric concepts. The representation given here has two novel features: first, it cleanly separates notions of geometry and reflection into distinct linear operators, and second, it employs an integral operator with a cosine-weighted measure. Both of these features simplify the subsequent analysis.

We first define the *local reflection operator*, denoted by $\mathbf{K}$, which is an integral operator whose kernel k accounts for the scattering of incident radiant energy at surfaces. This operator is most easily defined by specifying its action on an arbitrary field radiance function h:

$$(\mathbf{K}h)(\mathbf{r}, \mathbf{u}) \equiv \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, h(\mathbf{r}, \mathbf{u}')\, d\mu(\mathbf{u}'). \tag{6.5}$$

Equation (6.5) is essentially the definition of a *kernel operator* [114], with the minor difference that the position $\mathbf{r}$ acts as a parameter of the kernel; consequently, the integration is over a proper subset of the domain of h. The $\mathbf{K}$ operator maps the field radiance function h to the surface radiance function after one reflection (Figures 6.2b-c). The operation is *local* in that the transformation occurs at each surface point in isolation of the others. In equation (6.5) the measure-theoretic notation allows us to introduce a new measure μ that incorporates the cosine weighting. The measures μ and σ are related by

$$\mu_{\mathbf{r}}(E) \equiv \int_E |\mathbf{u} \cdot \mathbf{n}(\mathbf{r})|\ d\sigma(\mathbf{u}), \tag{6.6}$$

which holds for any measurable $E \subset \mathcal{S}^2$; henceforth, the dependence on the position $\mathbf{r}$ will be implicit. The new notation eliminates the proliferation of cosine factors that appear in radiative transfer computations and, more importantly, emphasizes that the cosine is an artifact of surface integration. By associating the cosine with the integral operator and not the kernel, the function k retains the reciprocity property of reflectance functions. Moreover, this observation is vital in defining appropriate norms and inner products for radiance functions.

Next, we define the *field radiance operator*, denoted by $\mathbf{G}$, which is a linear operator expressing the incident field radiance at each point in terms of the surface radiance of the surrounding environment (Figures 6.2a-b). Showing the action of $\mathbf{G}$ on a surface radiance function h, we have

$$(\mathbf{G}h)(\mathbf{r}, \mathbf{u}) \equiv \begin{cases} h(\mathbf{p}(\mathbf{r}, -\mathbf{u}), \mathbf{u}) & \text{when } \nu(\mathbf{r}, \mathbf{u}) < \infty \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

The $\mathbf{G}$ operator expresses the transport of radiant energy from surface to surface as a linear transformation on the space of surface radiance functions. The symbol chosen for this operator is intended as a mnemonic for "global" or "geometry". Defining $\mathbf{G}$ in this way allows us to factor out the implicit function $\mathbf{r}'(\mathbf{r}, \mathbf{u})$ from the integral in equation (6.1). It is easily verified that $\mathbf{KG}$ is equivalent to the original integral and that both $\mathbf{K}$ and $\mathbf{G}$ are linear. Therefore, we may write equation (6.1) as

$$f = f_0 + \mathbf{KG}\, f, \tag{6.8}$$

which is a linear operator equation of the *second kind.* This formulation retains the full generality of equation (6.1), yet is more amenable to some types of analysis. To complete the definition of equation (6.8) it is necessary to specify the function space X over which the operators are defined. We return to this point in section 6.3.

To emphasize the linear relationship between surface radiance and surface emission, equation (6.8) can be written more concisely as

$$\mathbf{M}\, f = f_0, \tag{6.9}$$

where the linear operator $\mathbf{M}$ is defined by

$$\mathbf{M} \equiv \mathbf{I} - \mathbf{KG}, \tag{6.10}$$

and $\mathbf{I}$ is the identity operator. The operator $\mathbf{M}$ embodies all information about surface geometry and surface reflectance. However, since $\mathbf{G}$ is determined by the manifold $\mathcal{M}$, a global illumination problem is completely specified by the 3-tuple $(\mathcal{M}, \mathbf{K}, f_0)$ of surfaces, local reflection operator, and surface emission function.

In the following sections we explore fundamental properties of the operators defined above. In chapter 7 these properties will be used to study numerical methods for solving equation (6.9) and identify sources of error.

6.3 Normed Linear Spaces

In this section we introduce the basic tools of analysis that will be used in deriving error bounds for approximate solutions for global illumination. Quantifying error requires a notion of "distance" between the members of the space X, which implies a metric of some form. We shall impose metric properties on X by making it a *normed linear space*, denoted by the ordered pair $(X, ||\cdot||)$, where $||\cdot||$ is a real-valued norm defined on X. We summarize a few of the essential properties of normed linear spaces that pertain to radiative transfer. More complete treatments are given by Rudin [131] and Kantorovich [75], for example.

6.3.1 Function Norms

Infinitely many norms can be defined on a space of radiance functions, each imposing a different topology on the space as well as a different notion of size, distance, and convergence. Since equation (6.8) requires only that radiance functions be integrable with respect to the kernel k in order to be well defined, it is natural to consider the L_p-norms and their corresponding function spaces. The appropriate definition for the L_p-norm of a radiance function is

$$|| f ||_p \equiv \left[\int_{\mathcal{M}} \int_{\mathcal{S}^2} | f(\mathbf{r}, \mathbf{u})|^p \; d\mu(\mathbf{u}) \; dm(\mathbf{r}) \right]^{1/p}, \tag{6.11}$$

where m denotes area measure, and p is a real number in $[1, \infty]$. Here f will typically denote either a surface or field radiance function, which is zero over one hemisphere. Definition (6.11) is also meaningful for $p < 1$, but the corresponding norms are not *strictly convex* and are normally excluded [83]. The function space $L_p(\mathcal{M} \times \mathcal{S}^2, m \times \sigma)$ is then defined to be the set of measurable functions over $\mathcal{M} \times \mathcal{S}^2$ with finite L_p-norms. Three of the L_p-norms are of particular interest. The L_1 and L_∞-norms have immediate physical interpretations, while the L_2-norm possesses algebraic properties that make it appropriate in some instances; for example, when the structure of a Hilbert space is required to define projection-based methods.

In the limiting case of $p = \infty$, the L_p-norm reduces to

$$\| f \|_\infty \equiv \operatorname*{ess\ sup}_{\mathbf{r} \in \mathcal{M}} \; \operatorname*{ess\ sup}_{\mathbf{u} \in \mathcal{S}^2} \; | f(\mathbf{r}, \mathbf{u}) | , \tag{6.12}$$

where " ess sup" is the *essential supremum*; that is, the least upper bound obtainable by ignoring a subset of the domain with measure zero [159]. More precisely, if h is a real function defined on a set A, then

$$\operatorname*{ess\ sup}_{x \in A} h(x) \;\equiv\; \inf \{ \, m \mid h(x) \le m \;\text{ for almost every }\; x \in A \, \} . \tag{6.13}$$

Thus, the L_∞-norm ignores isolated maximal points, for example. As the *maximum radiance* attained (or approached) over all surface points and in all directions, $\| f \|_\infty$ has the dimensions of radiance [watts/m^{2}sr]. In contrast, $\| f \|_1$ is the *total power* of the radiance function f, and consequently has the dimensions of power [watts]. The L_1-norm is related to vector irradiance by

$$\| f \|_1 = \int_{\mathcal{M}} | \, \Phi(\mathbf{r}) \cdot \mathbf{n}(\mathbf{r}) \, | \; dm(\mathbf{r}), \tag{6.14}$$

where $\Phi(\mathbf{r})$ is the vector irradiance at $\mathbf{r}$ due to the the radiance distribution function $f(\mathbf{r}, \cdot)$, and $\mathbf{n}(\mathbf{r})$ is the surface normal at $\mathbf{r}$.

To assign a meaning to the distance between two radiance functions f and f' in a normed linear space X, we may use the induced metric $\| f - f' \|$. Similarly, we may define the distance between a function and a subspace $Y \subset X$ by

$$\operatorname{dist}(f, Y) \equiv \inf_{f' \in Y} \| f - f' \| , \tag{6.15}$$

where "inf" denotes the greatest lower bound. When the subscript on the norm is omitted, it implies that the definition or relation holds for any choice of norm.

6.3.2 Operator Norms

To investigate the effects of perturbing the operators, as we shall do in the following chapter, we must also endow the linear space of operators with a norm. If X and

Y are normed linear spaces and $\mathbf{A} : X \to Y$ is a linear operator, then the *operator norm* of $\mathbf{A}$ is defined by

$$||\,\mathbf{A}\,|| \equiv \sup\,\{\,||\,\mathbf{A}h\,||\,:\,||\,h\,|| \leq 1\,\}\,, \tag{6.16}$$

where the norms appearing on the right are those associated with Y and X, respectively. The operator norm is said to be *induced* by the function norms. Although the theory of linear operators closely parallels that of matrices, there are important differences; for instance, matrix norms are necessarily finite while operator norms need not be. We therefore distinguish the class of *bounded* operators as those with finite norm.

Equation (6.16) implies that $||\,\mathbf{A}h\,|| \leq ||\,\mathbf{A}\,||\;||\,h\,||$ for all $h \in X$. Operator norms also satisfy $||\,\mathbf{AB}\,|| \leq ||\,\mathbf{A}\,||\;||\,\mathbf{B}\,||$ whenever the composition $\mathbf{AB}$ is meaningful; this property makes operator norms compatible with the multiplicative structure of operators [75]. Additional bounds pertaining to inverse operators can be deduced from these basic properties. For instance, given a bounded operator $\mathbf{A}$ with an inverse, any operator $\mathbf{B}$ sufficiently close to $\mathbf{A}$ is also invertible, with

$$||\,\mathbf{B}^{-1}\,|| \leq \frac{||\,\mathbf{A}^{-1}\,||}{1 - ||\,\mathbf{A} - \mathbf{B}\,||\;||\,\mathbf{A}^{-1}\,||}. \tag{6.17}$$

This holds whenever $||\,\mathbf{A} - \mathbf{B}\,|| < 1/||\,\mathbf{A}^{-1}\,||$. Inequality (6.17) is known as Banach's lemma [109, p. 32]. A useful corollary of this result is the inequality

$$||\,\mathbf{A}^{-1} - \mathbf{B}^{-1}\,|| \leq \frac{||\,\mathbf{A} - \mathbf{B}\,||\;||\,\mathbf{A}^{-1}\,||^2}{1 - ||\,\mathbf{A} - \mathbf{B}\,||\;||\,\mathbf{A}^{-1}\,||}, \tag{6.18}$$

which holds under the same conditions as inequality (6.17) [77, p. 31].

6.3.3 Norms of Special Operators K, G, and $\mathbf{M}^{-1}$

We now compute bounds for the operators $\mathbf{K}$, $\mathbf{G}$, and $\mathbf{M}^{-1}$, which will be essential in deriving all subsequent bounds. From definitions (6.5), (6.11), and (6.16), we may deduce an explicit formula for $||\,\mathbf{K}\,||_1$. A straightforward computation yields

$$||\,\mathbf{K}\,||_1 = \operatorname*{ess\,sup}_{\mathbf{r}\in\mathcal{M}}\;\operatorname*{ess\,sup}_{\mathbf{u}'\in\Omega_i}\int_{\Omega_o} k(\mathbf{r};\mathbf{u}' \to \mathbf{u})\;d\sigma(\mathbf{u}). \tag{6.19}$$

This norm is the least upper bound on the directional-hemispherical reflectivity of any surface in the environment, disregarding isolated points and directions, and other surfaces of measure zero. Equation (6.3) guarantees that $\| \mathbf{K} \|_1$ is bounded above by one in any physically realizable environment. If we disallow perfect reflectors, as well as sequences of materials approaching them, and ignore the wave optics effect of specular reflection near grazing, the norm has a bound strictly less than one; that is,

$$\| \mathbf{K} \|_1 \;=\; \omega \;<\; 1. \tag{6.20}$$

When the bound is less than one it is possible to use a Neumann series to derive a number of related bounds that would otherwise be difficult to obtain. Therefore, the motivation for imposing the above restrictions is convenience of analysis rather than physical constraints. Note that when transparent surfaces are involved, the phenomenon of total internal reflection must also be ignored to maintain a bound that is less than one.

The L_∞-norm of $\mathbf{K}$ can be computed in a similar fashion. In precise analogy with matrix norms, the L_∞-norm merely exchanges the roles of the two directions, which are the arguments of the kernel:

$$\| \mathbf{K} \|_\infty = \operatorname*{ess\,sup}_{\mathbf{r} \in \mathcal{M}} \; \operatorname*{ess\,sup}_{\mathbf{u} \in \Omega_o} \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \rightarrow \mathbf{u}) \, d\sigma(\mathbf{u}'). \tag{6.21}$$

See Appendix A.8 for complete derivations of both of these norms. By the reciprocity relation shown in equation (6.4) we have $\| \mathbf{K} \|_1 = \| \mathbf{K} \|_\infty$, which implies that reflected radiance is everywhere diminished by at least as much as the total power. It is interesting to note that the L_∞ bound on $\mathbf{K}$ implies that it is thermodynamically impossible for a passive optical system to increase the radiance of its input; this fact was previously proven by Milne using a very different argument [98].

The two bounds on $\mathbf{K}$ given above can be extended to a bound on $\| \mathbf{K} \|_p$ for all $1 < p < \infty$. The connection is provided by the following theorem.

Theorem 12 *If* $\mathbf{A}$*is a kernel operator, then* $\| \mathbf{A} \|_p \leq \max \{ \| \mathbf{A} \|_1, \| \mathbf{A} \|_\infty \}$.

Proof: See Appendix A.9.

Although theorem 12 applies to standard kernel operators it can be extended slightly to accommodate the local reflection operator $\mathbf{K}$, which is kernel operator with an extra parameter. We state the resulting bound as a theorem.

Theorem 13 *If* $\mathbf{K}$ *is a physically realizable local reflection operator for an environment with directional-hemispherical reflectance bounded away from 1, then*

$$\| \mathbf{K} \|_p \leq \omega \tag{6.22}$$

for all $1 \leq p \leq \infty$*, where* $\omega < 1$.

Proof: See Appendix A.10.

The field radiance operator $\mathbf{G}$ is also bounded, but we shall arrive at its norm in an entirely different manner. To study various properties of this operator and to compute its norm, we first prove three elementary identities, which are collected in the following lemma.

Lemma 2 *Let* f *and* g *be surface radiance functions and let* $f \cdot g$ *denote their pointwise product. Then the field radiance operator* $\mathbf{G}$ *satisfies*

$$\mathbf{G}(f \cdot g) = (\mathbf{G}f) \cdot (\mathbf{G}g) \tag{6.23}$$

$$\mathbf{G}\, f^p = (\mathbf{G}\, f)^p \tag{6.24}$$

$$\mathbf{G}\, |f| = |\, \mathbf{G}\, f \,| \tag{6.25}$$

where p *is a positive integer and* $|f|$ *denotes the pointwise absolute value.*

<u>Proof:</u> To prove equation (6.23), let $\mathbf{r} \in \mathcal{M}$ and $\mathbf{u} \in \mathcal{S}^2$ and suppose that $\mathbf{r}' = \mathbf{p}(\mathbf{r}, \mathbf{u})$ is defined. Then

$$\begin{aligned} [\mathbf{G}(\, f \cdot g\,)]\,(\mathbf{r}, \mathbf{u}) &= (f \cdot g)(\mathbf{r}', -\mathbf{u}) \\ &= f(\mathbf{r}', -\mathbf{u})\, g(\mathbf{r}', -\mathbf{u}) \\ &= [\mathbf{G}f(\mathbf{r}, \mathbf{u})]\,[\mathbf{G}g(\mathbf{r}, \mathbf{u})] \\ &= [\mathbf{G}f \cdot \mathbf{G}g]\,(\mathbf{r}, \mathbf{u}). \end{aligned}$$

When $\mathbf{p}(\mathbf{r}, \mathbf{u})$ is undefined, all of the above functions are zero, so equality holds trivially. Equation (6.24) follows by forming powers of f through repeated application of equation (6.23). Finally, equation (6.25) follows by observing that $\mathbf{G}f$ preserves every value assumed by the function f. Changing the sign of f at any point of its domain causes the corresponding sign change at a single point in the range of $\mathbf{G}f$. □□

The bound on $\mathbf{G}$ can now be obtained using the above lemma and the principle that radiance remains constant along straight lines in free space, which is applicable here since we are assuming that there is no participating medium present. The bound follows immediately from the following theorem.

Theorem 14 *If* $\mathbf{G}$ *is the field radiance operator associated with the manifold* $\mathcal{M}$*, then* $\|\, \mathbf{G}f \,\|_p \leq \|\, f \,\|_p$ *for all* $1 \leq p \leq \infty$ *and for all surface radiance functions* f*. Furthermore, when* $\mathcal{M}$ *forms an enclosure equality holds for all* f*, and when* $\mathcal{M}$ *does not form an enclosure, then either* $\|\, \mathbf{G}f \,\|_p = \|\, f \,\|_p$ *for some* f*, or* $\mathbf{G} = 0$*.*

<u>Proof:</u> Let us begin by assuming that $\mathcal{M}$ forms an enclosure, which is synonymous with $\mathbf{G}^2 = \mathbf{I}$. Then at any $\mathbf{r} \in \mathcal{M}$, the complete radiance distribution function over $\mathcal{S}^2$ is given by $f(\mathbf{r}, \mathbf{u}) + f(\mathbf{p}(\mathbf{r}, -\mathbf{u}), \mathbf{u})$. The net flux through the surface at $\mathbf{r}$

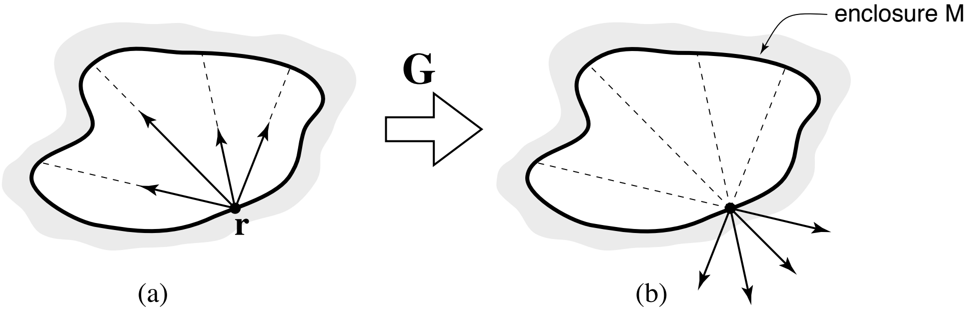

Figure 6.3: *The net flux at the point* $\mathbf{r}$ *is a function of (a) the surface radiance and (b) the field radiance constructed using* $\mathbf{G}$.

can then be obtained by integrating over $\mathcal{S}^2$, giving

$$\int_{\mathcal{S}^2} [\, f(\mathbf{r},\mathbf{u}) + f(\mathbf{p}(\mathbf{r},-\mathbf{u}),\mathbf{u})\,]\, d\mu(\mathbf{u}) \;=\; \Phi(\mathbf{r})\cdot\mathbf{n}(\mathbf{r}). \tag{6.26}$$

But $f(\mathbf{p}(\mathbf{r},\mathbf{u}),\mathbf{u})$ can be expressed in terms of the $\mathbf{G}$ operator, as shown in Figure 6.3. Integrating both sides over the entire surface $\mathcal{M}$, we have

$$\int_{\mathcal{M}}\int_{\mathcal{S}^2} [\, f(\mathbf{r},\mathbf{u}) - \mathbf{G}f(\mathbf{r},\mathbf{u})\,]\, d\mu(\mathbf{u})\; dm(\mathbf{r}) \;=\; \int_{\mathcal{M}} \Phi(\mathbf{r})\cdot\mathbf{n}(\mathbf{r})\; dm(\mathbf{r}). \tag{6.27}$$

But by Gauss's theorem it follows that

$$\int_{\mathcal{M}} \Phi(\mathbf{r})\cdot\mathbf{n}(\mathbf{r})\; dm(\mathbf{r}) \;=\; \int_V \nabla\cdot\Phi(\mathbf{r})\; dv(\mathbf{r}) \;=\; 0, \tag{6.28}$$

where the final equality is a result of equation (2.19), which states that the light field is solenoidal in empty space. This holds because we have excluded participating media from the volume V enclosed by $\mathcal{M}$. Thus, from equation (6.27) we have

$$\int f \;=\; \int \mathbf{G}\, f \tag{6.29}$$

for all surface radiance functions f, where the integration is taken to be over $\mathcal{M}\times\mathcal{S}^2$ and with respect to the measure $m \times \mu$. From equation (6.29) and lemma 2 we have the sequence of equalities

$$\int |f|^p \;=\; \int \mathbf{G}\,|f|^p \;=\; \int (\mathbf{G}\,|f|)^p \;=\; \int |\mathbf{G}f|^p, \tag{6.30}$$

which implies that $\| f \|_p = \| \mathbf{G} f \|_p$ for all f. Thus, $\mathbf{G}$ preserves all L_p-norms when $\mathcal{M}$ forms an enclosure. To show that the result also holds for non-enclosures we note that removing any part of $\mathcal{M}$ simply reduces the support of $\mathbf{G}f$, which can only decrease $\| \mathbf{G} f \|_p$. Consequently, $\| \mathbf{G} f \|_p$ remains bounded above by $\| f \|_p$. Furthermore, this bound can always be attained by a function f that defines a thin beam from one surface to another. When $\mathcal{M}$ is convex, no such beams exists and $\mathbf{G} = 0$, otherwise $\| \mathbf{G} f \|_p = \| f \|_p$. □□

Theorem 14 implies that $\| \mathbf{G} \|_p = 1$ for all $1 \leq p \leq \infty$, except when $\mathcal{M}$ is convex. This is a weaker statement, however, which does not subsume the theorem. Given the L_p bounds on $\mathbf{K}$ and $\mathbf{G}$, we may now bound the operator $\mathbf{M}^{-1}$, which maps surface emission functions to surface radiance functions at equilibrium. It follows from equation (6.22) and theorem 14 that $\mathbf{M}^{-1}$ exists and can be expressed as a Neumann series [77, p. 30]. Taking norms and summing the resulting geometric series, we have

$$\| \mathbf{M}^{-1} \|_p \leq 1 + \omega + \omega^2 + \ldots = \frac{1}{1-\omega} \tag{6.31}$$

for all $1 \leq p \leq \infty$, since $\| (\mathbf{KG})^n \| \leq \| \mathbf{KG} \|^n \leq \omega^n$. As a direct consequence of this bound, we have the following theorem.

Theorem 15 *The space $L_p(\mathcal{M} \times \mathcal{S}^2, m \times \mu)$ is closed under global illumination.*

Proof: Suppose that $f = \mathbf{M}^{-1} f_0$, where $f_0 \in L_p$. Because $\mathbf{M}^{-1}$ is bounded in all L_p-norms, it follows that $\| f \|_p \leq \| \mathbf{M}^{-1} \|_p \| f_0 \|_p < \infty$. Therefore the equilibrium solution f is also in L_p. □□

6.4 Related Operators

In this section we show that **K** and **G** can also be used to define and analyze other operators that arise in global illumination. In particular, we show that the adjoint of **M** with respect to the natural inner product follows directly from basic properties of **K** and **G**. We also establish the exact relationship between these operators and a similar operator decomposition proposed by Gershbein et al. [45].

6.4.1 Adjoint Operators

Adjoint operators have been studied in the context of global illumination for the purpose of computing solutions over portions of the environment, such as the surfaces that are visible from a given vantage point. This application of the adjoint transport operator has been used in simulating diffuse environments by Smits et al. [151] and in non-diffuse environments by Aupperle et al. [12]. The approach is known as *importance-driven* global illumination.

The problem of computing view-dependent solutions for global illumination is similar to the "flux at a point" problems that arise in reactor shielding calculations, where it is necessary to compute the flux of neutrons arriving at a small detector [29]. Such problems are solved much more efficiently using the adjoint transport equation [87,153]. For global illumination, the direct and adjoint solutions can be used in conjunction by computing them simultaneously. This approach allows view-dependent solutions to be computed far more efficiently than either component in isolation [151]. In this section we consider only the algebraic properties of adjoint operators and show how they relate to the operators **K** and **G**.

Because the adjoint of an operator is most naturally defined with respect to an inner product, we may proceed most easily in the setting of a Hilbert space. Consequently, we shall restrict our attention to the space L_2, since neither the L_1-norm nor the L_∞-norm can be defined in terms of an inner product. The latter fact can be shown very simply by observing that the *parallelogram law*

$$\| f - g \|^2 + \| f + g \|^2 = 2 \left(\| f \|^2 + \| g \|^2 \right) \tag{6.32}$$

holds for all f and g in a normed linear space if and only if the norm $\| \cdot \|$ is compatible with an inner product [126]. However, equation (6.32) is violated by both the L_1-norm and the L_∞-norm; for example, this is so when f and g are disjoint unit step-functions.

Both $\mathbf{K}$ and $\mathbf{G}$ have a natural symmetry, and they become *self-adjoint* in L_2 if we impose a simple identification. In particular, if $\mathbf{K}$ and $\mathbf{G}$ are to be identical with their Hilbert adjoints $\mathbf{K}^*$ and $\mathbf{G}^*$, it is necessary that they be mappings of a space into itself. Therefore, we must identify surface and field radiance functions so that they may be viewed as the same space.

To establish the required identification, we note that the space of field radiance functions is naturally isomorphic to the space of surface radiance functions. We denote the obvious isomorphism by $\mathbf{H}$, where

$$(\mathbf{H}f)(\mathbf{r}, \mathbf{u}) \equiv f(\mathbf{r}, -\mathbf{u}) \tag{6.33}$$

is the linear operator that simply reverses the direction associated with each radiance value. From this definition it is clear that $\mathbf{H}$ maps surface radiance functions to field radiance functions, and vise versa, and that $\mathbf{H} = \mathbf{H}^{-1}$. Using this isometry we may formally define new versions of both $\mathbf{K}$ and $\mathbf{G}$ that have additional symmetry. Specifically, we define

$$\begin{aligned} \bar{\mathbf{K}} &\equiv \mathbf{KH}, \\ \bar{\mathbf{G}} &\equiv \mathbf{HG}. \end{aligned}$$

Then $\mathbf{T} \equiv \mathbf{KG} = \bar{\mathbf{K}}\bar{\mathbf{G}}$. Note that $\mathbf{H}$ is an *isometric* isomorphism with respect to all L_p-norms; that is, $\| f \|_p = \| \mathbf{P}f \|_p$ for all p, which implies that $\| \bar{\mathbf{K}} \|_p = \| \mathbf{K} \|_p$ for all p. Using the new operators the space of surface radiance functions can be mapped onto itself without the intermediate space of field radiance functions. The actions of the various operators are depicted in the simple commutative diagram shown in Figure 6.4. By virtue of the isometry $\mathbf{H}$, it is clear that the distinction between the two types of radiance functions is only superficial, although it is an aid to visualizing the physical process. In the context of Hilbert adjoints, however, we must forgo this distinction so the operators may be defined on the same space.

$$\begin{tikzcd}
X_{\mathrm{s}} \arrow[r, "\mathbf{G}"] \arrow[d, "\bar{\mathbf{G}}"'] \arrow[dr, "\mathbf{T}"] & X_{\mathrm{f}} \arrow[d, "\mathbf{K}"] \\
X_{\mathrm{s}} \arrow[r, "\bar{\mathbf{K}}"'] & X_{\mathrm{s}}
\end{tikzcd}$$

Figure 6.4: *Operators connecting the spaces of surface and field radiance functions, which are denoted by X_{s} and X_{f} respectively.*

We shall now show that both $\bar{\mathbf{K}}$ and $\bar{\mathbf{G}}$ are "symmetric" with respect to the natural inner product on the space $L_2(\mathcal{M} \times \mathcal{S}^2, m \times \sigma)$, which is given by

$$\langle f \mid g \rangle \equiv \int_{\mathcal{M}} \int_{\mathcal{S}^2} f(\mathbf{r}, \mathbf{u})\, g(\mathbf{r}, \mathbf{u})\, d\mu(\mathbf{u})\, dm(\mathbf{r}). \tag{6.34}$$

More precisely, we shall show that $\left\langle \bar{\mathbf{G}} f \,\middle|\, g \right\rangle = \left\langle f \mid \bar{\mathbf{G}} g \right\rangle$ and $\left\langle \bar{\mathbf{K}} f \,\middle|\, g \right\rangle = \left\langle f \mid \bar{\mathbf{K}} g \right\rangle$, or equivalently, that $\bar{\mathbf{K}} = \bar{\mathbf{K}}^*$ and $\bar{\mathbf{G}} = \bar{\mathbf{G}}^*$. Operators with this property are said to be self-adjoint. We state and prove the result as a theorem.

Theorem 16 *The operators $\bar{\mathbf{K}}$ and $\bar{\mathbf{G}}$ are self-adjoint in the Hilbert space L_2; that is, they are self-adjoint with respect to the inner product $\langle \cdot \mid \cdot \rangle$ defined in equation (6.34).*

Proof: That $\bar{\mathbf{K}}$ is self-adjoint in L_2 follows from the symmetry of the kernel and

Fubini's theorem:

$$\begin{aligned}\left\langle f \mid \bar{\mathbf{K}} g \right\rangle &= \int_{\mathcal{M}} \int_{\Omega_o} f(\mathbf{r}, \mathbf{u}) \int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, g(\mathbf{r}, \mathbf{u}')\; d\mu(\mathbf{u}')\; d\mu(\mathbf{u})\; dm(\mathbf{r}) \\ &= \int_{\mathcal{M}} \int_{\Omega_o} g(\mathbf{r}, \mathbf{u}') \int_{\Omega_o} k(\mathbf{r}; \mathbf{u} \to \mathbf{u}')\, f(\mathbf{r}, \mathbf{u})\; d\mu(\mathbf{u})\; d\mu(\mathbf{u}')\; dm(\mathbf{r}) \\ &= \left\langle \bar{\mathbf{K}} f \,\middle|\, g \right\rangle.\end{aligned}$$

Therefore $\bar{\mathbf{K}}^* = \bar{\mathbf{K}}$ for any collection of surfaces $\mathcal{M}$ whose reflectance functions obey reciprocity. To obtain the corresponding result for $\bar{\mathbf{G}}$ it is convenient to begin by assuming that $\mathcal{M}$ forms an enclosure. Using the fact that $\bar{\mathbf{G}}^2 = \mathbf{I}$, followed by equations (6.23) and (6.29), we obtain the sequence of equalities

$$\begin{aligned}\left\langle \bar{\mathbf{G}} f \,\middle|\, g \right\rangle &= \int (\bar{\mathbf{G}} f) \cdot g = \int (\bar{\mathbf{G}} f) \cdot (\bar{\mathbf{G}}^2 g) \\ &= \int \bar{\mathbf{G}} \left(f \cdot \bar{\mathbf{G}} g \right) = \int f \cdot \bar{\mathbf{G}} g = \left\langle f \mid \bar{\mathbf{G}} g \right\rangle.\end{aligned}$$

Here the integration is once again taken to be over $\mathcal{M} \times \mathcal{S}^2$, as in equation (6.29). Hence, $\bar{\mathbf{G}}$ is self-adjoint when $\mathcal{M}$ forms an enclosure. To show that the result holds for arbitrary $\mathcal{M}$ we first express the space L_p as a direct sum of two orthogonal subspaces generated by $\bar{\mathbf{G}}$. In particular, we define the sets X_1 and X_2 by

$$\begin{aligned}X_1 &\equiv \left\{ f \in L_p : \bar{\mathbf{G}}^2 f = f \right\}, \\ X_2 &\equiv \left\{ f \in L_p : \bar{\mathbf{G}} f = 0 \right\}.\end{aligned}$$

It is easy to verify that X_1 and X_2 are subspaces of L_p such that $L_p = X_1 \oplus X_2$, $\bar{\mathbf{G}}^* = \bar{\mathbf{G}}$ on X_1, and $\bar{\mathbf{G}} X_1 \subset X_1$. Moreover, the functions of X_1 and X_2 have disjoint supports, so $X_1 \perp X_2$. Therefore, for all $f, g \in L_p$ we have

$$\left\langle \bar{\mathbf{G}} f \,\middle|\, g \right\rangle = \left\langle \bar{\mathbf{G}} f_1 \,\middle|\, g_1 \right\rangle = \left\langle f_1 \mid \bar{\mathbf{G}} g_1 \right\rangle = \left\langle f \mid \bar{\mathbf{G}} g \right\rangle, \tag{6.35}$$

where f_1 and g_1 denote the X_1-components of f and g respectively. From equation (6.35) we see that $\bar{\mathbf{G}}$ is self-adjoint on all of L_p. □□

It follows immediately from the previous theorem and basic properties of adjoint operators that the adjoint of $\mathbf{M}$ is

$$\mathbf{M}^* \;=\; \left(\mathbf{I} - \bar{\mathbf{K}}\bar{\mathbf{G}}\right)^* \;=\; \mathbf{I} - \bar{\mathbf{G}}\bar{\mathbf{K}}.$$

Therefore, we have shown that the operators $\bar{\mathbf{K}}$ and $\bar{\mathbf{G}}$ suffice to form both $\mathbf{M}$ and its adjoint. This is another indication that partitioning of the integral operator in equation (6.1) into the operators $\mathbf{K}$ and $\mathbf{G}$ is an extremely natural one.

6.4.2 Operators for Diffuse Environments

Gershbein et al. [45] proposed an operator decomposition similar to equation (6.8) but specifically tailored to diffuse environments. Their operators were chosen to exploit the very different behaviors exhibited by irradiance functions and surface reflectance functions; the former are nearly always smooth with respect to position, whereas the latter may vary rapidly due to high-frequency textures. We show that these operators are simply related to $\mathbf{K}$ and $\mathbf{G}$.

We begin by employing the notation of this chapter to rephrase the two essential operators used by Gershbein et al., which we shall denote here by $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{K}}$. The operator $\widehat{\mathbf{G}}$ is the *irradiance operator*, which is defined by

$$(\widehat{\mathbf{G}}b)(\mathbf{r}) \;\equiv\; \frac{1}{\pi}\int_{\Omega_o} b(\mathbf{p}(\mathbf{r},\mathbf{u}))\, d\mu(\mathbf{u}), \tag{6.36}$$

where $b : \mathcal{M} \to \mathrm{I\!R}$ is a surface radiosity function [watts/m^2], and $\mathbf{p}$ is the ray casting function defined on page 137. The factor of $1/\pi$ results from the conversion of radiosity to equivalent surface radiance on surrounding surfaces. The operator $\widehat{\mathbf{K}}$ is the *reflectance operator*, which is defined by

$$(\widehat{\mathbf{K}}b)(\mathbf{r}) \;\equiv\; \rho(\mathbf{r})\, b(\mathbf{r}), \tag{6.37}$$

where $\rho : \mathcal{M} \to \mathrm{I\!R}$ is the surface reflectivity function; that is, $\rho(\mathbf{r})$ is the constant of proportionality between irradiance and reflected radiosity at a point $\mathbf{r}$ on a diffuse surface. In analogy with equation (6.8), the operators $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{G}}$ lead to a governing

$$\begin{tikzcd}
\text{radiosity} \arrow[r, "\widehat{\mathbf{G}}"] \arrow[d, shift right, "\mathbf{U}"'] & \text{irradiance} \arrow[r, "\widehat{\mathbf{K}}"] & \text{radiosity} \arrow[d, shift right, "\mathbf{U}"'] \\
\text{surface radiance} \arrow[r, "\mathbf{G}"] \arrow[u, dashed, shift right, "\mathbf{U}^{-1}"'] & \text{field radiance} \arrow[r, "\mathbf{K}"] \arrow[u, dashed, "\mathbf{A}"] & \text{surface radiance} \arrow[u, dashed, shift right, "\mathbf{U}^{-1}"']
\end{tikzcd}$$

Figure 6.5: *The operators* $\widehat{\mathbf{K}}$ *and* $\widehat{\mathbf{G}}$ *play a role similar to* $\mathbf{K}$ *and* $\mathbf{G}$*.*

equation for global illumination in diffuse environments, where the unknown is now a radiosity function.

To illustrate the connections among $\mathbf{K}$, $\mathbf{G}$, $\widehat{\mathbf{K}}$, and $\widehat{\mathbf{G}}$, we define operators that convert between the corresponding radiometric quantities used in diffuse and non-diffuse settings. First, we define the *local irradiance operator* $\mathbf{A}$ by

$$(\mathbf{A}h)(\mathbf{r}) \equiv \int_{\Omega_i} h(\mathbf{r}, \mathbf{u})\; d\mu(\mathbf{u}), \tag{6.38}$$

where h is a field radiance function. The operator $\mathbf{A}$ converts field radiance to irradiance, which is the first moment about the surface normal at each point $\mathbf{r} \in \mathcal{M}$. Next, to convert between radiosity functions and surface radiance functions we define a prolongation operator $\mathbf{U}$ by

$$(\mathbf{U}b)(\mathbf{r}, \mathbf{u}) \;\equiv\; \frac{1}{\pi}\, b(\mathbf{r}). \tag{6.39}$$

The operator $\mathbf{U}$ elevates a radiosity function to the equivalent radiance function, which is independent of direction. Using the new operators we may express $\widehat{\mathbf{G}}$ as $\mathbf{AGU}$ in a diffuse environment. Similarly, at diffuse surfaces the kernel k satisfies $k(\mathbf{r}; \mathbf{u}' \to \mathbf{u}) \equiv \rho(\mathbf{r})/\pi$. In this case $\mathbf{K}$ simply averages over field radiance and produces a direction-independent surface radiance function. It follows that $\mathbf{U}^{-1}\mathbf{K} = \widehat{\mathbf{K}}\mathbf{A}$, where $\mathbf{U}^{-1}$ denotes the restriction operator that takes a direction-independent radiance function to the equivalent radiosity function. We are lead to the following simple relationship among the operators:

$$\begin{aligned} \widehat{\mathbf{K}}\widehat{\mathbf{G}} &= \widehat{\mathbf{K}}\,[\mathbf{AGU}] \\ &= \left[\widehat{\mathbf{K}}\mathbf{A}\right]\mathbf{GU} \\ &= \mathbf{U}^{-1}\mathbf{KGU}, \end{aligned}$$

which holds only for entirely diffuse environments. The actions of all the operators are shown in Figure 6.5, where the dashed arrows indicate transitions in which information is lost, except in the case of diffuse environments.

6.5 The Rendering Equation

The rendering equation, formulated by Kajiya [71], is the form in which global illumination problems are most frequently posed; we now show that the rendering equation is equivalent to equation (6.8). To conveniently express the new equation, Kajiya introduced a number of *multi-point transport* quantities that differ from standard radiometric quantities [72]. The rendering equation is

$$\widehat{f}(\mathbf{r},\mathbf{r}') = \widehat{g}(\mathbf{r},\mathbf{r}')\left[\,\widehat{f}_0(\mathbf{r},\mathbf{r}') + \int_{\mathcal{M}} \widehat{k}(\mathbf{r},\mathbf{r}',\mathbf{r}'')\,\widehat{f}(\mathbf{r}',\mathbf{r}'')\;dm(\mathbf{r}'')\right], \tag{6.40}$$

where $\mathbf{r},\mathbf{r}' \in \mathcal{M}$, and the non-standard transport quantities are denoted by hats; the corresponding terminology and physical units are summarized below.

Quantity	*Name*	*Physical Units*
$\widehat{f}(\mathbf{r},\mathbf{r}')$	transport intensity	watts m^{-4}
$\widehat{f}_0(\mathbf{r},\mathbf{r}')$	transport emittance	watts m^{-2}
$\widehat{k}(\mathbf{r},\mathbf{r}',\mathbf{r}'')$	scattering function	dimensionless
$\widehat{g}(\mathbf{r},\mathbf{r}')$	geometry term	m^{-2}

The *geometry term* $\widehat{g}$ in equation (6.40) is defined by

$$\widehat{g}(\mathbf{r}',\mathbf{r}) \equiv \begin{cases} \,||\,\mathbf{r}'-\mathbf{r}\,||^{-2} & \text{if } \mathbf{r}' \text{ and } \mathbf{r} \text{ are unoccluded} \\ \quad 0 & \text{otherwise.} \end{cases} \tag{6.41}$$

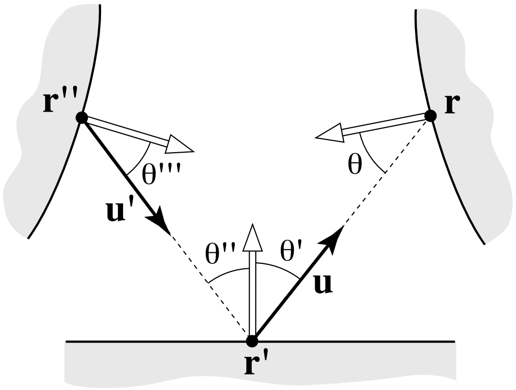

Figure 6.6: *The three points and four angles used in defining the multi-point transport quantities appearing in the rendering equation.*

This function encodes point-to-point visibility and is closely related to the ray casting function $\mathbf{p}$ defined earlier. The most significant difference between equation (6.40) and equation (6.1) is that the former is expressed exclusively in terms of surface points, whereas the latter uses points and solid angle. The physical interpretation of equation (6.40) is that radiant energy flows from the points $\mathbf{r}''$ on surrounding surfaces toward the point $\mathbf{r}'$, with some fraction reaching $\mathbf{r}$ after scattering. See Figure 6.6. The rendering equation can be derived from equation (6.1), or equivalently from equation (6.8), by a sequence of steps that include a change of variable. The derivation also leads naturally to the transport quantities.

6.5.1 Derivation from the Classical Formulation

Because the rendering equation is expressed in terms of surface points instead of directions, it is convenient to introduce the converse of the ray casting function. We define the *two-point direction function* by

$$\mathbf{q}(\mathbf{r}, \mathbf{r}') \equiv \frac{\mathbf{r}' - \mathbf{r}}{\|\, \mathbf{r}' - \mathbf{r} \,\|}. \tag{6.42}$$

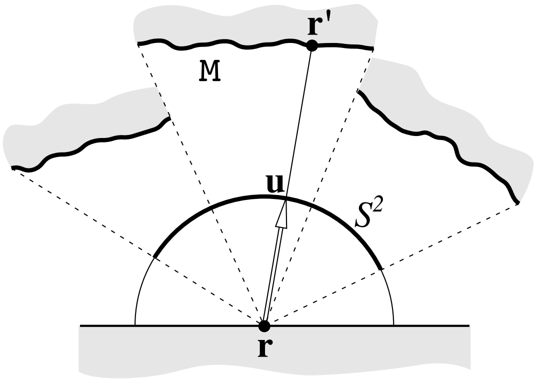

Figure 6.7: *The 2-form defined on the surface $\mathcal{M}$ is the pullback of the solid angle 2-form. The pullback is defined in terms of the two-point direction function, the inverse of the ray casting function.*

At a fixed point $\mathbf{r}$, the functions $\mathbf{q}$ and $\mathbf{p}$ are inverses of one another. Thus, if we define $\mathcal{M}' \equiv \mathbf{p}(\mathbf{r}, \Omega_o)$, which is the subset of $\mathcal{M}$ visible to the point $\mathbf{r}$, then

$$\mathbf{q}(\mathbf{r}, \mathcal{M}') \ \equiv \ \{\mathbf{q}(\mathbf{r}, \mathbf{r}') : \mathbf{r}' \in \mathcal{M}'\} \ = \ \Omega_o. \tag{6.43}$$

Note that $\mathcal{M}'$ is actually a function of $\mathbf{r}$. The two-point direction function is the key to deriving the rendering equation from the corresponding classical version

$$f(\mathbf{r}', \mathbf{u}) = f_0(\mathbf{r}', \mathbf{u}) + \int_{\Omega_i} k(\mathbf{r}'; \mathbf{u}' \to \mathbf{u})\, f(\mathbf{r}'', \mathbf{u}') \cos\theta''\ d\sigma(\mathbf{u}'), \tag{6.44}$$

where the variables appearing in this equation, and in the following discussion, are depicted in Figure 6.6. Note that only polar angles will enter into the equation, as these relate to surface geometry, and therefore to integration over surfaces. All dependence on the azimuth angles, on the other hand, is hidden within the kernel function k, which may correspond to an anisotropic BRDF.

Now, let $R(\mathbf{r}', \mathbf{u})$ denote the radiance at $\mathbf{r}'$ in the direction $\mathbf{u}$ due only to reflected light. That is,

$$R(\mathbf{r}', \mathbf{u}) = \int_{\mathbf{q}(\mathbf{r}, \mathcal{M}')} k(\mathbf{r}'; \mathbf{u}' \to \mathbf{u})\, f(\mathbf{r}'', \mathbf{u}') \cos\theta''\ d\sigma(\mathbf{u}'),$$

where we have used equation (6.43) to express the domain of integration in terms of $\mathbf{q}$; this was done in anticipation of a change of variables, which will express the integral in terms of the surface $\mathcal{M}'$ instead of solid angle. This transition is equivalent to changing a 2-form defined on the manifold $\mathcal{S}^2$ into a 2-form defined on the manifold $\mathcal{M}'$. The formal mechanism for accomplishing this is the *pullback* [57] defined by $\mathbf{q}$, which is denoted by $\mathbf{q}^*$. Since $\mathbf{q} : \mathcal{M}' \to \mathcal{S}^2$, the operation $\mathbf{q}^* d\omega$ pulls $d\omega$ back to $\mathcal{M}'$, where $d\omega$ is the solid angle 2-form. In particular, we have

$$\mathbf{q}^*\, d\omega = \frac{\cos\theta'''}{\|\,\mathbf{r}'' - \mathbf{r}'\,\|^2}\, d\mathbf{r}'', \tag{6.45}$$

where $d\mathbf{r}''$ denotes the volume element on $\mathcal{M}'$, which defines the measure m by

$$m(E) = \int_E d\mathbf{r}''$$

for all $E \in \mathcal{M}$. Equation (6.45) is a well-known formula for converting from solid angle to surface integration [117,32]. Performing the change of variables, we have

$$R(\mathbf{r}', \mathbf{u}) = \int_{\mathcal{M}'} k(\mathbf{r}'; \mathbf{q}(\mathbf{r}'', \mathbf{r}') \to \mathbf{u})\, f(\mathbf{r}'', \mathbf{q}(\mathbf{r}'', \mathbf{r}'))\frac{\cos\theta''\ \cos\theta'''}{\|\,\mathbf{r}'' - \mathbf{r}'\,\|^2}\, dm(\mathbf{r}'').$$

But the integral on the right may be expressed in terms of the complete manifold $\mathcal{M}$ by introducing the geometry term. This step can be viewed as constructing the characteristic function for the set $\mathcal{M}'$ in terms of the function $\widehat{g}(\cdot, \mathbf{r}')$, which has the added benefit of absorbing the factor of $\|\,\mathbf{r}'' - \mathbf{r}'\,\|^{-2}$. Thus, we may write

$$\begin{aligned} R(\mathbf{r}', \mathbf{u}) &= \int_{\mathcal{M}} k(\mathbf{r}'; \mathbf{q}(\mathbf{r}'', \mathbf{r}') \to \mathbf{u})\, f(\mathbf{r}'', \mathbf{q}(\mathbf{r}'', \mathbf{r}'))\, \widehat{g}(\mathbf{r}'', \mathbf{r}')\ \cos\theta'' \cos\theta'''\, dm(\mathbf{r}'') \\ &= \int_{\mathcal{M}} k(\mathbf{r}'; \mathbf{q}(\mathbf{r}'', \mathbf{r}') \to \mathbf{u})\, \widehat{f}(\mathbf{r}'', \mathbf{r}')\ dm(\mathbf{r}''), \end{aligned}$$

where the last equality follows from the definition of the two-point transport intensity function $\widehat{f}$, which is given below. Thus, equation (6.44) may be written

$$f(\mathbf{r}', \mathbf{u}) = f_0(\mathbf{r}', \mathbf{u}) + \int_{\mathcal{M}} k(\mathbf{r}'; \mathbf{q}(\mathbf{r}', \mathbf{r}'') \to \mathbf{u})\, \widehat{f}(\mathbf{r}'', \mathbf{r}')\ dm(\mathbf{r}''). \tag{6.46}$$

Since $\mathbf{u}$ is a free parameter, it may be expressed in different terms without altering the rest of the equation. In particular, we may obtain $\mathbf{u}$ from $\mathbf{r}'$ and a new parameter $\mathbf{r} \in \mathcal{M}$ by means of the two-point direction function. Thus, we have

$$f(\mathbf{r}', \mathbf{q}(\mathbf{r}', \mathbf{r})) = f_0(\mathbf{r}', \mathbf{q}(\mathbf{r}', \mathbf{r})) + \int_{\mathcal{M}} k(\mathbf{r}'; \mathbf{q}(\mathbf{r}'', \mathbf{r}') \to \mathbf{q}(\mathbf{r}', \mathbf{r}))\, \widehat{f}(\mathbf{r}'', \mathbf{r}')\; dm(\mathbf{r}''),$$

which now puts the equation in terms of $\mathbf{r}$ and $\mathbf{r}'$. Finally, to obtain the form of equation (6.40) we multiply both sides of the equation by $\widehat{g}(\mathbf{r}', \mathbf{r}) \cos\theta' \cos\theta$ and introduce the multi-point transport quantities, which are summarized below.

Quantity	*Defining expression*
$\widehat{f}(\mathbf{r}, \mathbf{r}')$	$\cos\theta' \cos\theta\; f(\mathbf{r}, \mathbf{q}(\mathbf{r}, \mathbf{r}'))\; \widehat{g}(\mathbf{r}, \mathbf{r}')$
$\widehat{f}_0(\mathbf{r}, \mathbf{r}')$	$\cos\theta' \cos\theta\; f_0(\mathbf{r}', \mathbf{q}(\mathbf{r}', \mathbf{r}))$
$\widehat{k}(\mathbf{r}, \mathbf{r}', \mathbf{r}'')$	$\cos\theta' \cos\theta\; k\,(\mathbf{r}'; \mathbf{q}(\mathbf{r}'', \mathbf{r}') \to \mathbf{q}(\mathbf{r}', \mathbf{r}))$

These quantities correspond to the definitions given by Kajiya [71], but are phrased here in terms of the two-point direction function $\mathbf{q}$. Note that the multi-point quantities could be defined more symmetrically if the emission term $\widehat{f}_0$ were to include a factor of $\widehat{g}(\mathbf{r}, \mathbf{r}')$; this would give both transport emission and transport intensity the same physical units as well as the same behavior with respect to occlusion; that is, both would be zero for occluded paths.

6.5.2 Measure-Theoretic Definition of Transport Intensity

The two-point transport intensity $\widehat{f}(\mathbf{r}, \mathbf{r}')$ is analogous to radiance $f(\mathbf{r}, \mathbf{u})$ and can be defined by means of a measure-theoretic argument similar to that given in chapter 2. We give an outline of the argument here.

Let A and B be subsets of $\mathcal{M}$, and let m denote Lebesgue measure over $\mathcal{M}$. We then define two natural measures over the product manifold $\mathcal{M} \times \mathcal{M}$: one that is purely geometrical, and one that characterizes the transfer of radiant energy from $\mathcal{M}$ to itself in steady state. To obtain two-point transport intensity, we define the geometrical measure to be simply the (completed) product measure $m \times m$. Let $\mathcal{F}$ denote the measure defined by the flow of radiant energy from $\mathcal{M}$ to itself; that is, $\mathcal{F}$ is the positive set function that assigns to each set $A \times B \subset \mathcal{M} \times \mathcal{M}$ the amount of power leaving A and reaching B directly. The positivity and additivity of the set function $\mathcal{F}$ satisfy the axioms of a measure on $\mathcal{M} \times \mathcal{M}$. The construction of $\mathcal{F}$ is similar to that used for the measure $\mathcal{E}$ defined in chapter 2, but here the emphasis is on surfaces, and the semantics of energy is present from the start.

Since either $m(A) = 0$ or $m(B) = 0$ implies that $\mathcal{F}(A \times B) = 0$, the measures are related by absolute continuity, denoted by $\mathcal{F} \ll m \times m$. This is the only connection we require between the measures to infer the existence of a density function. By the Radon-Nikodym theorem, there exists a function $\widehat{f} : \mathcal{M} \times \mathcal{M} \to \mathrm{I\!R}$ such that

$$\mathcal{F}(A \times B) = \int_A \int_B \widehat{f}(\mathbf{r}, \mathbf{r}')\; dm(\mathbf{r})\; dm(\mathbf{r}'). \tag{6.47}$$

The function $\widehat{f}$ is the two-point transport intensity. In view of equation (6.47), this quantity also corresponds to the Radon-Nikodym derivative of the measure $\mathcal{F}$ with respect to the product measure $m \times m$. This relationship is denoted by

$$\widehat{f} = \frac{d\mathcal{F}}{d(m \times m)}. \tag{6.48}$$

As in chapter 2, the dimensional relationship follows from this, yielding the dimensions of watts/m^4 for two-point transport intensity.

Equation (6.48) suggests that we may also define radiance directly from surface measures. To do so we proceed as above, but with a different geometric measure over $\mathcal{M} \times \mathcal{M}$. By replacing the product measure $m \times m$ with

$$F(A \times B) \equiv \int_A \int_B \frac{\cos\theta\, \cos\theta'}{||\, \mathbf{r} - \mathbf{r}'\, ||^2}\; dm(\mathbf{r})\, dm(\mathbf{r}'), \tag{6.49}$$

where the angles correspond to those shown in Figure 6.6, we obtain

$$f = \frac{d\mathcal{F}}{dF}, \tag{6.50}$$

where the density function f now corresponds to radiance. The quantity $F(A \times B)$ can be interpreted as the number of lines passing through both A and B, which is an important concept in the study of radiative transfer. In particular, the ratio

$$\frac{F(A \times B)}{\pi\, m(A)} \tag{6.51}$$

is the fraction of all lines passing through A that also meet B, which is precisely the *form factor* or *configuration factor* [69,148] from surface A to surface B. Thus, radiance is the Radon-Nikodym derivative of radiant power with respect line measure, which is closely related to the area-to-area form factors that appear in both radiative heat transfer and computer graphics.

Chapter 7

Error Analysis for Global Illumination

In this chapter we identify sources of error in global illumination algorithms and derive bounds for each distinct category. Errors arise from three sources: inaccuracies in the boundary data, discretization, and computation. Boundary data consist of surface geometry, reflectance functions, and emission functions, all of which may be perturbed by errors in measurement or simulation, or by simplifications made for computational efficiency. Discretization error is introduced by replacing the continuous radiative transfer equation with a finite-dimensional linear system, usually by means of boundary elements and a corresponding projection method. Finally, computational errors perturb the finite-dimensional linear system through roundoff error and imprecise form factors, inner products, visibility, etc., as well as by halting iterative solvers after a finite number of steps. Using the error taxonomy introduced in this chapter we examine existing global illumination algorithms.

There are many practical questions concerning error in the context of global illumination. For instance, in simulating a given physical environment, perhaps under varied lighting conditions, how accurately must the reflectance functions be measured? Or, when simulating radiant transfer among diffuse surfaces, how important is it to use higher-order elements? Can we expect higher accuracy by using analytic area-to-area instead of point-to-area form factors? Finally, how accurate must visibility computations be for global illumination? While the analysis presented in this chapter cannot provide definitive answers to these questions, it provides a formalism and a starting point for determining quantitative answers.

We shall only consider a well-defined class of global illumination problems. Specifically, we address the problem of approximating solutions to a form of the rendering equation [71] given imprecise data for geometry, reflectance functions, and emission functions. We further assume that the approximation is to be assessed quantitatively by its distance from the theoretical solution.

Given these restrictions, we derive error bounds in terms of potentially known quantities, such as bounds on emission and reflectivity, and bounds on measurement error. To bound the error of a numerical solution using this type of information we draw upon the general theory of integral equations [11] as well as the more abstract theory of operator equations [4,81].

7.1 Projection Methods

By far the most common methods for solving global illumination problems are those employing surface discretizations, which are essentially *boundary element* methods [95]. In more abstract terms, boundary element methods are themselves *projection methods* whose role is to recast infinite-dimensional problems in finite dimensions. In this section we pose the problem of numerical approximation for global illumination in terms of projections. This level of abstraction will allow us to clearly identify and categorize all sources of error while avoiding the details of specific implementations.

The idea behind boundary element methods is to construct an approximate solution from a known finite-dimensional subspace $X_n \subset X$, where the discretization parameter n typically denotes the dimension of the subspace. For global illumination the space X_n may consist of n boundary elements over which the radiance function is constant. Alternatively, it may consist of fewer boundary elements, but with internal degrees of freedom, such as tensor product polynomials [184], spherical harmonics [149], or wavelets [53]. In any case, each element of the function space X_n is a linear combination of a finite number of basis functions, $u_1, \ldots, u_n$. That is,

$$X_n = \operatorname{span}\left\{u_1, \ldots, u_n\right\}. \tag{7.1}$$

Given a set of basis functions, we seek an approximation f_n from X_n that is "close" to the exact solution f in some sense. By virtue of the finite-dimensional space, finding f_n is equivalent to determining n unknown coefficients $\alpha_1, \ldots, \alpha_n$ such that

$$f_n = \sum_{j=1}^{n} \alpha_j u_j. \tag{7.2}$$

There are many possible methods for selecting such an approximation from X_n, each motivated by a specific notion of closeness and the computational requirements of finding the approximation.

A universal feature of discrete boundary element approaches is that they operate using a finite amount of "information" gathered from the problem instance. For projection methods this is done in the following way. We select $f_n \in X_n$ by imposing a finite number of conditions on the *residual error*, which is defined by

$$r_n \equiv \mathbf{M} f_n - f_0. \tag{7.3}$$

Specifically, we attempt to find f_n such that r_n simultaneously satisfies n linear constraints. Since we wish to make the residual "small", we set

$$\phi_i(r_n) = 0, \tag{7.4}$$

for $i = 1, 2, \ldots n$, where the $\phi_i : X \to \mathbb{R}$ are linear functionals. The functionals and basis functions together define a projection operator, as we show in section 7.3.2.

Any collection of n linearly independent functionals defines an approximation f_n by "pinning down" the residual error with sufficiently many constraints to uniquely determine the coefficients. However, the choice of functionals has implications for the quality of the approximation as well as the computation required to obtain it. Combining equations (7.2), (7.3), and (7.4) we have

$$\phi_i \left(\mathbf{M} \sum_{j=1}^{n} \alpha_j u_j - f_0 \right) = 0, \tag{7.5}$$

for $i = 1, 2, \ldots, n$, which is a system of n equations for the unknown coefficients $\alpha_1, \ldots, \alpha_n$. By exploiting the linearity of ϕ_i and $\mathbf{M}$, we may express the above equations in matrix form:

$$\begin{bmatrix} \phi_1 \mathbf{M} u_1 & \cdots & \phi_1 \mathbf{M} u_n \\ \vdots & \ddots & \vdots \\ \phi_n \mathbf{M} u_1 & \cdots & \phi_n \mathbf{M} u_n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \phi_1 f_0 \\ \vdots \\ \phi_n f_0 \end{bmatrix}. \tag{7.6}$$

As we shall demonstrate below, most global illumination algorithms described in recent literature are special cases of the formulation in equation (7.6). Specific projection-based algorithms are characterized by the following components:

1. A finite set of basis functions $u_1, \ldots, u_n$

2. A finite set of linear functionals $\phi_1, \ldots, \phi_n$

3. Algorithms for evaluating $\phi_i \mathbf{M} u_j$ for $i, j = 1, 2, \ldots, n$

4. Algorithms for solving the discrete linear system

These choices do not necessarily coincide with the sequential steps of an algorithm. Frequently information obtained during evaluation of the linear functionals or during the solution of the discrete linear system is used to alter the choice of basis functions. The essence of adaptive meshing lies in this form of feedback. Regardless of the order in which the steps are carried out, the approximation generated ultimately rests upon specific choices in each of the above categories. Consequently, we can determine conservative bounds on the accuracy of the final solution by studying the impact of these choices independently.

7.2 A Taxonomy of Errors

In the previous section we characterized the fundamental features that distinguish projection-based global illumination algorithms. In this section we introduce a higher-level organization motivated by distinct categories of error; this subsumes the previous ideas and adds the notion of imprecise problem instances. Assuming that accuracy is measured by comparing with the exact solution to equation (6.1), all sources of error incurred by projection methods fall into one of three categories:

- **Perturbed Boundary Data:**

  Both the operator $\mathbf{M}$ *and the source term* f_0 *may be inexact due to measurement errors and/or simplifications made for efficiency.*

- **Discretization Error:**

  The finite-dimensional space X_n *may not include the exact solution. In addition, satisfying the constraints* $\phi_1, \ldots, \phi_n$ *may not select the best possible approximation from* X_n*.*

- **Computational Errors:**

  The matrix elements $\phi_i \mathbf{M} u_j$ *may not be computed exactly, thus perturbing the discrete linear system. Finally, the perturbed linear system may not be solved exactly.*

It is important to note that the above categories are mutually exclusive and account for all types of errors incurred in solving equation (6.1) with a projection method. The conceptual error taxonomy is shown schematically in Figure 7.1. In the remainder of this section we illustrate each of these categories of error with examples from existing algorithms.

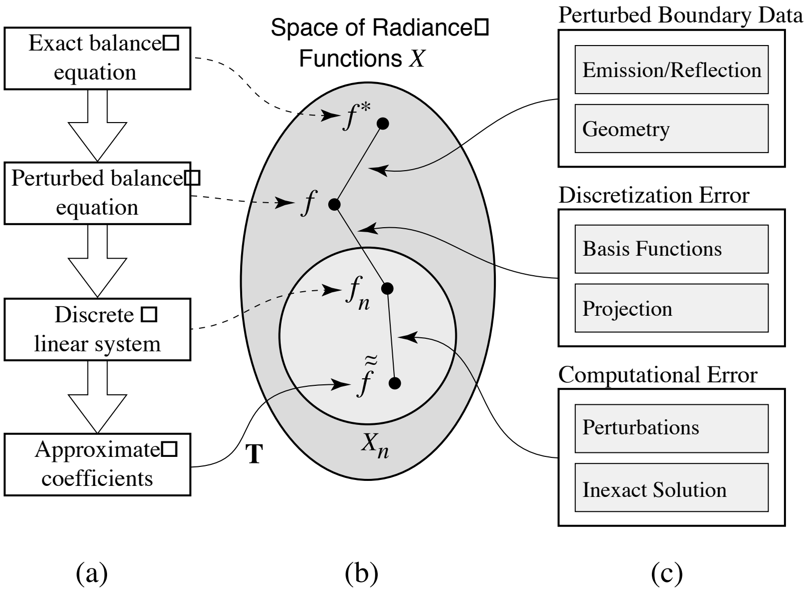

Figure 7.1: *(a) The conceptual stages for computing an approximate solution. (b) The exact solution at each stage is an approximation for the previous stage. (c) Approximations specific to each stage introduce new errors.*

7.2.1 Perturbed Boundary Data

The idealized problems that we solve in practice are rarely as realistic as we would like. As a rule, we settle for solving "near by" problems for several reasons. First, the data used as input may only be approximate. For instance, reflectance and emission functions obtained through simulation [177] or empirically through measurement of actual materials or light sources [162,173] are inherently contaminated by error.

A second reason that boundary data may be perturbed is that use of the exact data may be prohibitively expensive or even impossible. Thus, a surface may be treated as a smooth Lambertian reflector for the purpose of simulation, although the actual geometry and material exhibits directional scattering. Regardless of the source of the discrepancy, the near-by problem can be viewed as a *perturbation* of the original problem.

7.2.2 Discretization Error

To make the problem of global illumination amenable to solution by a digital computer, we must recast the problem in terms of finite-dimensional quantities and finite processes. This transition is referred to as *discretization.* In general, the discrete finite-dimensional problem cannot entirely capture the behavior of the infinite-dimensional problem, and the discrepancy is called *discretization error.* This type of error is particularly difficult to analyze as it inherently involves infinite-dimensional spaces and their relationship to finite-dimensional subspaces. Consequently, most global illumination algorithms rely upon heuristics rather than error bounds to perform discretization tasks such as adaptive meshing.

We now look at how two aspects of discretization have been treated in various algorithms: 1) the choice of basis functions, which determines the finite-dimensional space containing the approximation, and 2) projection, the finite process by which the approximation is selected from this space.

Basis Functions

For many global illumination algorithms, the basis functions are completely determined by the geometry of the boundary elements. This is true of any piecewise-constant approximation, such as those employed by Goral et al. [52], Cohen et al.[31], and Hanrahan et al. [62]. Often a smoothing step is applied to piecewise-constant approximations for display purposes, although this does not necessarily improve the accuracy of the approximation. When non-constant elements are employed, such as tensor product polynomials [184,164] or wavelets [53], the degrees of freedom of the elements add to the dimension of the approximating subspace.

For non-diffuse environments the finite-dimensional space must also account for the directional dependence of surface radiance functions; several avenues have been explored for doing this. Immel et al. [70] used subdivided cubes centered at a finite number of surface points to simultaneously discretize directions and positions. Sillion et al. [149] used a truncated series of spherical harmonics to capture directional dependence and a quadrilateral mesh of surface elements for the spatial dependence. As a third contrasting approach, Aupperle et al. [12] used piecewise-constant functions defined over pairs of patches to account for both directional and spatial variations.

The accuracy of the approximation is limited by the space of basis functions; in general, the exact solution cannot be formed by a finite linear combination of polynomials or other basis functions. The error can be reduced either by expanding the subspace X_n, or by selecting basis functions that fit the solution more closely. Discontinuity meshing [64,93] is an example of the latter strategy.

Projections

Given a set of basis functions, the next conceptual step is to construct an approximation of the exact solution from it. We now demonstrate how the major projection methods follow from specific choices of the linear functionals $\phi_1, \ldots, \phi_n$ described in section 7.1. The first technique is *collocation* [82], which follows by defining ϕ_i to be an *evaluation functional* at the ith collocation point; that is

$$\phi_i h \equiv h(x_i), \tag{7.7}$$

where $x_1, \ldots, x_n$ are distinct points in the domain of the radiance functions chosen so that $\det[u_i(x_j)] \neq 0$. Given these functionals, the ijth element of the matrix on the left of equation (7.6) has the form

$$u_j(x_i) - (\mathbf{KG}u_j)(x_i). \tag{7.8}$$

Collocation has been widely used in global illumination because of the relative simplicity of evaluating expressions of this form. In the case of constant basis functions over planar boundary elements, the resulting matrix has a unit diagonal with point-to-area form factors off the diagonal, which is a widely used radiosity formulation [31]. In general, methods based on a finite number of point-to-area interactions are collocation methods.

A second technique is the *Galerkin method*, which follows by defining ϕ_i to be an inner product functional with the ith basis function; that is

$$\phi_i h \equiv \langle u_i \mid h \rangle \,, \tag{7.9}$$

where the inner product $\langle \cdot \mid \cdot \rangle$ denotes the integral of the product of two functions. The ijth element of the resulting linear system has the form

$$\langle u_i \mid u_j \rangle - \langle u_i \mid \mathbf{KG} u_j \rangle \,. \tag{7.10}$$

The Galerkin method was first employed in global illumination by Goral et al. [52] using a uniform mesh and constant basis functions. The use of higher-order basis functions was investigated by Heckbert [64] and later by Zatz [184] and Troutman and Max [164]. For diffuse environments with constant elements, the second inner product in equation (7.10) reduces to an area-to-area form factor.

There are other possibilities for the linear functionals $\phi_1, \ldots, \phi_n$. For instance, the inner products in equation (7.9) may be taken with respect to a different set of basis functions. If we set

$$\phi_i h \equiv \langle \mathbf{M} u_i \mid h \rangle \,, \tag{7.11}$$

we obtain the *least squares* method. With the above functional, the solution to the linear system in equation (7.6) has a residual error that is orthogonal to the space X_n, which minimizes the residual error with respect to the L_2-norm. However, the matrix elements for the least squares method include terms of the form $\langle \mathbf{KG} u_i \mid \mathbf{KG} u_j \rangle$, which are formidable to evaluate even with trivial basis functions [66]. As a result, there are currently no practical global illumination algorithms based on the least squares method.

7.2.3 Computational Errors

Given a particular method of discretization, error may be incurred in constructing the discrete (finite-dimensional) linear system, and once formulated, we may fail to solve even the discrete problem exactly. These facts illustrate a third distinct class of errors. Because these errors arise from the practical limits of computational procedures, this class is called *computational errors* [91]. Computational errors perturb the discrete problem, and then preclude exact solution of the perturbed problem. We now look at examples of each of these.

Perturbation of the Linear System

The most computationally expensive operation of global illumination is the evaluation of the matrix elements in equation (7.6); this is true even of algorithms that do not store an explicit matrix [30]. Furthermore, only in very special cases can the matrix elements be formed exactly, as they entail visibility calculations coupled with multiple integration. Consequently, the computed matrix is nearly always perturbed by computational errors.

A common example of an error in this category is imprecise form factors. The algorithm introduced by Cohen and Greenberg [31] for diffuse environments computed form factors at discrete surface points by means of a hemicube, which introduced a number of errors specific to this approach [18]. Errors are also introduced when form factors are approximated through ray tracing [171] or by using simpler geometries [62]. These errors can be mitigated to some extent by the use of analytic form factors [18,140], yet there remain many cases for which no analytic expression is known, particularly in the presence of occluders.

For non-constant basis functions the form factor computations are replaced by more general inner products, which require different approximations. For instance, the matrix elements in the Galerkin approach of Zatz [184] required four-fold integrals, which were approximated using Gaussian quadrature. Non-diffuse environments pose a similar difficulty in that the matrix elements entail integration with reflectance functions [149].

Another form of matrix perturbation arises from simplifications made for the sake of efficiency. For example, small entries may be set to zero [94], or the entire matrix may be approximated by one with a more efficient structure, such as a block matrix [62] or a wavelet decomposition [53].

Inexact Solution of the Linear System

Once the discrete linear system is formed, we must solve for the coefficients $\alpha_1, \ldots, \alpha_n$. This has been done in a number of ways, including Gaussian elimination [52], Gauss-Seidel [31], Southwell relaxation [30,49], which is also known as *shooting*, and Jacobi iteration [62]. In any such method, there will be some error introduced in the solution process: direct solvers like Gaussian elimination are prone to roundoff error, whereas iterative solvers like Gauss-Seidel must halt after a finite number of iterations.

In approaches where the matrix is constructed in advance, such as "full matrix" radiosity [31] or hierarchical radiosity [62], the iterative solution can be carried out to essentially full convergence. Other approaches, such as progressive radiosity [30], construct matrix elements on the fly and then discard them. The cost of computing these elements generally precludes complete convergence, making this source of error significant.

7.3 Error Bounds

With the results of the previous section we can obtain bounds for each category of error. However, because the methods employed here take account of very little information about an environment, the bounds tend to be quite conservative. We now examine each source of error shown in Figure 7.1.

7.3.1 Bounding Errors Due to Perturbed Boundary Data

We solve global illumination problems using inexact or noisy data in hopes of obtaining a solution that is close to that of the original problem. But under what circumstances is this a reasonable expectation? To answer this question we analyze the mapping from problem instances $(\mathcal{M}, \mathbf{K}, f_0)$ to solutions f. We shall assume that any input to a global illumination problem may be contaminated by error, and determine the impact of these errors on the solution. We shall show that the problem of global illumination is well-posed in all physically realizable environments; that is, "small" perturbations of the input data produce "small" errors in the solution in physically meaningful problems.

To bound the effects of input data perturbations, we examine the quantity $||\, f^* - f \,||$, where f^* is the solution to the exact or unperturbed system, and f is the solution to the perturbed system

$$\widetilde{\mathbf{M}} f = \widetilde{f_0}, \tag{7.12}$$

where perturbed entities are denoted with tildes.

Perturbed Reflectance and Emission Functions

We first investigate the effect of perturbing the reflectance and emission functions. The former is equivalent to perturbing the local reflection operator $\mathbf{K}$. Consider a perturbed operator $\widetilde{\mathbf{K}}$ and a perturbed emission function $\widetilde{f_0}$ such that

$$\left\| \mathbf{K} - \widetilde{\mathbf{K}} \right\| \le \delta_k \tag{7.13}$$

and

$$\left|\left| f_0 - \widetilde{f_0} \right|\right| \le \delta_g. \tag{7.14}$$

Since $||\, \mathbf{G} \,|| = 1$, and $\mathbf{G}$ is assumed to be exact, it follows that inequality (7.13) also applies to the corresponding perturbed $\mathbf{M}$ operator:

$$\left\| \widetilde{\mathbf{M}} - \mathbf{M} \right\| = \left\| \mathbf{KG} - \widetilde{\mathbf{K}}\mathbf{G} \right\| \le \left\| \mathbf{K} - \widetilde{\mathbf{K}} \right\| \, ||\, \mathbf{G} \,|| \le \delta_k.$$

Intuitively, the above inequality holds because the worst-case behavior over all possible field radiance functions is unaffected by $\mathbf{G}$, which merely redistributes radiance.

To bound the error $||\, f^* - f \,||$ due to perturbations in the reflection and emission functions according to the inequalities (7.13) and (7.14) we write

$$\begin{aligned} ||\, f^* - f \,|| &= \left\| \, \mathbf{M}^{-1} f_0 - \widetilde{\mathbf{M}}^{-1} \widetilde{f_0} \, \right\| \\ &\leq \left\| \, \mathbf{M}^{-1} - \widetilde{\mathbf{M}}^{-1} \, \right\| \, \left|\left| \widetilde{f_0} \right|\right| + \left\| \, \mathbf{M}^{-1} \, \right\| \, \left|\left| f_0 - \widetilde{f_0} \right|\right| . \end{aligned}$$

From inequalities (6.18), (6.31) and (7.13), we have

$$\left\| \, \mathbf{M}^{-1} - \widetilde{\mathbf{M}}^{-1} \, \right\| \leq \left(\frac{\delta_k}{1 - \omega - \delta_k} \right) \left(\frac{1}{1 - \omega} \right) . \tag{7.15}$$

Combining the above and noting that $\left|\left| \widetilde{f_0} \right|\right| \leq ||\, f_0 \,|| + \delta_g$ by inequality (7.14), we arrive at the bound

$$||\, f^* - f \,|| \leq \left(\frac{\delta_k}{1 - \omega - \delta_k} \right) \left(\frac{||\, f_0 \,|| + \delta_g}{1 - \omega} \right) + \frac{\delta_g}{1 - \omega}, \tag{7.16}$$

which contains terms accounting for perturbations in the reflectance and emission functions individually as well as a second-order term involving $\delta_k \delta_g$. Note that the reflectance term requires $\delta_k < 1 - \omega$, indicating that the problem may become less stable as the maximum reflectivity approaches 1. For highly reflective environments, the results may be arbitrarily bad if the input data are not correspondingly accurate. In general, the worst-case absolute error in f depends upon the maximum reflectivity of the environment ω, the perturbation of the reflection functions δ_k, and the error in the emission function, δ_g.

Perturbed Surface Geometry

The effects of imprecise surface geometry are more difficult to analyze than those due to imprecise reflection or emission. While the space of radiance functions has a linear algebraic structure that underlies all of the analysis, no analogous structure exists on the set of possible surface geometries. The analysis must therefore proceed along different lines.

One possible alternative is to study imprecise surface geometry indirectly, by means of the $\mathbf{G}$ operator. That is, we can express the effect of surface perturbations on the field radiance at each point as a perturbation of $\mathbf{G}$. Given a perturbed operator $\widetilde{\mathbf{G}}$, and a bound on its distance from the exact $\mathbf{G}$, the same analysis used for perturbed reflectance functions can be applied. Such an approach may be useful for analyzing schemes in which geometry is simplified to improve the efficiency of global illumination, such as the algorithm proposed by Rushmeier et al. [133]. However, relating changes in geometry to bounds on the perturbation of $\mathbf{G}$ is an open problem, and solutions to this problem are likely to be norm-dependent.

7.3.2 Bounding Discretization Error

In this section we study discretization errors introduced by projection methods. Clearly, the discretization error $\| f - f_n \|$ is bounded from *below* by $\operatorname{dist}(f, X_n)$, the distance to the best approximation attainable within the space X_n. To obtain an upper bound, we express equation (7.4) using an explicit projection operator:

$$\mathbf{P}_n r_n = 0, \tag{7.17}$$

where 0 represents the zero function, and $\mathbf{P}_n$ is the projection operator corresponding to the subspace X_n. That is, $\mathbf{P}_n$ is a linear operator with $\mathbf{P}_n^2 = \mathbf{P}_n$ and $\mathbf{P}_n h = h$ for all $h \in X_n$. Such a projection can be defined in terms of the basis functions $u_1, \ldots, u_n$ and the linear functionals $\phi_1, \ldots, \phi_n$ described in section 7.1. The form of $\mathbf{P}_n$ is particularly simple when

$$\phi_i(u_j) = \delta_{ij}, \tag{7.18}$$

which is commonly the case. For instance, this condition is met whenever there is exactly one collocation point within the support of each basis function, or when orthogonal polynomials are used in a Galerkin-based approach. When equation (7.18) holds, the projection operator $\mathbf{P}_n$ is given by

$$\mathbf{P}_n h = \sum_{i=1}^{n} \phi_i(h) u_i \tag{7.19}$$

for any function $h \in X$ [50]. It is easy to see that equation (7.19) defines a projection onto X_n, and that $\mathbf{P}_n h = 0$ if and only if $\phi_i(h) = 0$ for $i = 1, 2, \ldots n$; that is, equation (7.17) is a valid replacement for equation (7.4). To produce the desired bound, we write equation (7.17) as

$$\mathbf{P}_n \mathbf{M}(f - f_n) = 0, \tag{7.20}$$

then isolate the quantity $\|\, f - f_n \,\|$. Adding $(\mathbf{I} - \mathbf{P}_n)(f - f_n)$ to both sides of equation (7.20) and simplifying, using the fact that $(\mathbf{I} - \mathbf{P}_n) f_n = 0$, we arrive at

$$\left[\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G}\right](f - f_n) = (\mathbf{I} - \mathbf{P}_n) f. \tag{7.21}$$

When the operator on the left of equation (7.21) is invertible, we obtain the bound

$$\|\, f - f_n \,\| \leq \left\| (\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G})^{-1} \right\| \; \|\, f - \mathbf{P}_n f \,\|. \tag{7.22}$$

A more meaningful bound can be obtained by simplifying both factors on the right hand side of the above equation [112]. Since $\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G}$ is an approximation of the operator $\mathbf{M}$, let δ_P be such that

$$\|\, \mathbf{M} - (\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G}) \,\| \leq \delta_P, \tag{7.23}$$

which simplifies to

$$\|\, \mathbf{K} - \mathbf{P}_n \mathbf{K} \,\| \leq \delta_P. \tag{7.24}$$

Because $\mathbf{M}$ is invertible, so is $\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G}$ when δ_P is sufficiently small. Banach's lemma then provides the bound

$$\left\| (\mathbf{I} - \mathbf{P}_n \mathbf{K} \mathbf{G})^{-1} \right\| \leq \frac{1}{1 - \omega - \delta_P}.$$

The second norm on the right of inequality (7.22) can be simplified as follows. Let $h \in X_n$. Noting that $\mathbf{P}_n h = h$, we have

$$\begin{aligned} \|\, f - \mathbf{P}_n f \,\| &= \|\, (f - h) + (h - \mathbf{P}_n f) \,\| \\ &\leq \|\, f - h \,\| + \|\, \mathbf{P}_n (h - f) \,\| \\ &\leq (1 + \|\, \mathbf{P}_n \,\|) \,\|\, f - h \,\| . \end{aligned}$$

Since $h \in X_n$ was chosen arbitrarily, the inequality holds for the greatest lower bound over X_n, which results in the bound

$$\|\, f - \mathbf{P}_n f \,\| \leq (1 + \|\, \mathbf{P}_n \,\|) \operatorname{dist}(f, X_n).$$

From the above inequalities we obtain the upper and lower bounds

$$\operatorname{dist}(\, f, X_n) \leq \|\, f - f_n \,\| \leq \left(\frac{\operatorname{dist}(\, f, X_n)}{1 - \omega - \delta_P} \right) (1 + \|\, \mathbf{P}_n \,\|) \,, \tag{7.25}$$

where the constant δ_P is such that

$$\|\, \mathbf{K} - \mathbf{P}_n \mathbf{K} \,\| \leq \delta_P. \tag{7.26}$$

The bounds in equation (7.25) depend on both the subspace X_n and the projection method. Note that if f is in the space X_n, then $\operatorname{dist}(\, f, X_n) = 0$, and the upper bound implies that $f_n = f$. Thus, all projection methods find the exact solution when it is achievable with a linear combination of the given basis functions. On the other hand, when $\operatorname{dist}(\, f, X_n)$ is large, then the lower bound implies that the approximation will be poor even when all other steps are exact. Unfortunately, this distance is difficult to estimate *a priori*, as it depends on the actual solution.

The dependence on the type of projection appears in the factor of $1 + \|\, \mathbf{P}_n \,\|$ and in the constant δ_P. For all projections based on inner products, $\|\, \mathbf{P}_n \,\| = 1$ when the basis functions are orthogonal [11, p. 64]. For other methods, such as collocation, the norm of the projection may be greater than one [11, p. 56]. The meaning of the constant δ_P is more subtle. The norm in equation (7.26) is a measure of how well the projection $\mathbf{P}_n$ captures features of the reflected radiance.

7.3.3 Bounding Computational Errors

The effects of computational errors can be estimated by treating them as perturbations of the *discrete* linear system. The analysis therefore parallels that of perturbed boundary data, although it is carried out in a finite dimensional space. We shall denote the linear system in equation (7.6) by $\mathbf{A}\alpha = \mathbf{b}$. In general, the matrix elements as well as the vector $\mathbf{b}$ will be inexact due to errors or simplifications. We denote the perturbed system and its solution by

$$\widetilde{\mathbf{A}}\,\tilde{\alpha} = \tilde{\mathbf{b}}. \tag{7.27}$$

Although the exact matrix is unknown, it is frequently possible to bound the error present in each element. For instance, this can be done for approximate form factors and block matrix approximations [62].

Given element-by-element error bounds, the impact on the final solution can be bounded. From the element perturbations we can find δ_A and δ_b such that

$$\left\| \,\mathbf{A} - \widetilde{\mathbf{A}}\, \right\| \leq \delta_A, \tag{7.28}$$

and

$$\left\| \,\mathbf{b} - \tilde{\mathbf{b}}\, \right\| \leq \delta_b, \tag{7.29}$$

for some vector norm and the matrix norm it induces. Also, since the perturbed matrix is known, the norm of its inverse can be estimated:

$$\left\| \,\widetilde{\mathbf{A}}^{-1}\, \right\| \leq \beta. \tag{7.30}$$

With the three bounds described above, essentially the same steps used in bounding the effects of perturbed boundary data can be applied here and yield the bound

$$||\,\alpha - \tilde{\alpha}\,|| \leq \left(\frac{\delta_A \beta^2}{1 - \delta_A \beta} \right) \left(\left\| \,\tilde{\mathbf{b}}\, \right\| + \delta_b \right) + \beta\delta_b. \tag{7.31}$$

The form of this bound is somewhat different for relative errors, which are more conveniently expressed in terms of condition numbers [109, p. 34].

Computing values for δ_A and δ_b that are reasonably tight is almost always difficult, requiring error bounds for each step in forming the matrix elements. There is no universal method by which this can be done; each approach to estimating visibility or computing inner products, for example, requires a specialized analysis. In contrast, the bound in equation (7.30) is more accessible, as it is purely a problem of linear algebra.

Given that equation (7.27) generally cannot be solved exactly, the solution process is yet another source of error. The result is an approximation of $\tilde{\alpha}$, which we denote by $\tilde{\tilde{\alpha}}$. This last source of error is bounded by

$$\left\| \, \tilde{\alpha} - \tilde{\tilde{\alpha}} \, \right\| \leq \beta \left\| \, \widetilde{\mathbf{A}} \, \tilde{\tilde{\alpha}} - \tilde{\mathbf{b}} \, \right\| . \tag{7.32}$$

When $\widetilde{\mathbf{A}}$ and $\tilde{\mathbf{b}}$ are stored explicitly, the above expression can be used as the stopping criterion for an iterative solver.

To relate errors in the coefficients $\alpha_1, \ldots, \alpha_n$ to errors in the resulting radiance function, consider the mapping $\mathbf{T} : \mathrm{I\!R}^n \rightarrow X_n$ where

$$\mathbf{T}x \equiv \sum_{j=1}^{n} x_j \, u_j . \tag{7.33}$$

As a finite-dimensional linear operator, $\mathbf{T}$ is necessarily bounded; its norm supplies the connection between coefficients in $\mathrm{I\!R}^n$ and functions in X_n. Observe that

$$\begin{aligned} \left\| \, f_n - \tilde{\tilde{f}} \, \right\| &= \left\| \, \mathbf{T}\alpha - \mathbf{T}\tilde{\tilde{\alpha}} \, \right\| \\ &\leq \, || \, \mathbf{T} \, || \left\| \, \alpha - \tilde{\tilde{\alpha}} \, \right\| \\ &\leq \, || \, \mathbf{T} \, || \left(\left\| \, \alpha - \tilde{\alpha} \, \right\| + \left\| \, \tilde{\alpha} - \tilde{\tilde{\alpha}} \, \right\| \right) . \end{aligned} \tag{7.34}$$

Equation (7.34) relates the computational error present in the final solution to inequalities (7.31) and (7.32). The value of $|| \, \mathbf{T} \, ||$ will depend on the basis functions $u_1, \ldots, u_n$ and the choice of norms for both $\mathrm{I\!R}^n$ and X_n, which need not be related.

7.4 The Combined Effect of Errors

In the previous sections we derived inequalities to bound the errors introduced into the solution of a global illumination problem. Using these inequalities we can now bound the distance between the exact solution and the computed solution. By the triangle inequality we have

$$\left\| \, f^* - \tilde{\tilde{f}} \, \right\| \leq \left\| \, f^* - f \, \right\| + \left\| \, f - f_n \, \right\| + \left\| \, f_n - \tilde{\tilde{f}} \, \right\| , \tag{7.35}$$

which is the numerical analogue of the chain of approximations shown in Figure 7.1. The terms on the right correspond to errors arising from perturbed boundary data, discretization, and computation; sections 7.3.1, 7.3.2, and 7.3.3 provide bounds for each of these errors. The first and third terms can be reduced in magnitude by decreasing errors in the emission and reflectance functions and in the computational methods for forming and solving the linear system.

Chapter 8

Conclusions

8.1 Summary

Chapter 2 presented a measure-theoretic development of phase space density, the abstract counterpart of radiance. The approach identified the physical and mathematical principles upon which radiance is based, and provided a simple proof of the constancy of radiance along rays in free space.

Chapter 3 presented a closed-form expression for the irradiance Jacobian due to polygonal sources of uniform brightness in the presence of arbitrary polygonal blockers. The expression can be evaluated in much the same way as Lambert's formula for irradiance. When blockers are present, a minor extension of standard polygon clipping is required. Several applications that make use of the irradiance Jacobian were demonstrated, including the direct computation of isolux contours and local irradiance extrema, both in the presence of polygonal occluders.

Chapter 4 presented a number of new closed-form expressions for computing illumination from luminaires (area light sources) with directional distributions as well as reflections from and transmissions through surfaces with a wide range of non-diffuse finishes. The expressions can be evaluated efficiently for arbitrary nonconvex polygons in O(nk) time, where n is related to the directionality of the luminaire or glossiness of the surface, and k is the number of edges in the polygon.

These are the first such expressions available.

The new expressions were derived using a proposed generalization of radiance called *irradiance tensors.* These tensors were shown to satisfy a simple recurrence relation that generalizes Lambert's well-known formula; expressions for *axial moments* and *double-axis moments* of polygonal luminaires were derived directly from this recurrence relation. The latter quantities have direct applications in simulating non-diffuse phenomena whose distributions are defined in terms of moments. The formulas give rise to efficient and easily implemented algorithms. The new algorithms were verified by means of Monte Carlo in chapter 5, where a new method was derived for generating stratified samples over spherical triangles.

Chapter 6 introduced a new operator equation describing the transfer of monochromatic radiant energy among opaque surfaces. The new formulation is appropriate for global illumination and has several theoretical advantages over previous formulations. First, it is based on standard radiometric concepts, which allows for the direct application of thermodynamic constraints. Secondly, it is well-suited to the *a priori* error analysis presented in this thesis.

Chapter 7 identified three sources of error in global illumination algorithms: inaccuracies in the input data, discretization, and computational errors. Input errors result from noise in measurement or simulation, or from simplifications. Discretization errors result from restricting the space of possible approximations and from the method of selecting the approximation. Computational errors form a large class that includes imprecise form factors and visibility, as well as errors introduced by block matrix approximations and iterative matrix methods. To produce a reliable solution, each of these sources of error must be accounted for. Using standard methods of analysis, we have derived worst-case bounds for each category based on properties of the new operators.

8.2 Future Work

The techniques of chapter 3 can be applied to other problems and extended. For instance, the expression for the irradiance Jacobian can be applied to geometric optimization problems involving illumination; that is, in selecting optimal shape parameters for luminaires and blockers with respect to some lighting objective function. The benefit of analytic Jacobians over finite difference approximations increases with the number of parameters. Also, the algorithms that incorporate irradiance gradients can be extended to apply to non-diffuse luminaires by means of the expression for double axis moments given in chapter 4.

There are a number of natural extensions to the work presented in chapter 4. Given axial moments with respect to two or more axes, all of arbitrary orders, it is possible to combine the effects demonstrated in the chapter. For example, it would be possible to compute glossy reflections of directional luminaires. Another application would be the simulation of non-diffuse surfaces illuminated by skylight using the polynomial approximation of a skylight distribution proposed by Nimroff et al. [106]. According to equation (4.59), all moments of this form admit closed-form expressions; however, practical algorithms for their evaluation do not yet exist. Another important extension is to accommodate more general polynomial functions over the sphere; in particular, polynomial approximations of realistic BRDFs obtained from theory [161] or measurement [173].

The analysis of chapter 7 would be more practical if it were based on more detailed information about environments; constants such as maximum reflectivity are much too coarse to obtain tight bounds. A deficiency of the L_p function norms employed in the chapter is that they do not adequately handle the wave optics effect of specular reflection at grazing angles. Other norms should be explored, including those that are in some sense perceptually-based. Finally, reliable bounds are needed for a wide assortment of standard computations, such as form factors between partially occluded surfaces and inner products involving higher-order elements.

Appendix A

Additional Proofs and Derivations

A.1 Proof of theorem 6

To prove theorem 6 on page 67, we first impose a parametrization on $\mathcal{S}^2$ by means of the standard coordinate charts

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

which map $[0, \pi] \times [0, 2\pi]$ onto $\mathcal{S}^2$. Then $d\omega = \sin\theta \, d\theta \, d\phi$, and

$$\eta(i, j, k) = \frac{1}{4\pi} \left[\int_0^{2\pi} \cos^i \phi \ \sin^j \phi \ d\phi \right] \left[\int_0^{\pi} \cos^k \theta \ \sin^{i+j+1} \theta \ d\theta \right]. \tag{A.1}$$

To simplify this expression, we introduce the function $p(i, j)$ defined by

$$p(i, j) \equiv \int_0^{\pi/2} \cos^i \theta \, \sin^j \theta \ d\theta, \tag{A.2}$$

which is positive for all i and j, and the function $\mathbf{e}(i)$, defined by

$$\mathbf{e}(i) \ \equiv \ \begin{cases} 1 & \text{if } i \text{ is even} \\ 0 & \text{otherwise.} \end{cases} \tag{A.3}$$

We now employ symmetries of sine and cosine to simplify the integrals on the right hand side of equation (A.1). First, because sine is an odd function, we have $\sin\theta = -\sin(\pi+\theta)$. It follows that

$$\int_0^{2\pi} \cos^i\phi\ \sin^j\phi\ d\phi \ = \ 2\,\mathbf{e}(j)\int_0^{\pi} \cos^i\phi\ \sin^j\phi\ d\phi.$$

Similarly, cosine is odd, so $\cos\theta = -\cos(\pi-\theta)$. It follows that

$$\int_0^{\pi} \cos^i\phi\ \sin^j\phi\ d\phi \ = \ 2\,\mathbf{e}(i)\int_0^{\pi/2} \cos^i\phi\ \sin^j\phi\ d\phi.$$

By means of these reductions, both factors on the right hand side of equation (A.1) may be expressed in terms of the function $p(i,j)$, yielding

$$\eta(i,j,k) = \frac{2}{\pi}\,\mathbf{e}(i)\,\mathbf{e}(j)\,\mathbf{e}(k)\,p(i,j)\,p(k,i+j+1). \tag{A.4}$$

The right hand side of equation (A.4) is clearly zero if and only if at least one of i, j, or k is odd. □□

A.2 Proof of theorem 7

In this appendix we supply the proof of theorem 7 on page 67, which gives a closed-form expression for $\eta(i,j,k)$. We first state and prove two useful lemmas.

Lemma 3 *If i and j are non-negative integers, then*

$$p(i+2,j) = \frac{i+1}{i+j+2}\,p(i,j) \tag{A.5}$$

$$p(i,j+2) = \frac{j+1}{i+j+2}\,p(i,j), \tag{A.6}$$

where the function $p(i,j)$ is defined by equation (A.2) on page 185.

Proof: Integrating $p(i+2,j)$ by parts, we have

$$\begin{aligned} p(i+2,j) &= -\int_0^{\pi/2} \left[j\cos^2\theta \;-\; (i+1)\,\sin^2\theta\right] \cos^i\theta\,\sin^j\theta\,d\theta \\ &= \int_0^{\pi/2} \left[i+1-(i+j+1)\,\cos^2\theta\right] \cos^i\theta\,\sin^j\theta\,d\theta \\ &= (i+1)\,p(i,j) \;-\; (i+j+1)\,p(i+2,j). \end{aligned}$$

Thus, $(i+j+2)\,p(i+2,j) = (i+1)\,p(i,j)$, which verifies equation (A.5). Equation (A.6) follows in precisely the same manner. □□

Lemma 4 *If i, j, and k are non-negative integers, and $n = i+j+k$, then*

$$\eta(i+2,j,k) = \frac{i+1}{n+3}\,\eta(i,j,k), \tag{A.7}$$

and the analogous result also holds for the second and third arguments.

Proof: Equation (A.7) clearly holds when any of i, j, or k are odd, as both sides are then zero. When i, j, and k are all even, it follows from equation (A.4) that

$$\eta(i,j,k) = \frac{2}{\pi}\, p(i,j)\, p(k, i+j+1). \tag{A.8}$$

From equation (A.8) and recurrence relations (A.5) and (A.6), we have

$$\begin{aligned} \eta(i+2,j,k) &= \frac{2}{\pi}\, p(i+2,j)\, p(k,i+j+3) \\ &= \frac{2}{\pi}\left[\frac{i+1}{i+j+2}\, p(i,j)\right]\left[\frac{i+j+2}{i+j+k+3}\, p(k,i+j+1)\right] \\ &= \frac{i+1}{n+3}\, \eta(i,j,k), \end{aligned} \tag{A.9}$$

which proves the lemma. □□

The proof of theorem 7 follows easily from lemma 4. We start by deriving an expression for $\eta(i,j,k)$ in terms of double factorials. Assuming that i, j, and k are all even integers, repeated application of recurrence relation (A.7) gives

$$\begin{aligned} \eta(i,j,k) &= \frac{i-1}{n+1}\, \eta(i-2,j,k) \\ &= \frac{(i-1)(i-3)\cdots 1}{(n+1)(n-1)\cdots(j+k+1)}\, \eta(0,j,k) \\ &= \frac{[\,(i-1)(i-3)\cdots 1\,]\ [\,(j-1)(j-3)\cdots 1\,]\ [\,(k-1)(k-3)\cdots 1\,]}{(n+1)(n-1)(n-3)\cdots 1} \\ &= \frac{(i-1)!!\,(j-1)!!\,(k-1)!!}{(n+1)!!}, \end{aligned}$$

which proves the first part of the theorem. The alternate expression for $\eta(i,j,k)$, based on multinomial coefficients, can now be derived using the two identities

$$\begin{aligned} n! &= n!!\,(n-1)!! \\ (2n)!! &= n!\,2^n, \end{aligned}$$

which follow immediately from the definition of the double factorial. Thus,

$$\begin{aligned}\frac{(i-1)!!\,(j-1)!!\,(k-1)!!}{(n+1)!!} &= \frac{1}{n+1}\left(\frac{n!!}{n!}\right)\left(\frac{i!}{i!!}\right)\left(\frac{j!}{j!!}\right)\left(\frac{k!}{k!!}\right)\\ &= \frac{1}{n+1}\left[\frac{(n/2)!}{(i/2)!\,(j/2)!\,(k/2)!}\right]\left[\frac{i!\,j!\,k!}{n!}\right]\\ &= \frac{1}{n+1}\binom{n'}{i'\quad j'\quad k'}\Big/\binom{n}{i\quad j\quad k},\end{aligned}$$

where the primes indicate division by two. This proves the second part of the theorem. □□

A.3 Proof of theorem 8

In this appendix we provide the proof of theorem 8 on page 77.

Consider a single element of the n-tensor Ξ^n corresponding to a given multi-index $\mathrm{I} = (i_1, ..., i_n)$, and let $(i, j, k) = (\kappa^1_{\mathrm{I}}, \kappa^2_{\mathrm{I}}, \kappa^3_{\mathrm{I}})$. When n is odd, it follows from theorem 6 that $\Xi^n_{\mathrm{I}} = 0$, since at least one of i, j, or k is odd; this verifies the first part of the theorem. Now, suppose that n is even and consider the summation on the right of equation (4.36). If any of i, j, or k are odd, then at least one δ-function in each term will have non-matching indices, making $\Xi^n_{\mathrm{I}} = 0$ as required. When i, j, and k are all even, the only terms in the sum that are non-zero are those in which the permutation $(j_1, \ldots, j_n)$ forms a sequence of matching pairs; that is, $j_{2m-1} = j_{2m}$ for $m = 1, 2, \ldots, n/2$. To count the number of times this occurs for the given index I, observe that there are

$$\begin{pmatrix} & n' & \\ i' & j' & k' \end{pmatrix} \tag{A.10}$$

distinct arrangements of the $n' \equiv n/2$ matching pairs, and each of these arrangements occurs with a multiplicity of $i!\, j!\, k!$ within the set of all $n!$ permutations of $(i_1, \ldots, i_n)$. Therefore

$$\begin{aligned} \Xi^n_{\mathrm{I}} &= \frac{i!\, j!\, k!}{(n+1)!} \begin{pmatrix} & n' & \\ i' & j' & k' \end{pmatrix} \\ &= \left(\frac{1}{n+1}\right) \begin{pmatrix} & n' & \\ i' & j' & k' \end{pmatrix} \Big/ \begin{pmatrix} & n & \\ i & j & k \end{pmatrix} \\ &= \eta(i, j, k), \end{aligned} \tag{A.11}$$

which proves the theorem. □□

A.4 Proof that $x^i y^j z^k \, d\omega$ is closed

We show that the 2-form $x^i y^j z^k \, d\omega$ is closed, where x, y, and z are the direction cosines, and $d\omega$ is the solid angle 2-form defined on page 72. We shall show that

$$d \left[\frac{x^i y^j z^k}{r^n} \, d\omega \right] = 0, \tag{A.12}$$

where $r \equiv \sqrt{x^2 + y^2 + z^2}$ and $n \equiv i + j + k$. The computation is purely mechanical. By the definition of $d\omega$ we have

$$\frac{x^i y^j z^k}{r^n} \, d\omega \;=\; \frac{x^{i+1} y^j z^k \, dy \wedge dz \;+\; x^i y^{j+1} z^k \, dz \wedge dx \;+\; x^i y^j z^{k+1} \, dx \wedge dy}{r^{n+3}}.$$

In computing the differential of the expression on the right, two thirds of the terms vanish since the wedge product of a 1-form with itself is zero. For example, the expression $d(x^{i+1} \, y^j \, z^k \, dy \wedge dz)$ results in only one non-zero term; the partial with respect to x. Performing all such simplifications, we have

$$\begin{aligned} d \left[\frac{x^i y^j z^k}{r^n} \, d\omega \right] \;&=\; \frac{(i+1) + (j+1) + (k+1)}{r^{n+3}} \, x^i \, y^j \, z^k \; dx \wedge dy \wedge dz \\ &\quad - (n+3) \frac{x^2 + y^2 + z^2}{r^{n+5}} \, x^i \, y^j \, z^k \; dx \wedge dy \wedge dz \\ &=\; 0. \end{aligned}$$

Thus, the initial 2-form is closed for all exponents i, j, and k. □□

A.5 Proof of lemma 1

In this appendix we derive the recurrence relation presented as lemma 1 on page 108. To begin, we write

$$F_n \;\equiv\; \int (a + b\,\cos\theta)^n\, d\theta \;=\; a\,F_{n-1} \;+\; b\int (a + b\,\cos\theta)^{n-1}\,\cos\theta\, d\theta.$$

Integrating the final term by parts, we have

$$F_n = a\,F_{n-1} \;+\; b\,(a + b\,\cos\theta)^{n-1}\sin\theta \;+\; (n-1)\,b^2\int (a + b\,\cos\theta)^{n-2}\,\sin^2\!\theta\, d\theta.$$

Next, we express $b^2\,\sin^2\!\theta$ in terms of $(a + b\,\cos\theta)$, resulting in

$$\begin{aligned} b^2\,\sin^2\!\theta \;&=\; (b \;+\; b\,\cos\theta)(b \;-\; b\,\cos\theta) \\ &=\; [\;(b-a) + (a + b\cos\theta)\;] \times [\;(b+a) - (a + b\cos\theta)\;]\,. \end{aligned}$$

Substituting the above into the expression for F_n, we have

$$\begin{aligned} F_n \;&=\; a\,F_{n-1} + b\,(a + b\,\cos\theta)^{n-1}\sin\theta \\ &\quad + (n-1)\int \Big[-(a + b\cos\theta)^n + 2a(a + b\cos\theta)^{n-1} + (b^2 - a^2)(a + b\cos\theta)^{n-2}\Big]\, d\theta \\ &=\; a\,F_{n-1} + b\,(a + b\,\cos\theta)^{n-1}\sin\theta \\ &\quad + (n-1)\Big[\,-F_n \;+\; 2a\,F_{n-1} \;+\; (b^2 - a^2)\,F_{n-2}\,\Big]\,. \end{aligned}$$

Simplifying, we arrive at

$$n\,F_n = a(2n-1)\,F_{n-1} \;+\; (n-1)(b^2 - a^2)\,F_{n-2} \;+\; (a + \cos\theta)^{n-1}\sin\theta,$$

with the base cases $n = 0$ and $n = 1$ following immediately from the definition of F_n, which proves the lemma. □□

A.6 Proof of theorem 11

In this appendix we provide the proof of theorem 11 on page 112. The steps of the proof are identical to those used in the proof of theorem 10, with a crucial difference appearing in the formulation of equation (4.60). To account for the introduction of the $\mathbf{w} \cdot \mathbf{u}$ expression in the denominator, we define

$$\mathbf{A}^{n,q} \equiv \left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \mathbf{A}^n, \tag{A.13}$$

where $\mathbf{A}^n$ is the nth-order tensor defined in equation (4.61). We then proceed in the same fashion as before, but with the new n-tensor $\mathbf{A}^{n,q}$. First, we compute the partials of $\mathbf{A}^{n,q}$. Differentiating the product, we have

$$\mathbf{A}^{n,q}_{\mathrm{J},m} = \frac{\partial}{\partial\, \mathbf{r}_m}\left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \mathbf{A}^n_{\mathrm{J}} \;+\; \left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \mathbf{A}^n_{\mathrm{J},m}, \tag{A.14}$$

where J is an n-index. As in equation (4.65), we compute the product of ε_{kml} with $\mathbf{A}^{n,q}_{\mathrm{J},m}$; for clarity, we perform this step for each of the two terms on the right of equation (A.14) individually. Let I denote an $(n-2)$-index. Using the expression for $\mathbf{A}^n$ given in equation (4.61), we have

$$\begin{aligned}
\varepsilon_{kml}\, \frac{\partial}{\partial\, \mathbf{r}_m}\left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \mathbf{A}^n_{\mathrm{I}jl} &= \left[\frac{q\, r^{q-2}}{(\mathbf{r} \cdot \mathbf{w})^{q+1}}\right] \left[r^2 \mathbf{w}_m - \mathbf{r}_m(\mathbf{r} \cdot \mathbf{w})\right] \varepsilon_{kml}\, \mathbf{A}^n_{\mathrm{I}jl} \\
&= \left[\frac{q\, \mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{r}_p}{(\mathbf{r} \cdot \mathbf{w})^{q+1}}\right] \left[\frac{\mathbf{r}_m(\mathbf{r} \cdot \mathbf{w}) - r^2 \mathbf{w}_m}{(n+1) r^{n+3-q}}\right] \left(\delta_{pk}\delta_{jm} - \delta_{pm}\delta_{jk}\right) \\
&= \left[\frac{q\, \mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{r}_k}{(\mathbf{r} \cdot \mathbf{w})^{q+1}}\right] \left[\frac{\mathbf{r}_j(\mathbf{r} \cdot \mathbf{w}) - r^2 \mathbf{w}_j}{(n+1) r^{n+3-q}}\right] \\
&= \frac{q}{n+1} \left[\frac{\mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{r}_j}{(\mathbf{r} \cdot \mathbf{w})^q\, r^{n-q}} - \frac{\mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{w}_j}{(\mathbf{r} \cdot \mathbf{w})^{q+1}\, r^{n-2-q}}\right] \left(\frac{\mathbf{r}_k}{r^3}\right).
\end{aligned}$$

Next, we consider the second term on the right of equation (A.14). Multiplying by ε_{kml} and simplifying, we have

$$\begin{aligned}
\varepsilon_{kml} \left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \mathbf{A}^n_{\mathrm{I}jl,m} &= \left(\frac{r}{\mathbf{r} \cdot \mathbf{w}}\right)^q \left[\frac{\delta_{i_2 j}\, \mathbf{r}_{\mathrm{I}/2}^{n-3} + \cdots \delta_{i_1 j}\, \mathbf{r}_{\mathrm{I}/1}^{n-3} +}{r^{n+1}} \;-\; \frac{\mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{r}_j}{r^{n+3}}\right] \mathbf{r}_k \\
&= \left[\frac{\left(\delta_{i_1 j}\, \mathbf{r}_{\mathrm{I}/1}^{n-3} + \delta_{i_2 j}\, \mathbf{r}_{\mathrm{I}/2}^{n-3} + \cdots\right) r^2 \;-\; (n+1)\, \mathbf{r}_{\mathrm{I}}^{n-2}\, \mathbf{r}_j}{(\mathbf{r} \cdot \mathbf{w})^q\, (n+1)\, r^{n-q}}\right] \left(\frac{\mathbf{r}_k}{r^3}\right).
\end{aligned}$$

Note that both of the final expressions above include a factor of $\mathbf{r}_k/r^3$; this factor will be used to complete the 2-form corresponding to $d\omega$, thereby converting the surface integral to an integral over solid angle. Thus, we have

$$\begin{aligned}\int_{\partial A} \mathbf{A}^{n,q}_{\mathrm{I}jl}\, d\mathbf{r}_l &= \int_{\partial A} \mathbf{A}^{n,q}_{\mathrm{I}jl,m}\, d\mathbf{r}_m \wedge d\mathbf{r}_l \\ &= \frac{q}{n+1}\left[\mathbf{T}^{n,q}_{\mathrm{I}j} \;-\; \mathbf{w}_j\, \mathbf{T}^{n-1,q+1}_{\mathrm{I}}\right] - \mathbf{T}^{n,q}_{\mathrm{I}j} + \frac{1}{n+1}\sum_{k=1}^{n-1} \delta_{\mathrm{I}_k j}\, \mathbf{T}^{n-2,q}_{\mathrm{I}/k}.\end{aligned}$$

Simplifying and rearranging terms, we obtain

$$\mathbf{T}^{n,q}_{\mathrm{I}j} = \frac{1}{n+1-q}\left[\sum_{k=1}^{n-1} \delta_{\mathrm{I}_k j}\, \mathbf{T}^{n-2,q}_{\mathrm{I}/k} \;-\; q\, \mathbf{w}_j\, \mathbf{T}^{n-1,q+1}_{\mathrm{I}} \;-\; \int_{\partial A} \frac{\mathbf{u}^{n-1}_{\mathrm{I}}\, \mathbf{n}_j}{(\mathbf{u}\cdot\mathbf{w})^q}\, ds\right],$$

which proves the theorem. □□

A.7 Reducing Υ to Known Special Functions

We show that the special function $\Upsilon(a, x)$ defined on page 112 can be expressed in terms of the well-known *dilogarithm* function, denoted by $\mathrm{Li}_2(z)$ and defined over the complex plane. We shall make use of two intermediate functions: Lobachevsky's function [55, p. 941], defined by

$$L(x) \equiv \int_0^x \log(\cos\theta)\, d\theta, \tag{A.15}$$

and a two-parameter generalization of this function, denoted by M, where

$$M(\alpha, x) \equiv \int_0^x \log(\cos\theta + \cos\alpha)\, d\theta. \tag{A.16}$$

Both of these integrals can be expressed in terms of the dilogarithm function. Gröbner and Hofreiter [56, vol. II, pp. 69–70] provide the following essential formulas:

$$\begin{aligned} L(x) &= \frac{i}{2}\,\mathrm{Li}_2\left(e^{i(2x-\pi)}\right) - x\log 2 - \frac{i}{2}\left(x^2 - \frac{\pi^2}{12}\right), & \text{(A.17)} \\ M(\alpha, x) &= i\,\mathrm{Li}_2\left(e^{i(\alpha+x-\pi)}\right) - i\,\mathrm{Li}_2\left(e^{i(\alpha-x-\pi)}\right) - x(\log 2 + i\alpha), & \text{(A.18)} \end{aligned}$$

where the parameter α may be an arbitrary complex number. Since a variety of distinct definitions of the dilogarithm have appeared in the literature [1,56,86], we note that here $\mathrm{Li}_2(z)$ is given by the series

$$\mathrm{Li}_2(z) \equiv \sum_{n=1}^{\infty} \frac{z^n}{n^2}, \tag{A.19}$$

for all complex z such that $|z| < 1$. An alternate definition of the dilogarithm is given by the integral

$$\mathrm{Li}_2(z) \equiv -\int_0^z \frac{\log(1-t)}{t}\, dt, \tag{A.20}$$

which is equivalent to the series on the unit disk, but extends the definition to the entire complex plane [86]. Only elementary identities are needed to express $\Upsilon(a, x)$

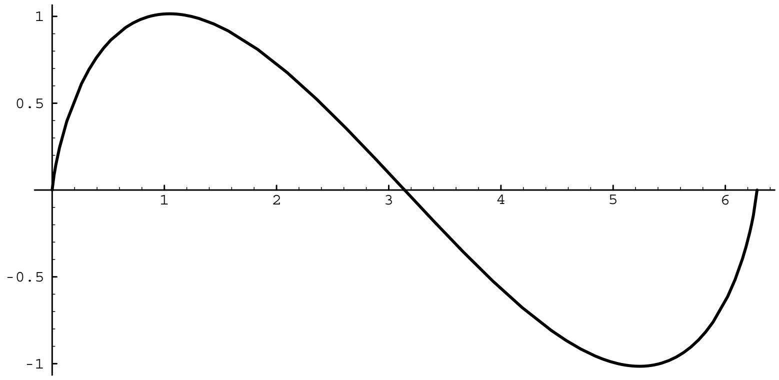

Figure A.1: *Clausen's integral over the domain* $[0, 2\pi]$.

in terms of the functions L and M above. The reduction proceeds as follows:

$$\begin{aligned}
\Upsilon(a, x) &\equiv \int_0^x \log\left(1 + \frac{a^2}{\cos^2\theta}\right) d\theta \\
&= \int_0^x \log\left(\cos^2\theta + a^2\right) d\theta \;-\; \int_0^x \log\left(\cos^2\theta\right) d\theta \\
&= \frac{1}{2}\int_0^{2x} \log\left[\frac{\cos\theta + 2a^2 + 1}{2}\right] d\theta \;-\; 2\int_0^x \log(\cos\theta)\, d\theta \\
&= \frac{1}{2} M(\, i\, \cosh^{-1}(2a^2+1),\, 2x) \;-\; x\,\log 2 \;-\; 2\, L(x), \qquad \text{(A.21)}
\end{aligned}$$

where the last step follows from the fact that $\cos(ix) = \cosh x$ for any real number x. A disadvantage of this formulation is that it involves dilogarithms with complex arguments, which requires a two-parameter special function [86, p. 121]. Furthermore, the arguments have modulus greater than one, which precludes the use of the power series in equation (A.19). An alternative is to express $\Upsilon(a, x)$ in terms of the *Clausen integral*, denoted by $\mathrm{Cl}_2(\theta)$ and defined by

$$\mathrm{Cl}_2(x) \;\equiv\; -\int_0^x \log\left(2\sin\frac{\theta}{2}\right) d\theta \;=\; \sum_{n=1}^{\infty} \frac{\sin nx}{x^2}. \qquad \text{(A.22)}$$

Here it is the series representation that is defined over the entire complex plane. From the series it is clear that the function is a cyclic, with $\mathrm{Cl}_2(x + 2n\pi) = \mathrm{Cl}_2(x)$ and $\mathrm{Cl}_2(x) = -\mathrm{Cl}_2(2\pi - x)$. A plot of the function over the domain $[0, 2\pi]$ is shown in Figure A.1. Numerous connections exist among Clausen's integral, the dilogarithm, and Lobachevsky's function. For instance, from the integral representation of Clausen's function it is easy to see that

$$L(x) \;=\; \frac{1}{2}\,\mathrm{Cl}_2(\pi - 2x) \;-\; x\,\log 2. \tag{A.23}$$

The counterpart to equation (A.18) that we shall use in reducing $\Upsilon(a,x)$ is

$$F(\alpha, x) \;\equiv\; \int_0^x \log(1 + \sin\alpha\,\cos\theta)\,d\theta. \tag{A.24}$$

In the mid 19th century F. W. Newman showed that this and many closely related integrals can be reduced to Clausen's integral [104]. One such identity derived by Newman [105, p. 88] (Also see Lewin [86, p. 308]) is

$$\begin{aligned} F(\alpha, x) \;&=\; x\,\log\left(\sin^2\frac{\alpha}{2}\right) - \eta\,\log\left(\tan^2\frac{\alpha}{2}\right) \\ &\quad -\mathrm{Cl}_2(2x) + \mathrm{Cl}_2(2x - 2\eta) + \mathrm{Cl}_2(2\eta), \end{aligned} \tag{A.25}$$

where $\tan\eta \equiv \sin x/(\tan\frac{x}{2} + \cos x)$. By means of this identity, the reduction of $\Upsilon(a,x)$ is straightforward:

$$\begin{aligned} \Upsilon(a,x) \;&=\; \frac{1}{2}\int_0^{2x} \log\left(1 + \frac{\cos\theta}{2a^2+1}\right)d\theta \;+\; x\,\log\left(\frac{2a^2+1}{2}\right) \;-\; 2\,L(x) \\ &=\; \frac{1}{2}\,F\left(\sin^{-1}\frac{1}{2a^2+1}\,,\,2x\right) \;+\; x\,\log\left(\frac{2a^2+1}{2}\right) \;-\; 2\,L(x). \end{aligned}$$

From the above and equations (A.22) and (A.25) it follows that $\Upsilon(a,x)$ can be expressed in terms of elementary functions and Clausen's integral. This representation rests upon a single special function of a real variable over a finite range.

A.8 The L_1 and L_∞ Norms of K

To compute the operator norm $||\,\mathbf{K}\,||_1$ we begin by considering the function norm $||\,\mathbf{K}f\,||_1$ for an arbitrary function $f \in L_1(\mathcal{M} \times \mathcal{S}^2, m \times \sigma)$. The derivation proceeds by manipulating the expression for $||\,\mathbf{K}f\,||_1$ to produce a product of $||\,f\,||_1$ and a new expression, which will be the operator norm $||\,\mathbf{K}\,||_1$. Since all quantities are positive, we shall drop the absolute value signs. Let f denote a field radiance function and observe that

$$\begin{aligned} ||\,\mathbf{K}f\,||_1 &= \int_{\mathcal{M}} \int_{\Omega_o} \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \rightarrow \mathbf{u})\, f(\mathbf{r}, \mathbf{u}')\; d\mu(\mathbf{u}')\; d\mu(\mathbf{u})\; dm(\mathbf{r}) \\ &= \int_{\mathcal{M}} \int_{\Omega_i} \int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \rightarrow \mathbf{u})\, f(\mathbf{r}, \mathbf{u}')\; d\mu(\mathbf{u})\; d\mu(\mathbf{u}')\; dm(\mathbf{r}) \\ &= \int_{\mathcal{M}} \int_{\Omega_i} f(\mathbf{r}, \mathbf{u}') \left[\, \int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \rightarrow \mathbf{u})\; d\mu(\mathbf{u}) \,\right] d\mu(\mathbf{u}')\; dm(\mathbf{r}), \end{aligned}$$

where the first equality follows from the definitions. The second equality holds by Fubini's theorem [132, p. 164], which states that the order of integration can be changed when the integral exists and is finite. In the final equality, we have isolated the kernel of the integral operator $\mathbf{K}$ to the extent possible.

Next, we introduce the constant ω defined by

$$\omega \;\equiv\; \operatorname*{ess\,sup}_{\mathbf{r} \in \mathcal{M}}\; \operatorname*{ess\,sup}_{\mathbf{u}' \in \Omega_i} \int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \rightarrow \mathbf{u})\, d\mu(\mathbf{u}), \tag{A.26}$$

which is independent of the function f. Since this quantity bounds the effect of the bracketed expression in the context of the outer double integral, we have

$$\begin{aligned} ||\,\mathbf{K}f\,||_1 &\leq \omega \int_{\mathcal{M}} \int_{\Omega_i} f(\mathbf{r}, \mathbf{u}')\; d\mu(\mathbf{u}')\; dm(\mathbf{r}) \\ &= \omega\, ||\,f\,||_1, \end{aligned} \tag{A.27}$$

which shows that ω is an *upper bound* on $||\,\mathbf{K}\,||_1$. To show that it is a lower bound as well we must show that this bound is either attained by some function f, or approached from below by a sequence of functions $f_1, f_2, \ldots$. We can accomplish the latter by a sequence of beams that approach perfect collimation about the incident direction $\mathbf{u}'$ in which

$$\int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, d\mu(\mathbf{u}) \tag{A.28}$$

is maximal. If this maximum is not attainable (e.g. the reflectivity may increase as the incident beam approaches grazing), then we let the sequenct $f_1, f_2, \ldots$ approach perfect collimation while simultaneously approaching grazing. From this it follows that ω is also a *lower bound* on $||\,\mathbf{K}\,||_1$. Therefore,

$$||\,\mathbf{K}\,||_1 = \underset{\mathbf{r}\in\mathcal{M}}{\text{ess sup}}\ \underset{\mathbf{u}'\in\Omega_i}{\text{ess sup}} \int_{\Omega_o} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, d\mu(\mathbf{u}). \tag{A.29}$$

The computation for the L_∞-norm proceeds similarly. For any $f \in L_\infty$ we have

$$\begin{aligned} ||\,\mathbf{K}f\,||_\infty &= \underset{\mathbf{r}\in\mathcal{M}}{\text{ess sup}}\ \underset{\mathbf{u}\in\Omega_o}{\text{ess sup}} \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, f(\mathbf{r}, \mathbf{u}')\; d\mu(\mathbf{u}') \\ &\leq \left[\ \underset{\mathbf{r}\in\mathcal{M}}{\text{ess sup}}\ \underset{\mathbf{u}\in\Omega_o}{\text{ess sup}} \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\; d\mu(\mathbf{u}')\ \right] ||\,f\,||_\infty, \end{aligned}$$

so the expression in brackets is an upper bound on $||\,\mathbf{K}\,||_\infty$. Again, this bound can be approached from below by considering a sequence of increasingly collimated beams about a direction of maximal reflectance. Therefore, we have

$$||\,\mathbf{K}\,||_\infty = \underset{\mathbf{r}\in\mathcal{M}}{\text{ess sup}}\ \underset{\mathbf{u}\in\Omega_o}{\text{ess sup}} \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, d\mu(\mathbf{u}'), \tag{A.30}$$

which differs from the L_1 norm by exchanging the arguments of the kernel. □□

A.9 Proof of theorem 12

In this appendix we provide the proof of theorem 12 on page 146. Although the proof of this theorem is given by Kato for finite-dimensional linear operators [77, p. 144], and is left as an exercise by Dunford and Schwartz [41, p. 518], the general theorem for infinite-dimensional operators does not appear to be widely known. For completeness, a proof which parallels that given by Kato is included below. To simplify notation the theorem is proved using functions of a single real variable.

Let $\mathbf{K}$ be the integral operator defined by

$$(\mathbf{K}\,u)(x) \equiv \int_{\mathcal{D}} k(x,y)\,u(y)\;d\mu(y),$$

where u and v are real-valued functions on some domain $\mathcal{D}$, μ is a positive measure on a σ-algebra over $\mathcal{D}$, and the kernel $k : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ is such that both $||\,\mathbf{K}\,||_1$ and $||\,\mathbf{K}\,||_\infty$ are finite. Letting $v = \mathbf{K}u$, we have

$$\begin{aligned} \frac{|v(x)|}{||\,\mathbf{K}\,||_\infty} &\le \int_{\mathcal{D}} \frac{|k(x,y)|}{||\,\mathbf{K}\,||_\infty}\,|u(y)|\;d\mu(y) \\ &= \int_{\mathcal{D}} |u(y)|\;d\widehat{\mu}_x(y), \end{aligned} \tag{A.31}$$

where $\widehat{\mu}_x$ is the positive measure defined by

$$\widehat{\mu}_x(E) \equiv \int_E \frac{|k(x,y)|}{||\,\mathbf{K}\,||_\infty}\,d\mu(y)$$

for all measurable $E \subset \mathcal{D}$. By the definition of $||\,\mathbf{K}\,||_\infty$ it follows that $\widehat{\mu}_x(\mathcal{D}) \le 1$ for all $x \in \mathcal{D}$. Now, let $p \ge 1$ be fixed and let ϕ denote the function $\phi(x) \equiv |x|^p$. Since ϕ is convex and $\widehat{\mu}_x(\mathcal{D}) \le 1$, we may apply Jensen's inequality [132, p. 63] to obtain

$$\phi\left[\,\int_{\mathcal{D}} |u(y)|\;d\widehat{\mu}_x(y)\,\right] \;\le\; \int_{\mathcal{D}} \phi\left(u(y)\right)\;d\widehat{\mu}_x(y), \tag{A.32}$$

for all $x \in \mathcal{D}$. By equation (A.31) we may rewrite equation (A.32) as

$$\left[\frac{|v(x)|}{||\,\mathbf{K}\,||_\infty}\right]^p \;\le\; \int_{\mathcal{D}} |u(y)|^p\,d\widehat{\mu}_x(y)$$

for all $x \in \mathcal{D}$. Multiplying both sides by $||\,\mathbf{K}\,||_\infty^p$ and integrating, we have

$$\begin{aligned}\int_{\mathcal{D}} |v(x)|^p \; d\mu(x) \;&\leq\; ||\,\mathbf{K}\,||_\infty^{p-1} \int_{\mathcal{D}}\int_{\mathcal{D}} |k(x,y)| \; |u(y)|^p \, d\mu(y) \, d\mu(x) \\ &=\; ||\,\mathbf{K}\,||_\infty^{p-1} \int_{\mathcal{D}} \left[\int_{\mathcal{D}} |k(x,y)| \; d\mu(x) \right] |u(y)|^p \, d\mu(y) \\ &\leq\; ||\,\mathbf{K}\,||_\infty^{p-1} \, ||\,\mathbf{K}\,||_1 \int_{\mathcal{D}} |u(y)|^p \, d\mu(y),\end{aligned}$$

where the interchange of the integrals in the middle equality is justified by the bounds on $\mathbf{K}$ and Fubini's theorem [132, p. 150]. Identifying the remaining integrals as L_p-norms, we have

$$||\,v\,||_p^p \;\leq\; ||\,\mathbf{K}\,||_\infty^{p-1} \; ||\,\mathbf{K}\,||_1 \; ||\,u\,||_p^p. \tag{A.33}$$

Raising both sides of equation (A.33) to the power $1/p$ yields

$$||\,v\,||_p \;\leq\; ||\,\mathbf{K}\,||_\infty^{(p-1)/p} \; ||\,\mathbf{K}\,||_1^{1/p} \; ||\,u\,||_p \tag{A.34}$$

for all u. Since $v = \mathbf{K}u$, equation (A.34) implies that

$$||\,\mathbf{K}\,||_p \;\leq\; ||\,\mathbf{K}\,||_\infty^{(p-1)/p} \; ||\,\mathbf{K}\,||_1^{1/p}, \tag{A.35}$$

by the definition of an operator norm. Finally, for any $A \geq 0$, $B \geq 0$, and $x \geq 1$, we have

$$A^{1-\frac{1}{x}} \; B^{\frac{1}{x}} \;\leq\; \max\{A, B\}^{1-\frac{1}{x}} \;\times\; \max\{A, B\}^{\frac{1}{x}} \;=\; \max\{A, B\}\,.$$

Thus, from equation (A.35) it follows that

$$||\,\mathbf{K}\,||_p \;\leq\; \max\{\; ||\,\mathbf{K}\,||_\infty, \; ||\,\mathbf{K}\,||_1 \;\},$$

for any $p \geq 1$, which completes the proof. □□

A.10 Proof of theorem 13

We shall prove theorem 13 on page 146 by extending theorem 12 slightly to accommodate the form of the local reflection operator $\mathbf{K}$. To begin, let $\mathbf{K_r}$ denote the local reflection operator restricted to the point $\mathbf{r} \in \mathcal{M}$. That is,

$$(\mathbf{K_r}\, h)(\mathbf{u}) \equiv \int_{\Omega_i} k(\mathbf{r}; \mathbf{u}' \to \mathbf{u})\, h(\mathbf{u}')\, d\mu(\mathbf{u}'), \tag{A.36}$$

where h is a radiance distribution function at $\mathbf{r}$. Then $\mathbf{K}$ and $\mathbf{K_r}$ are related by

$$(\mathbf{K}f)(\mathbf{r}, \mathbf{u}) = [\mathbf{K_r} f(\mathbf{r}, \cdot)]\, (\mathbf{u}). \tag{A.37}$$

From the known bounds on $\mathbf{K}$ it follows that $\|\, \mathbf{K_r} \,\|_1 = \|\, \mathbf{K_r} \,\|_\infty = \omega$ for almost every $\mathbf{r} \in \mathcal{M}$. From theorem 12 it then follows that $\|\, \mathbf{K_r} \,\|_p \le \omega$ for almost every $\mathbf{r} \in \mathcal{M}$ and for all $1 \le p \le \infty$, since $\mathbf{K_r}$ is a kernel operator in standard form. Because the L_p-norm corresponding to the restricted function h is

$$||\, h \,||_p \equiv \left[\int_{\mathcal{S}^2} |\, h(\mathbf{u}) |^p \, d\mu(\mathbf{u}) \right]^{1/p}, \tag{A.38}$$

we may express the L_p-norm for radiance functions as

$$||\, f \,||_p = \left[\int_{\mathcal{M}} ||\, f(\mathbf{r}, \cdot) \,||_p^p \, dm(\mathbf{r}) \right]^{1/p}. \tag{A.39}$$

The corresponding bound on $\mathbf{K}$ now follows easily by observing that

$$\begin{aligned} ||\, \mathbf{K}f \,||_p^p &= \int_{\mathcal{S}^2} ||\, \mathbf{K_r} f(\mathbf{r}, \cdot) \,||_p^p \, dm(\mathbf{r}) \\ &\le \int_{\mathcal{S}^2} \omega^p \, ||\, f(\mathbf{r}, \cdot) \,||_p^p \, dm(\mathbf{r}) \\ &= \omega^p \, ||\, f \,||_p^p \end{aligned}$$

for all f. Therefore, $||\, \mathbf{K} \,||_p \le \omega$. □□

Bibliography

[1] Abramowitz, M., and Stegun, I. A., Eds. *Handbook of Mathematical Functions.* Dover Publications, New York, 1965.

[2] Acton, F. S. *Numerical Methods that Work.* Harper & Row, New York, 1970.

[3] Amanatides, J. Ray tracing with cones. *Computer Graphics 18*, 3 (July 1984), 129–135.

[4] Anselone, P. M. Convergence and error bounds for approximate solutions of integral and operator equations. In *Error in Digital Computation*, L. B. Rall, Ed., vol. 2. John Wiley & Sons, 1965, pp. 231–252.

[5] Appel, A. Some techniques for shading machine renderings of solids. In *Proceedings of the Spring Joint Computer Conference* (1968), pp. 37–45.

[6] Arvo, J. Transfer equations in global illumination. In *Global Illumination, SIGGRAPH '93 Course Notes* (August 1993), vol. 42.

[7] Arvo, J. Applications of irradiance tensors to the simulation of non-lambertian phenomena. In *Computer Graphics* Proceedings (1995), Annual Conference Series, ACM SIGGRAPH.

[8] Arvo, J., and Kirk, D. Particle transport and image synthesis. *Computer Graphics 24*, 4 (August 1990), 63–66.

[9] Arvo, J., and Novins, K. Iso-contour volume rendering. In *1994 Symposium on Volume Visualization* (October 1994).

[10] Ashour, A., and Sabri, A. Tabulation of the function $\psi(\theta) = \sum_{n=1}^{\infty} \sin n\theta/n^2$. *Mathematical Tables and other Aids to Computation 10*, 54 (April 1956), 57–65.

[11] Atkinson, K. E. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind.* Society for Industrial and Applied Mathematics, Philadelphia, 1976.

[12] Aupperle, L., and Hanrahan, P. A hierarchical illumination algorithm for surfaces with glossy reflection. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 155–162.

[13] Baltes, H. P. Radiometry and coherence. In *Inverse Source Problems in Optics*, H. P. Baltes, Ed. Springer-Verlag, New York, 1978, ch. 5.

[14] Banach, S., and Tarski, A. Sur la décomposition des ensembles de points en parties respectivement congruentes. *Fundamenta Mathematicae 6* (1924), 244–277.

[15] Banach, S. S. *Theory of Linear Operations.* North-Holland, Amsterdam, 1987. Translated by F. Jellett.

[16] Bastos, R. M., de Sousa, A. A., and Ferreira, F. N. Reconstruction of illumination functions using hermite bicubic interpolation. In *Proceedings of the Fourth Eurographics Workshop on Rendering,* Paris, France (June 1993).

[17] Bateman, H. Report on the history and present state of the theory of integral equations. *Report of the 18th Meeting of the British Association for the Advancement of Science* (1910), 345–424.

[18] Baum, D. R., Rushmeier, H. E., and Winget, J. M. Improving radiosity solutions through the use of analytically determined form-factors. *Computer Graphics 23*, 3 (July 1989), 325–334.

[19] Berger, M. *Geometry, Volume II.* Springer-Verlag, New York, 1987. Translated by M. Cole and S. Levy.

[20] Bird, R. B., Stewart, W. E., and Lightfoot, E. N. *Transport Phenomena.* John Wiley & Sons, New York, 1960.

[21] Bishop, R. L., and Goldberg, S. I. *Tensor Analysis on Manifolds.* Dover Publications, New York, 1980.

[22] Borel, C. C., Gerstl, S. A. W., and Powers, B. J. The radiosity method in optical remote sensing of structured 3-d surfaces. *Remote Sensing of the Environment 36* (1991), 13–44.

[23] Buckalew, C., and Fussell, D. Illumination networks: Fast realistic rendering with general reflectance functions. *Computer Graphics 23*, 3 (July 1989), 89–98.

[24] Carlson, B. C. Invariance of an integral average of a logarithm. *The American Mathematical Monthly 82*, 4 (April 1975), 379–382.

[25] Case, K. M., de Hoffmann, F., and Placzek, G. *Introduction to the Theory of Neutron Diffusion*, vol. 1. Los Alamos Scientific Laboratory, Los Alamos, New Mexico, 1953.

[26] Chandrasekar, S. *Radiative Transfer.* Dover Publications, New York, 1960.

[27] Chew, L. P. Constrained delaunay triangulations. In *Proceedings of the Third Annual Symposium on Computational Geometry* (Waterloo, Ontario, Canada, June 1987), pp. 215–222.

[28] Christensen, P. H., Stollnitz, E. J., Salesin, D. H., and DeRose, T. D. Wavelet radiance. In *Proceedings of the Fifth Eurographics Workshop on Rendering,* Darmstadt, Germany (1994), pp. 287–302.

[29] Clark, F. H. Methods and data for reactor shield calculations. In *Advances in Nuclear Science and Technology*, E. J. Henley and J. Lewins, Eds., vol. 5. Plenum Press, New York, 1971, pp. 95–183.

[30] Cohen, M. F., Chen, S. E., Wallace, J. R., and Greenberg, D. P. A progressive refinement approach to fast radiosity image generation. *Computer Graphics 22*, 4 (August 1988), 75–84.

[31] Cohen, M. F., and Greenberg, D. P. The hemi-cube: A radiosity solution for complex environments. *Computer Graphics 19*, 3 (July 1985), 75–84.

[32] Cohen, M. F., and Wallace, J. R. *Radiosity and Realistic Image Synthesis.* Academic Press, New York, 1993.

[33] Conway, D. M., and Cottingham, M. S. The isoluminance contour model. In *Proceedings of Ausgraph '88* (Melbourne, Australia, September 1988), pp. 43–50.

[34] Cook, R. L. Distributed ray tracing. *Computer Graphics 18*, 3 (July 1984), 137–145.

[35] Cook, R. L. Stochastic sampling in computer graphics. *ACM Transactions on Graphics 5*, 1 (1986), 51–72.

[36] DiLaura, D. L., and Quinlan, J. Non-diffuse radiative transfer I: Planar area sources and point receivers. In *Illumination Engineering Society of North America, Annual Conference Technical Papers* (Miami, Florida, 1994), pp. 633–645.

[37] Dorsey, J. O., Sillion, F. X., and Greenberg, D. P. Design and simulation of opera lighting and projection effects. *Computer Graphics 25*, 4 (August 1991), 41–50.

[38] Drettakis, G., and Fiume, E. A fast shadow algorithm for area light sources using backprojection. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 223–230.

[39] Drettakis, G., and Fiume, E. L. Concrete computation of global illumination using structured sampling. In *Proceedings of the Third Eurographics Workshop on Rendering,* Bristol, United Kingdom (1992).

[40] Duderstadt, J. J., and Martin, W. R. *Transport Theory.* John Wiley & Sons, New York, 1979.

[41] Dunford, N., and Schwartz, J. T. *Linear Operators. Part I: General Theory.* John Wiley & Sons, New York, 1967.

[42] Flanders, H. *Differential Forms with Applications to the Physical Sciences.* Dover Publications, New York, 1989.

[43] Fok, V. A. The illumination from surfaces of arbitrary shape. *Transactions of the Optical Institute, Leningrad 28* (1924), 1–11. (Russian).

[44] Freund, J. E., and Walpole, R. E. *Mathematical Statistics*, fourth ed. Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[45] Gershbein, R., Schröder, P., and Hanrahan, P. Textures and radiosity: Controlling emission and reflection with texture maps. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 51–58.

[46] Gershun, A. The light field. *Journal of Mathematics and Physics 18*, 2 (May 1939), 51–151. Translated by P. Moon and G. Timoshenko.

[47] Glassner, A. *Principles of Digital Image Synthesis.* Morgan Kaufmann, New York, 1995.

[48] Goel, N. S., Rozehnal, I., and Thompson, R. L. A computer graphics based model for scattering from objects of arbitrary shapes in the optical region. *Remote Sensing of the Environment 36* (1991), 73–104.

[49] Goertler, S., Cohen, M., and Slusallek, P. Radiosity and relaxation methods. *IEEE Computer Graphics and Applications 14* (November 1994), 48–58.

[50] Golberg, M. A. A survey of numerical methods for integral equations. In *Solution methods for integral equations: Theory and applications*, M. A. Golberg, Ed. Plenum Press, New York, 1979, pp. 1–58.

[51] Gondek, J. S., Meyer, G. W., and Newman, J. G. Wavelength dependent reflectance functions. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 213–220.

[52] Goral, C. M., Torrance, K. E., Greenberg, D. P., and Battaile, B. Modeling the interaction of light between diffuse surfaces. *Computer Graphics 18*, 3 (July 1984), 213–222.

[53] Gortler, S. J., Schröder, P., Cohen, M. F., and Hanrahan, P. Wavelet radiosity. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 221–230.

[54] Grad, H. Note on N-dimensional Hermite polynomials. *Communications on Pure and Applied Mathematics 2* (1949), 325–330.

[55] Gradshteyn, I. S., and Ryzhik, I. M. *Table of Integrals, Series, and Products*, fifth ed. Academic Press, New York, 1994.

[56] Gröbner, W., and Hofreiter, N. *Integraltafel zweiter teil Bestimmte Integrale*, third ed. Springer-Verlag, Wien, 1961.

[57] Guillemin, V., and Pollack, A. *Differential Topology.* Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[58] Haas, A. The multiple prime random number generator. *ACM Transactions on Mathematical Software 13*, 4 (December 1987), 368–381.

[59] Halmos, P. R. *Measure Theory*, vol. 18 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1974.

[60] Halton, J. H. A retrospective and prospective survey of the Monte Carlo method. *SIAM Review 12*, 1270 (1970), 1–63.

[61] Hammersley, J. M., and Handscomb, D. C. *Monte Carlo Methods.* Chapman and Hall, New York, 1964.

[62] Hanrahan, P., Salzman, D., and Aupperle, L. A rapid hierarchical radiosity algorithm. *Computer Graphics 25*, 4 (July 1991), 197–206.

[63] He, X. D., Torrance, K. E., Sillion, F. X., and Greenberg, D. P. A comprehensive physical model for light reflection. *Computer Graphics 25*, 4 (July 1991), 175–186.

[64] Heckbert, P. S. *Simulating Global Illumination Using Adaptive Meshing.* PhD thesis, University of California, Berkeley, June 1991.

[65] Hendry, W. L., Lathrop, K. D., Vandervoort, S., and Wooten, J. Bibliography on neutral particle transport theory. Tech. Rep. LA-4287-MS, United States Atomic Energy Commission, 1970.

[66] Hildebrand, F. B. *Methods of Applied Mathematics.* Prentice-Hall, New York, 1952.

[67] Holmes, J. G. A method of plotting isolux curves. *Light and Lighting 39* (1946), 158–160.

[68] Hottel, H. C., and Sarofim, A. F. *Radiative Transfer.* McGraw-Hill, New York, 1967.

[69] Howell, J. R. *A Catalog of Radiation Configuration Factors.* McGraw-Hill, New York, 1982.

[70] Immel, D. S., Cohen, M. F., and Greenberg, D. P. A radiosity method for non-diffuse environments. *Computer Graphics 20*, 4 (August 1986), 133–142.

[71] Kajiya, J. T. The rendering equation. *Computer Graphics 20*, 4 (August 1986), 143–150.

[72] Kajiya, J. T. Radiometry and photometry for computer graphics. In *Advanced Topics in Ray Tracing, SIGGRAPH 90 Course Notes* (August 1990), vol. 24.

[73] Kajiya, J. T., and Von Herzen, B. P. Ray tracing volume densities. *Computer Graphics 18*, 3 (July 1984), 165–173.

[74] Kalos, M. H., and Whitlock, P. A. *Monte Carlo Methods, Volume I: Basics.* John Wiley & Sons, New York, 1986.

[75] Kantorovich, L., and Akilov, G. P. *Functional Analysis in Normed Spaces.* Pergamon Press, New York, 1964.

[76] Kantorovich, L. V., and Krylov, V. I. *Approximate Methods of Higher Analysis.* Interscience Publishers, New York, 1958. Translated by C. D. Benster.

[77] Kato, T. *Perturbation Theory for Linear Operators.* Springer-Verlag, New York, 1966.

[78] Kline, M., and Kay, I. W. *Electromagnetic Theory and Geometrical Optics.* John Wiley & Sons, New York, 1965.

[79] Kourganoff, V. *Basic Methods in Transfer Problems: Radiative Equilibrium and Neutron Diffusion.* Oxford University Press, New York, 1952.

[80] Kourganoff, V. *Introduction to the General Theory of Particle Transfer.* Gordon and Breach, New York, 1969.

[81] Krasnosel'skii, M. A., Vainikko, G. M., Zabreiko, P. P., Rutitskii, Y. B., and Stetsenko, V. Y. *Approximate Solution of Operator Equations.* Wolters-Noordhoff, Groningen, The Netherlands, 1972.

[82] Kress, R. *Linear Integral Equations.* Springer-Verlag, New York, 1989.

[83] Kreyszig, E. *Introductory Functional Analysis with Applications.* John Wiley & Sons, New York, 1978.

[84] Krook, M. On the solution of equations of transfer, I. *Astrophysical Journal 122*, 3 (November 1955), 488–497.

[85] Lambert, J. H. *Photometria, sive De mensura et gradibus luminis, colorum et umbrae.* No. 31-33 in Ostwald's Klassiker der exakten Wissenschaften. W. Engelmann, Leipzig, 1892.

[86] Lewin, L. *Polylogarithms and associated functions.* North Holland, New York, 1981.

[87] Lewins, J. *Importance, The Adjoint Function: The Physical Basis of Variational and Perturbation Theory in Transport and Diffusion Problems.* Pergamon Press, New York, 1965.

[88] Lewis, E. E., and W. F. Miller, Jr. *Computational Methods of Neutron Transport.* John Wiley & Sons, New York, 1984.

[89] Lichtenberg, A. J. *Phase-Space Dynamics of Particles.* John Wiley & Sons, New York, 1969.

[90] Lifshitz, L. M., and Pizer, S. M. A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*, 6 (June 1990), 529–540.

[91] Linz, P. *Theoretical Numerical Analysis, an Introduction to Advanced Techniques.* John Wiley & Sons, New York, 1979.

[92] Lischinski, D., Smits, B., and Greenberg, D. P. Bounds and error estimates for radiosity. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 67–74.

[93] Lischinski, D., Tampieri, F., and Greenberg, D. P. Discontinuity meshing for accurate radiosity. *IEEE Computer Graphics and Applications 12*, 6 (November 1992), 25–39.

[94] Lischinski, D., Tampieri, F., and Greenberg, D. P. Combining hierarchical radiosity and discontinuity meshing. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 199–208.

[95] Mackerle, J., and Brebbia, C. A., Eds. *The Boundary Element Reference Book.* Springer-Verlag, New York, 1988.

[96] Menzel, D. H., Ed. *Selected Papers on the Transfer of Radiation.* Dover Publications, New York, 1966.

[97] Mihalas, D., and Mihalas, B. W. *Foundations of Radiation Hydrodynamics.* Oxford University Press, New York, 1984.

[98] Milne, E. A. Thermodynamics of the stars. In *Handbuch der Astrophysik,* volume 3, part I, E. A. Milne, A. Pannekoek, S. Rosseland, and W. Westphal, Eds. Springer-Verlag, Berlin, 1930, ch. 2, pp. 65–255.

[99] Modest, M. F. *Radiative Heat Transfer.* McGraw-Hill, New York, 1993.

[100] Moon, P. *The Scientific Basis of Illuminating Engineering.* McGraw-Hill, New York, 1936.

[101] Moon, P., and Spencer, D. E. *Lighting Design.* Addison-Wesley, Cambridge, Massachusetts, 1948.

[102] Myneni, R. B., and Ross, J., Eds. *Photon-Vegetation Interactions.* Springer-Verlag, New York, 1991.

[103] Myneni, R. B., Ross, J., and Asrar, G. A review on the theory of photon transport in leaf canopies. *Agricultural and Forest Meterology 45*, 1–2 (February 1989), 1–153.

[104] Newman, F. W. On logarithmic integrals of the second order. *The Cambridge and Dublin Mathematical Journal II* (1847), 77–100.

[105] Newman, F. W. *The Higher Trigonometry and Superrationals of Second Order.* Macmillan and Bowes, 1892.

[106] Nimroff, J. S., Simoncelli, E., and Dorsey, J. Efficient re-rendering of naturally illuminated environments. In *Proceedings of the Fifth Eurographics Workshop on Rendering,* Darmstadt, Germany (1994), pp. 359–373.

[107] Nishita, T., and Nakamae, E. Half-tone representation of 3-D objects illuminated by area sources or polyhedron sources. In *Proceedings of the IEEE Computer Software and Applications Conference* (Chicago, November 1983), pp. 237–242.

[108] Nishita, T., and Nakamae, E. Continuous tone representation of 3-D objects taking account of shadows and interreflection. *Computer Graphics 19*, 3 (July 1985), 23–30.

[109] Ortega, J. M. *Numerical Analysis, a Second Course.* Academic Press, New York, 1972.

[110] Özisik, M. N. *Radiative Transfer and Interactions with Conduction and Convection.* John Wiley & Sons, New York, 1973.

[111] Pai, S. *Radiation Gas Dynamics.* Springer-Verlag, New York, 1966.

[112] Phillips, J. L. The use of collocation as a projection method for solving linear operator equations. *SIAM Journal on Numerical Analysis 9*, 1 (1972), 14–28.

[113] Phong, B. T. Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (June 1975), 311–317.

[114] Pietsch, A. *Operator Ideals.* North-Holland, New York, 1980.

[115] Planck, M. *The Theory of Heat Radiation.* Dover Publications, New York, 1988.

[116] Plantinga, H., and Dyer, C. R. Visibility, occlusion, and the aspect graph. *International Journal of Computer Vision 5*, 2 (1990), 137–160.

[117] Polyak, G. L. Radiative transfer between surfaces of arbitrary spatial distribution of reflection. In *Convective and Radiative Heat Transfer.* Publishing House of the Academy of Sciences of the USSR, Moscow, 1960.

[118] Pomraning, G. C. *The Equations of Radiation Hydrodynamics.* Pergamon Press, New York, 1973.

[119] Preisendorfer, R. W. *A Mathematical Foundation for Radiative Transfer Theory.* PhD thesis, University of California at Los Angeles, May 1956.

[120] Preisendorfer, R. W. A mathematical foundation for radiative transfer theory. *Journal of Mathematics and Mechanics 6*, 6 (November 1957), 685–730.

[121] Preisendorfer, R. W. Radiative transfer axioms. Tech. Rep. 57-44, Scripps Institution of Oceanography, Visibility Laboratory, University of California, San Diego, October 1957.

[122] Preisendorfer, R. W. *Radiative Transfer on Discrete Spaces.* Pergamon Press, New York, 1965.

[123] Preisendorfer, R. W. *Hydrologic Optics, Volumes I–VI.* National Oceanic & Atmospheric Administration, Honolulu, Hawaii, 1976. (Available as NTIS PB-259 793).

[124] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. *Numerical Recipes.* Cambridge University Press, New York, 1986.

[125] Rea, M. S. *Lighting Handbook*, eighth ed. Illuminating Engineering Society of North America, New York, 1993.

[126] Retherford, J. R. *Hilbert Space: Compact Operators and the Trace Theorem.* Cambridge University Press, New York, 1993.

[127] Robinson, R. M. On the decomposition of spheres. *Fundamenta Mathematicae 34* (1947), 246–260.

[128] Royden, H. L. *Real Analysis*, second ed. The Macmillan Company, New York, 1968.

[129] Rubinstein, R. Y. *Simulation and the Monte Carlo Method.* John Wiley & Sons, New York, 1981.

[130] Rudin, W. *Principles of Mathematical Analysis*, third ed. McGraw-Hill, New York, 1964.

[131] Rudin, W. *Functional Analysis.* McGraw-Hill, New York, 1973.

[132] Rudin, W. *Real and Complex Analysis*, second ed. McGraw-Hill, New York, 1974.

[133] Rushmeier, H. E., Patterson, C., and Veerasamy, A. Geometric simplification for indirect illumination calculations. *Graphics Interface '93* (May 1993), 227–236.

[134] Rushmeier, H. E., and Torrance, K. E. The zonal method for calculating light intensities in the presence of a participating medium. *Computer Graphics 21*, 4 (July 1987), 293–302.

[135] Rushmeier, H. E., and Torrance, K. E. Extending the radiosity method to include specularly reflecting and translucent materials. *ACM Transactions on Graphics 9*, 1 (January 1990), 1–27.

[136] Salesin, D., Lischninski, D., and DeRose, T. Reconstructing illumination functions with selected discontinuities. In *Proceedings of the Third Eurographics Workshop on Rendering,* Bristol, United Kingdom (May 1992), pp. 99–112.

[137] Schmid, E. *Beholding as in a Glass.* Herder and Herder, New York, 1969.

[138] Schrage, L. A more portable fortran random number generator. *ACM Transactions on Mathematical Software 5*, 2 (June 1979), 132–138.

[139] Schreiber, M. *Differential Forms: A Heuristic Introduction.* Springer-Verlag, New York, 1984.

[140] Schröder, P., and Hanrahan, P. On the form factor between two polygons. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 163–164.

[141] Schröder, P., and Hanrahan, P. Wavelet methods for radiance computations. In *Proceedings of the Fifth Eurographics Workshop on Rendering,* Darmstadt, Germany (1994), pp. 303–311.

[142] Schuster, A. Radiation through a foggy atmosphere. *Astrophysical Journal 21*, 1 (January 1905), 1–22.

[143] Schwarzschild, K. On the equilibrium of the sun's atmosphere. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen. Math.-phys. Klasse 195* (1906), 41–53. (See Menzel [96] for English translation).

[144] Segal, I. E., and Kunze, R. A. *Integrals and Operators*, second ed. Springer-Verlag, New York, 1978.

[145] Sherman, M. P. Moment methods in radiative transfer problems. *Journal of Quantitative Spectroscopy and Radiative Transfer 7*, 89–109 (1967).

[146] Shirley, P. A ray tracing algorithm for global illumination. *Graphics Interface '90* (May 1990), 205–212.

[147] Shirley, P., and Wang, C. Distribution ray tracing: Theory and practice. In *Proceedings of the Third Eurographics Workshop on Rendering,* Bristol, United Kingdom (1992), pp. 33–43.

[148] Siegel, R., and Howell, J. R. *Thermal Radiation Heat Transfer*, second ed. Hemisphere Publishing Corp., New York, 1981.

[149] Sillion, F., Arvo, J., Westin, S., and Greenberg, D. P. A global illumination solution for general reflectance distributions. *Computer Graphics 25*, 4 (July 1991), 187–196.

[150] Sillion, F., and Puech, C. A general two-pass method integrating specular and diffuse reflection. *Computer Graphics 23*, 4 (August 1989).

[151] Smits, B., Arvo, J., and Salesin, D. An importance-driven radiosity algorithm. *Computer Graphics 26*, 4 (July 1992), 273–282.

[152] Smits, B., and Meyer, G. W. Newton's colors: Simulating interference phenomena in realistic image synthesis. In *Proceedings of Eurographics Workshop on Photosimulation, Realism, and Physics in Computer Graphics,* Rennes, France (June 1990), pp. 185–194.

[153] Spanier, J., and Gelbard, E. M. *Monte Carlo Principles and Neutron Transport Problems.* Addison-Wesley, Reading, Massachusetts, 1969.

[154] Sparrow, E. M. Application of variational methods to radiative heat-transfer calculations. *ASME Journal of Heat Transfer 82*, 4 (November 1960), 375–380.

[155] Sparrow, E. M. A new and simpler formulation for radiative angle factors. *ASME Journal of Heat Transfer 85*, 2 (May 1963), 81–88.

[156] Spivak, M. *Calculus on Manifolds.* Benjamin/Cummings, Reading, Massachusetts, 1965.

[157] Stewart, A. J., and Ghali, S. Fast computation of shadow boundaries using spatial coherence and backprojections. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 231–238.

[158] Stewart, J. C. Some topics in radiative transfer. In *Developments in Transfer Theory*, E. Inönü and P. F. Zweifel, Eds. Academic Press, New York, 1967, pp. 113–148.

[159] Taylor, A. E., and Lay, D. C. *Introduction to Functional Analysis*, second ed. John Wiley & Sons, New York, 1980.

[160] Teller, S. Computing the antipenumbra of an area light source. *Computer Graphics 26*, 2 (July 1992), 139–148.

[161] Torrance, K. E., and Sparrow, E. M. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America 57*, 9 (September 1967), 1105–1114.

[162] Touloukian, Y. S., Ed. *Retrieval Guide to Thermophysical Properties Research Literature*, second ed. McGraw-Hill, New York, 1968.

[163] Trotter, A. P. *Illumination: Its Distribution and Measurement.* The Macmillan Company, London, 1911.

[164] Troutman, R., and Max, N. L. Radiosity algorithms using higher-order finite element methods. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 209–212.

[165] Turk, G. Generating random points in triangles. In *Graphics Gems*, A. S. Glassner, Ed. Academic Press, New York, 1990, pp. 24–28.

[166] Vedel, C. Improved storage and reconstruction of light intensities on surfaces. In *Proceedings of the Third Eurographics Workshop on Rendering,* Bristol, United Kingdom (May 1992), pp. 113–121.

[167] Vedel, C. Computing illumination from area light sources by approximate contour integration. In *Graphics Interface '93* (1993), pp. 237–244.

[168] Verbeck, C. P., and Greenberg, D. P. A comprehensive light-source description for computer graphics. *IEEE Computer Graphics and Applications 4*, 7 (July 1984), 66–75.

[169] Vincenti, W. G., and Kruger, Jr., C. H. *Introduction to Physical Gas Dynamics.* John Wiley & Sons, New York, 1965.

[170] Wallace, J., Cohen, M. F., and Greenberg, D. P. A two-pass solution to the rendering equation: A synthesis of ray tracing and radiosity methods. *Computer Graphics 21*, 3 (July 1987), 311–320.

[171] Wallace, J., Elmquist, K., and Haines, E. A ray tracing algorithm for progressive radiosity. *Computer Graphics 23*, 3 (July 1989), 315–324.

[172] Wang, C. Physically correct direct lighting for distribution ray tracing. In *Graphics Gems III*, D. Kirk, Ed. Academic Press, New York, 1992, pp. 301–312.

[173] Ward, G. J. Measuring and modeling anisotropic reflection. *Computer Graphics 26*, 2 (July 1992), 265–272.

[174] WARD, G. J. The RADIANCE lighting simulation and rendering system. In *Computer Graphics* Proceedings (1994), Annual Conference Series, ACM SIGGRAPH, pp. 459–472.

[175] WARD, G. J., AND HECKBERT, P. S. Irradiance gradients. In *Proceedings of the Third Eurographics Workshop on Rendering,* Bristol, United Kingdom (May 1992), pp. 85–98.

[176] WEILER, K., AND ATHERTON, P. Hidden surface removal using polygon area sorting. *Computer Graphics 11*, 3 (1977), 214–222.

[177] WESTIN, S., ARVO, J., AND TORRANCE, K. Predicting reflectance functions from complex surfaces. *Computer Graphics 26*, 2 (July 1992), 255–264.

[178] WHITTED, T. An improved illumination model for shaded display. *Communications of the ACM 32*, 6 (June 1980), 343–349.

[179] WIGNER, E. P. Mathematical problems of nuclear reactor theory. In *Nuclear Reactor Theory. Proceedings of the Eleventh Symposium in Applied Mathematics.* American Mathematical Society, Providence, Rhode Island, 1961, pp. 89–104.

[180] WOLF, L. B. Diffuse reflection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 23*, 4 (June 1992), 472–478.

[181] YAMAUTI, Z. Further study of geometrical calculations of illumination due to light from lumnious surface sources of simple forms. Tech. Rep. 194, Researches of the Electrotechnical Laboratory, Ministry of Communications, Tokyo, Japan, 1927.

[182] YAMAUTI, Z. Theory of field of illumination. Tech. Rep. 339, Researches of the Electrotechnical Laboratory, Ministry of Communications, Tokyo, Japan, October 1932.

[183] YOSIDA, K. *Functional Analysis*, sixth ed. Springer-Verlag, New York, 1980.

[184] ZATZ, H. Galerkin radiosity: A higher order solution method for global illumination. In *Computer Graphics* Proceedings (1993), Annual Conference Series, ACM SIGGRAPH, pp. 213–220.