Phase Space Ray Tracing for Illumination Optics

Carmela Filosa

Cover art:
Photography:

A catalogue record is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-3972-7

title

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen

door

Carmela Filosa

geboren te Torre del greco, Italië

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.
1e promotor:	prof.dr. W.L. IJzerman
copromotor:	dr. J.H.M. ten Thije Boonkkamp
leden:	

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

List of symbols

τ	time
Q	Total energy emitted from a light sorce or received by a target
Φ_r	Radiant flux
Φ	Luminous flux
λ	Wavelength
Ψ_r	Power per wavelength
$\bar{y}(\lambda)$	Luminousity function
E	Illuminance
$\mathrm{d}\Omega$	Solid angle
I	Intensity
L	Luminance
U	étendue
$\boldsymbol{\nu}$	Surface normal
n	Index of refraction of the medium in which a surface is immersed
θ	Angle between the direction of the solid angle and the normal $\boldsymbol{\nu}$
n_i	Index of refraction of the medium in which the incident ray travels
$n_\mathrm{r} = n_\mathrm{i}$	Index of refraction of the medium in which the reflected ray is located
n_t	Index of refraction of the medium in which the transmitted ray travels
$n_\mathrm{i,t}$	$\frac{n_\mathrm{i}}{n_\mathrm{t}}$
θ_i	Angle between the incident ray and the normal $\boldsymbol{\nu}$
θ_r	Angle between the reflected ray and the normal $\boldsymbol{\nu}$
θ_t	Angle between the transmitted ray and the normal $\boldsymbol{\nu}$
θ_c	Critical angle
$\boldsymbol{t}_\mathrm{i}$	Direction of the incident ray
$\boldsymbol{t}_\mathrm{r}$	Direction of the reflected ray
$\boldsymbol{t}_\mathrm{t}$	Direction of the transmitted ray
$\boldsymbol{\nu}_j$	Normal to the line j
t_j	Angle that the ray located on line j forms with respect to the optical axis
θ_j	Angle between the ray and the normal $\boldsymbol{\nu}_j$ to line j
n_j	Index of refraction of the medium in which line j is located

Chapter 1

Introduction

1.1 Motivation

1.2 Methods and results

1.3 Content of this thesis

Chapter 2

Illumination optics

This chapter provides some concepts of illumination optics used in this thesis. We start explaining the difference between radiometry and photometry. In particular, we focus on the photometric variables, defining them both in three and two dimensions. The reflection and refraction laws and the phenomenon of total internal reflection are explained. The last paragraph of the chapter gives a brief introduction to Fresnel reflection.

2.1 Radiometric and photometric variables

Radiometry is concerned with the measurement of electromagnetic radiation across the entire electromagnetic spectrum. Photometry is the subfield of radiometry that takes into account only the portion of the electromagnetic spectrum corresponding to the visible light, [1]. Radiometry deals with radiometric quantities. An important radiometric quantity is the radiant flux Φ_{r} (unit watt, [W]) which is the total energy emitted from a source or received by a target per unit time:

$$\Phi_{\mathrm{r}} = \frac{\mathrm{d}Q}{\mathrm{d}T}\,, \tag{2.1.1}$$

where Q is the energy and T the time.

In illumination optics the measurement of light is given in terms of the impression that it gives on the human eye. Therefore, illumination optics deals with photometric variables. The most important photometric variables are defined as follows using the same notation adopted by Chaves in [2]. The luminous flux Φ (unit lumen, [lm]) is defined as the perceived power of light by the human eye. The radiant and the luminous flux are related by the luminous efficacy function, unit [lm/W], which tells us how many lumen there are for each Watt of power at a given wavelength. The luminous efficacy reaches its maximum at a wavelength of 555 nm where it is equal to 683 lm/W. We may normalize the luminous efficacy function with its maximum value of 683. This normalized function is the dimensionless luminosity function $\bar{y}(\lambda)$ shown in Figure 2.1 where λ is the wavelength.

The luminous flux corresponding to one Watt of radiation power at any wavelength is given by the product of 683 lm/W and the luminosity function at the same wavelength,

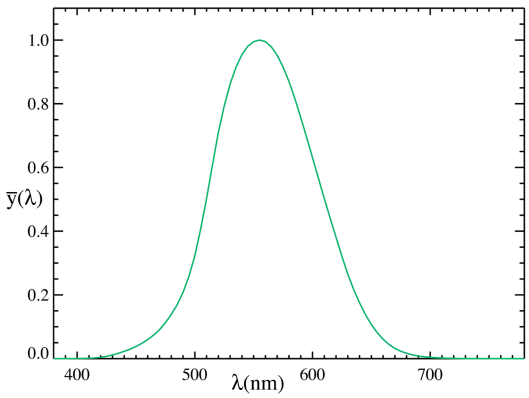


Figure 2.1: Luminosity function $\bar{y}(\lambda)$: relation between the eye's sensitivity and the wavelength of light. The luminosity function is dimensionless, [3].

i.e. $683\,\bar{y}(\lambda)$. Hence, the total luminous flux Φ has unit lumen [lm] and it is defined as:

$$\Phi = 683 \int_0^\infty \Psi_\mathrm{r}(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda \tag{2.1.2}$$

where $\Psi_\mathrm{r}(\lambda)$ is the power in watt per unit wavelength (unit [W/m]).

A beam of light can be described as a collection of parallel light rays, where a light ray can be interpreted as a path along which the energy travels. The luminous flux $\mathrm{d}\Phi$ incident on a surface is called illuminance E $\left(\text{unit } \left[\mathrm{lm/m^2}\right]\right)$ and is defined as:

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}A}\,, \tag{2.1.3}$$

where $\mathrm{d}A$ is an infinitesimal area receiving radiation. The density of light emitted by a point source in a given direction is determined by the solid angle. The solid angle on a given direction is defined by the infinitesimal surface area $\mathrm{d}S$ of a sphere subtended by the radius of that sphere and by the rays emitted by the center on that direction, [4]. Indicating with r the radius of the sphere, the infinitesimal solid angle $\mathrm{d}\Omega$ defined by $\mathrm{d}S$ is given by:

$$\mathrm{d}\Omega = \frac{\mathrm{d}S}{r^2}\,. \tag{2.1.4}$$

The solid angle on the entire sphere is $\Omega = 4\pi$, its unit is the steradian [sr] and it is usually defined on a unit sphere. The luminous intensity I (unit candela (cd), [cd = lm/sr]) is defined as the luminous flux $\mathrm{d}\Phi$ per solid angle $\mathrm{d}\Omega$ and is given by:

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega}\,. \tag{2.1.5}$$

The luminance L $\left(\text{unit } \left[\mathrm{cd/m^2}\right]\right)$ is the luminous flux per unit solid angle $\mathrm{d}\Omega$ and per unit projected area $\cos\theta \mathrm{d}A$ where θ is the angle that the normal $\boldsymbol{\nu}$ to the area $\mathrm{d}A$ makes with the direction of the solid angle $\mathrm{d}\Omega$, as shown in Figure 2.2. L is given by:

$$L = \frac{\mathrm{d}\Phi}{\cos\theta \mathrm{d}A \mathrm{d}\Omega}\,. \tag{2.1.6}$$

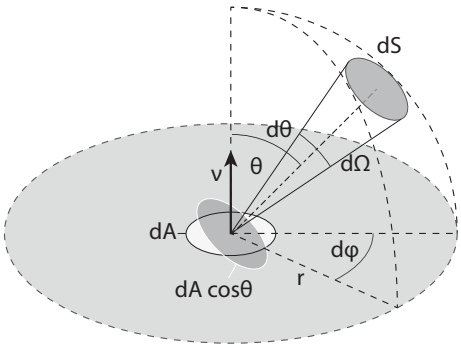


Figure 2.2: Solid angle $\mathrm{d}\Omega$ in a direction making an angle θ with the normal to the area $\mathrm{d}A$.

Note that from (2.1.5) and (2.1.6) we can derive a relation between the intensity and the luminance. The intensity I emitted by the infinitesimal area $\mathrm{d}A$ is given by:

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} = L(\boldsymbol{x}, \theta) \cos\theta \mathrm{d}A\,. \tag{2.1.7}$$

When the luminance is uniform over a finite area A, the luminous intensity emitted in the direction θ is equal to:

$$I(\theta) = L(\theta) A \cos\theta\,. \tag{2.1.8}$$

Thus, when $L(\boldsymbol{x}, \theta)$ does not depend on the position and the direction (i.e. $L(\boldsymbol{x}, \theta) = L$), we obtain Lambert's cosine law:

$$I(\theta) = I_0 \cos\theta\,. \tag{2.1.9}$$

where $I_0 = I(0) = LA$.

Finally, the étendue U (unit [m sr]) describes the ability of a source to emitt light or the capability of an optical system to receive light, [5]. The quantity $\mathrm{d}U$ is defined as:

$$\mathrm{d}U = n^2 \cos\theta \mathrm{d}A \mathrm{d}\Omega. \tag{2.1.10}$$

where n is the index of refraction of the medium in which the surface A is immersed. In optics the étendue is considered to be a volume in phase space (or an area for two-dimensional systems). This concept will be clarified in Chapter 4 in which we treat the phase space in more detail. An important property of the étendue is that it is conserved within an optical system in absence of absorption. We now show, using the approach of Chaves in [2], how conservation of this quantity can be derived. Consider a light ray emitted from an infinitesimal area $\mathrm{d}A_1$ to the area $\mathrm{d}A_2$. Suppose that the centers of $\mathrm{d}A_1$ and $\mathrm{d}A_2$ are located at a distance d to each other, see Figure 2.3. Indicating with $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ the normals to the surfaces $\mathrm{d}A_1$ and $\mathrm{d}A_2$, respectively and with θ_1 and θ_2 the angles that the central ray forms with $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, respectively, the flux $\mathrm{d}\Phi_1$ passing through $\mathrm{d}A_2$ coming from $\mathrm{d}A_1$ and the corresponding solid angle

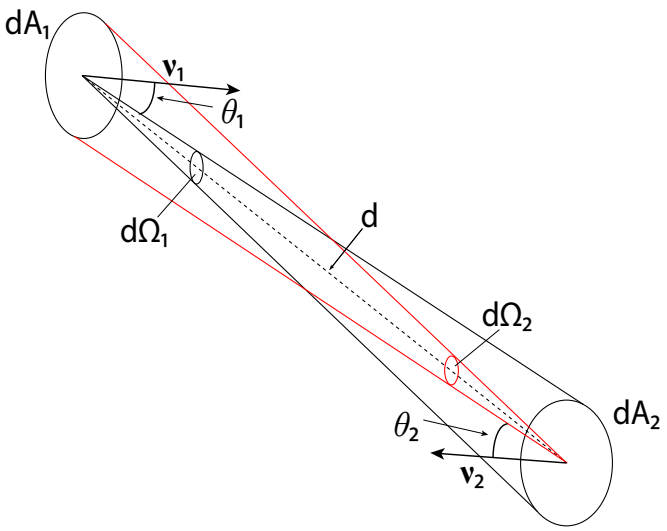


Figure 2.3: $\mathrm{d}A_1$ and $\mathrm{d}A_2$ are two surfaces with normals ν_1 and ν_2, respectively. Their centers are located at a distance d. θ_1 and θ_2 are the angles made by the central ray with the normals ν_1 and ν_2, respectively.

$\mathrm{d}\Omega_1$ are defined as:

$$\begin{aligned} \mathrm{d}\Phi_1 &= L\cos\theta_1 \mathrm{d}A_1 \mathrm{d}\Omega_1, \\ \mathrm{d}\Omega_1 &= \frac{\mathrm{d}A_2\cos(\theta_2)}{d^2}\,. \end{aligned} \tag{2.1.11}$$

Similarly, the flux $\mathrm{d}\Phi_2$ passing through $\mathrm{d}A_1$ coming from $\mathrm{d}A_2$ is equal to:

$$\begin{aligned} \mathrm{d}\Phi_2 &= L\cos\theta_2 \mathrm{d}A_2 \mathrm{d}\Omega_2 \\ \mathrm{d}\Omega_2 &= \frac{\mathrm{d}A_1\cos\theta_1}{d^2}\,. \end{aligned} \tag{2.1.12}$$

Then from Eq. (2.1.10) we obtain the following relations:

$$\begin{aligned} \mathrm{d}U_1 &= n^2 \mathrm{d}A_1\cos\theta_1 \mathrm{d}\Omega_1 = \frac{n^2 \mathrm{d}A_1\cos\theta_1 \mathrm{d}A_2\cos\theta_2}{d^2}, \\ \mathrm{d}U_2 &= n^2 \mathrm{d}A_2\cos\theta_2 \mathrm{d}\Omega_2 = \frac{n^2 \mathrm{d}A_2\cos\theta_2 \mathrm{d}A_1\cos\theta_1`}{d^2} \end{aligned} \tag{2.1.13}$$

for $\mathrm{d}A_1$ and $\mathrm{d}A_2$, respectively. From the previous equations we can conclude that $\mathrm{d}U_1 = \mathrm{d}U_2$ and therefore the étendue $\mathrm{d}U$ is conserved along a beam of light. Since also the flux through the areas $\mathrm{d}A_1$ and $\mathrm{d}A_2$ is conserved, the following relation holds:

$$L := n^2 \frac{\mathrm{d}\Phi}{\mathrm{d}U} = constant\,. \tag{2.1.14}$$

In the optical systems we will consider in this work, the source and the target are located in the same medium (air) with $n = 1$, so the luminance L equals the basic

luminance $L^* = L/n^2$ at the source and the target of the system.

In this thesis we consider two-dimensional optical systems. Hence, the definitions of the photometric parameters have to be given in two dimensions. An infinitesimal line segment of length $\mathrm{d}a$ that emits a light beam and the ray that makes an angle θ with the normal $\boldsymbol{\nu}$ are considered, see Fig. 2.4.

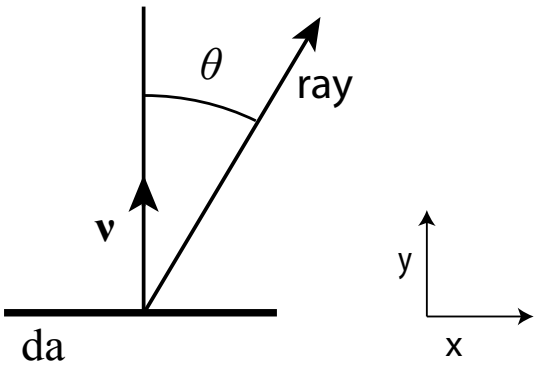


Figure 2.4: Ray emitted by an infinitesimal line segment $\mathrm{d}a$ that makes an angle θ with respect to the line normal $\boldsymbol{\nu}$.

The two-dimensional illuminance $\big(\text{unit } \big[\mathrm{lm/m}\big]\big)$ denotes the luminous flux falling on an infinitesimal line segment of length $\mathrm{d}a$ and it is given by:

$$E = \frac{\mathrm{d}\Phi}{\mathrm{d}a} \,. \tag{2.1.15}$$

The luminous intensity $\big(\text{unit } \big[\mathrm{lm/rad}\big]\big)$ is the luminous flux per angle $\mathrm{d}\theta$:

$$I = \frac{\mathrm{d}\Phi}{\mathrm{d}\theta} \,. \tag{2.1.16}$$

The two-dimensional luminance $\big(\text{unit } \big[\mathrm{lm/(rad \cdot m)}\big]\big)$ is given by:

$$L = \frac{\mathrm{d}\Phi}{\cos\theta \, \mathrm{d}a \, \mathrm{d}\theta}. \tag{2.1.17}$$

Thus the following relation holds:

$$I = L(x, \theta) \cos\theta \, \mathrm{d}a \tag{2.1.18}$$

where x is a certain position at the light source $\mathrm{d}a$. Finally, the étendue $\mathrm{d}U$ (unit $[\mathrm{m} \cdot \mathrm{rad}]$) in two dimensions is given by:

$$\mathrm{d}U = n \cos\theta \mathrm{d}a \, \mathrm{d}\theta. \tag{2.1.19}$$

In order to determine the light distribution on a surface and to compute the photometric variables on that surface, we need to understand how the light emitted from the source propagates. In the field of geometric optics the light propagation is described by light rays. The propagation of a light ray traveling through different media is determined by the reflection and refraction law. In the following we introduce these two laws and we explain the total internal reflection phenomenon.

2.2 Reflection and refraction law

A light ray is described by a position vector $\boldsymbol{x}$ on a surface and a direction vector $\boldsymbol{t}$ and can be parameterized by the arc length s. Light rays travel in a homogeneous medium along straight lines, once they hit a reflective surface their direction changes. Denoting with $\boldsymbol{t}_\mathrm{i}$ the direction of the incident ray and with $\boldsymbol{\nu}$ the unit normal to the surface at the location of incidence, the direction $\boldsymbol{t}_\mathrm{r}$ of the reflected ray is given by:

$$\boldsymbol{t}_\mathrm{r} = \boldsymbol{t}_\mathrm{i} - 2(\boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{\nu})\boldsymbol{\nu}\,, \tag{2.2.1}$$

where the vectors $\boldsymbol{t}_\mathrm{i}$ and $\boldsymbol{\nu}$ are unit vectors and $\boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{\nu}$ indicates the scalar product between $\boldsymbol{t}_\mathrm{i}$ and $\boldsymbol{\nu}$. From Eq. (2.2.1) it follows that the vector $\boldsymbol{t}_\mathrm{r}$ is a unit vector too, indeed considering the scalar product $(\boldsymbol{t}_\mathrm{r}, \boldsymbol{t}_\mathrm{r})$ we conclude:

$$\boldsymbol{t}_\mathrm{r} \cdot \boldsymbol{t}_\mathrm{r} = \boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{t}_\mathrm{i} - 4(\boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{\nu})(\boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{\nu}) + 4(\boldsymbol{t}_\mathrm{i} \cdot \boldsymbol{\nu})^2(\boldsymbol{\nu} \cdot \boldsymbol{\nu}) = 1. \tag{2.2.2}$$

The vectors $\boldsymbol{t}_\mathrm{i}$, $\boldsymbol{t}_\mathrm{r}$ and $\boldsymbol{\nu}$ live all in the same plane. Defining the incident angle θ_i and the reflective angle θ_r such that θ_i, $\theta_\mathrm{r} \in [0, \pi/2)$. the reflection law states that $\theta_\mathrm{i} = \theta_\mathrm{r}$, see Fig. 2.5.

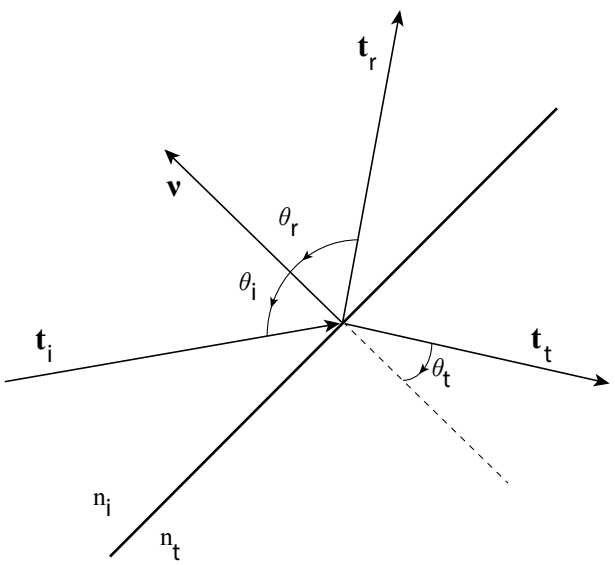


Figure 2.5: Propagation of a ray through two different media with index of refraction n_i and n_t.

When a ray propagates through two different media, its direction changes according to the law of refraction. Indicating with n_i the index of refraction of the medium in which the incident ray travels and with n_t the index of refraction of the medium of the transmitted ray, the direction $\boldsymbol{t}_\mathrm{t}$ of the transmitted ray is given by:

$$\boldsymbol{t}_\mathrm{t} = n_{\mathrm{i,t}}\, \boldsymbol{t}_\mathrm{i} + \left[\sqrt{1 - n_{\mathrm{i,t}}^2 + n_{\mathrm{i,t}}^2(\boldsymbol{\nu} \cdot \boldsymbol{t}_\mathrm{i})^2} - n_{\mathrm{i,t}}(\boldsymbol{\nu} \cdot \boldsymbol{t}_\mathrm{i})\right]\boldsymbol{\nu}\,, \tag{2.2.3}$$

where $n_{\mathrm{i,t}} = n_{\mathrm{i}}/n_{\mathrm{t}}$, [2]. Note that in Eq. (2.2.1) the direction of the normal $\boldsymbol{\nu}$ to the surface is not relevant for the computation of the direction of the reflective ray, since:

$$\boldsymbol{t}_{\mathrm{r}} = \boldsymbol{t}_{\mathrm{i}} - 2(\boldsymbol{t}_{\mathrm{i}} \cdot \boldsymbol{\nu})\boldsymbol{\nu} = \boldsymbol{t}_{\mathrm{i}} - 2(\boldsymbol{t}_{\mathrm{i}} \cdot -\boldsymbol{\nu})(-\boldsymbol{\nu}), \tag{2.2.4}$$

however, this is not the case for Eq. (2.2.3), therefore in the latter case we need to specify the direction of $\boldsymbol{\nu}$ which is usually chosen in such a way that the angle that it forms with the incident ray $\boldsymbol{t}_{\mathrm{i}}$ is smaller than or equal to $\pi/2$. Hence, if $(\boldsymbol{t}_{\mathrm{i}}, \boldsymbol{\nu}) \leq 0$ the normal $\boldsymbol{\nu}$ directed inside the same medium in which travels the incident ray is taken as in Fig. 2.5, otherwise the normal $-\boldsymbol{\nu}$ directed inside the same medium in which the transmitted ray will travel has to be considered.

Eq. (2.2.3) is only valid for

$$1 - n_{\mathrm{i,t}}^2 + n_{\mathrm{i,t}}^2(\boldsymbol{\nu} \cdot \boldsymbol{t}_{\mathrm{i}})^2 \geq 0 \tag{2.2.5}$$

which implies that

$$\frac{n_{\mathrm{t}}}{n_{\mathrm{i}}} \geq \sqrt{1 - (\boldsymbol{\nu} \cdot \boldsymbol{t}_{\mathrm{i}})^2} \tag{2.2.6}$$

from which we obtain:

$$n_{\mathrm{t}} \geq n_{\mathrm{i}} \sin\theta_{\mathrm{i}}\,. \tag{2.2.7}$$

The angle θ_{c} for which the equality holds is

$$\theta_{\mathrm{c}} = \arcsin\left(\frac{n_{\mathrm{t}}}{n_{\mathrm{i}}}\right) \tag{2.2.8}$$

and it is called the critical angle, [2]. When the incident angle θ_{i} is exactly equal to the critical angle θ_{c}, the square root in Eq. (2.2.3) is zero and the inner product $(\boldsymbol{t}_{\mathrm{t}}, \boldsymbol{\nu}) = 0$, hence the transmitted ray propagates parallel to the refractive surface. When $\theta_{\mathrm{i}} > \theta_{\mathrm{c}}$ the light ray is no longer refracted but is only reflected by the surface. This phenomenon is called total internal reflection (TIR). When TIR occurs, 100% of light is reflected and there is no loss of energy. Therefore, optical systems designed such that rays are reflected by TIR are very efficient. Lght that hits an ordinary refractive surface can be reflected and refracted. The energy that is reflected and refracted is determined by the Fresnel's coefficients. In the next paragraph an overview of the Fresnel coefficients is given.

2.3 Fresnel's equations

In order to derive Fresnel's equations we need to describe light as an electromagnetic wave. It is therefore useful to study the light propagation from the perspective of electromagnetic theory which gives information about the incident, reflected and transmitted radiant flux density that are denoted with E_i, E_r and E_t, respectively. Any component of the electric field $\boldsymbol{\mathcal{E}}$ can be written as

$$\boldsymbol{\mathcal{E}}(\boldsymbol{x}, \tau) = \boldsymbol{\mathcal{E}}_0(\boldsymbol{x}) e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega\tau)} \tag{2.3.1}$$

where $\boldsymbol{x}$ is the position vector and T is the time. The amplitude $\boldsymbol{\mathcal{E}}_0(\boldsymbol{x})$ is constant in time and $\omega = \frac{ck}{n}$ is the value of the angular frequency with c the velocity of light

and n the index of refraction in which the wave is traveling, which is the ratio of the speed of light c in vacuum and the speed of light v in the material. Note that the angular frequency can be also written as $\omega = vk$, in particular when a wave travels in vacuum $n = 1$ and $\omega = ck$. The vector $\boldsymbol{k}$ has the same direction of the wave and its absolute value $|\boldsymbol{k}| = k = \frac{2\pi}{\lambda}$ is the wave number in vacuum, with λ the wavelength. Similarly, the magnetic field has the form:

$$\boldsymbol{\mathcal{B}}(\boldsymbol{x}, \tau) = \boldsymbol{\mathcal{B}}_0(\boldsymbol{x}) e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega\tau)} \,. \tag{2.3.2}$$

Light can be seen as an electromagnetic wave, that is an oscillating electric field $\boldsymbol{\mathcal{E}}$ and an oscillating magnetic filed $\boldsymbol{\mathcal{B}}$ which propagates always perpendicular to $\boldsymbol{\mathcal{E}}$. Th eelectric field oscillates perpendicular to the wave propagation. Light is said to be polarized if the direction of the electic field is well defined. When the electric field propagates in different directions we talk about unpolarized light. By convention, we refer to the light's polarization as the direction of the electric field $\boldsymbol{\mathcal{E}}$, [6] with respect to the incident plane that is defined by the incident and reflected rays as is shown in Fig. 2.6.

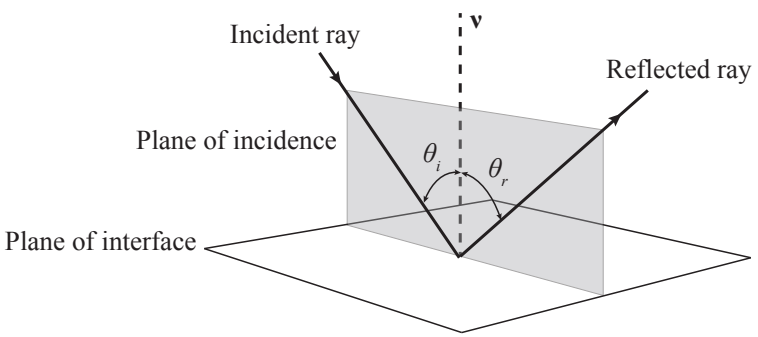


Figure 2.6: Light ray that hits a mirror located on the reflecting plane. The incident and the reflected ray leave in the same plane of the normal to the mirror that is called plane of incident.

In order to derive the Fresnel's coefficients the polarization of light must to be taken into account. Those coefficients are obtained considering Maxwell's equations and the boundary conditions due to the conservation of energy. The details of Fresnel's equations are widely explained in the literature. In the following we provide Fresnel coefficients and we briefly explain their physical interpretation. We refer the reader to [7, 8] for more details. Fresnel's coefficients can also be derived using a different approach that does not involve Maxwell's equations, this method is explained in [9]. The following particular cases of light's polarization need are considered.

1. $\boldsymbol{\mathcal{E}}$ is perpendicular to the plane of incidence (see Fig. 2.7). In this case light is said to be s-polarized.

2. $\boldsymbol{\mathcal{E}}$ is parallel to the plane of incidence (see Fig. 2.8). In this case light is said to be p-polarized.

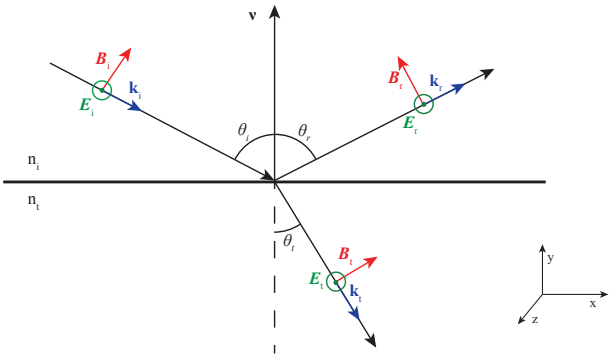


Figure 2.7: Propagation of an electromagnetic wave where $\boldsymbol{\mathcal{E}}$ is perpendicular to the incident plane. The components of $\boldsymbol{\mathcal{E}}$ are is indicated with the green circles. The components of $\boldsymbol{\mathcal{B}}$ are indicated with red arrows.

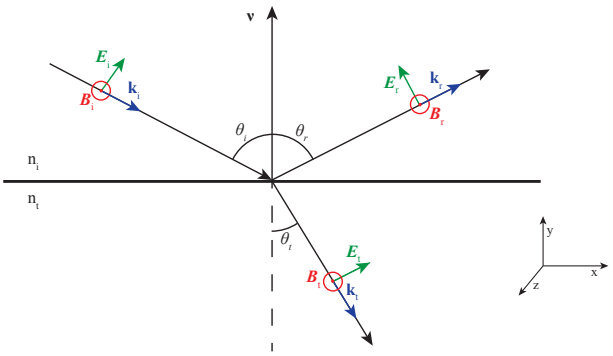


Figure 2.8: Propagation of an electromagnetic wave where $\boldsymbol{\mathcal{E}}$ is parallel to the incident plane. The components of $\boldsymbol{\mathcal{B}}$ are indicated with the red circle. The components of $\boldsymbol{\mathcal{E}}$ are indicated with greed arrows.

Energy conservation gives the boundary conditions of the electromagnetic field at the plane of the interface (which is perpendicular to the incident plane). In the following we derive Fresnel's coefficients for case 1. Similarly, the Fresnel's coefficients can be derived for the second case.

For s-polarized light the tangential components of $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}/\mu$ across the boundary between the two different media must be continuous. The continuity of the tangential component of $\boldsymbol{\mathcal{E}}$ leads to:

$$|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}| + |\boldsymbol{\mathcal{E}}_{0\mathrm{r}}| = |\boldsymbol{\mathcal{E}}_{0\mathrm{t}}|, \tag{2.3.3}$$

while the continuity of the tangential component of $\boldsymbol{\mathcal{B}}/\mu$ gives:

$$-\frac{|\boldsymbol{\mathcal{B}}_{0,\mathrm{i}}|}{\mu_\mathrm{i}}\cos\theta_\mathrm{i} + \frac{|\boldsymbol{\mathcal{B}}_{0,\mathrm{r}}|}{\mu_\mathrm{r}}\cos\theta_\mathrm{r} = -\frac{|\boldsymbol{\mathcal{B}}_{0,\mathrm{t}}|}{\mu_\mathrm{t}}\cos\theta_\mathrm{t}, \tag{2.3.4}$$

where the negative sign in front of $|\boldsymbol{\mathcal{B}}_{0,\mathrm{i}}|$ and $|\boldsymbol{\mathcal{B}}_{0,\mathrm{t}}|$ is due to the convention that a positive direction is considered with increasing x. Since $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{E}}/v$, Eq. (2.3.4) can be written as

$$\frac{1}{\mu_\mathrm{i} v_\mathrm{i}}(|\boldsymbol{\mathcal{E}}_{0,\mathrm{i}}| - |\boldsymbol{\mathcal{E}}_{0,\mathrm{r}}|)\cos\theta_\mathrm{i} = \frac{1}{\mu_\mathrm{t} v_\mathrm{t}}|\boldsymbol{\mathcal{E}}_{0,\mathrm{t}}|\cos\theta_\mathrm{t}, \tag{2.3.5}$$

where we employed the fact that $v_\mathrm{i} = v_\mathrm{r}$, and $\theta_\mathrm{i} = \theta_\mathrm{r}$. Using Eq. (2.3.1) and $n = c/v$, the previous equation becomes:

$$\frac{n_\mathrm{i}}{\mu_\mathrm{i}}(|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}| - |\boldsymbol{\mathcal{E}}_{0\mathrm{r}}|)\cos\theta_\mathrm{i} = \frac{n_\mathrm{t}}{\mu_\mathrm{i}}|\boldsymbol{\mathcal{E}}_{0\mathrm{t}}|\cos\theta_\mathrm{t} \tag{2.3.6}$$

Finally, assuming that $\mu_\mathrm{i} = \mu_\mathrm{t} = \mu_0$ and employing Eq. (2.3.3) we obtain:

$$\begin{aligned} r_s &= \frac{|\boldsymbol{\mathcal{E}}_{0\mathrm{r}}|_s}{|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}|_s} = \frac{n_\mathrm{i}\cos\theta_\mathrm{i} - n_\mathrm{t}\cos\theta_\mathrm{t}}{n_\mathrm{i}\cos\theta_\mathrm{i} + n_\mathrm{t}\cos\theta_\mathrm{t}}, \\ t_s &= \frac{|\boldsymbol{\mathcal{E}}_{0\mathrm{t}}|_s}{|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}|_s} = \frac{2n_\mathrm{i}\cos\theta_\mathrm{i}}{n_\mathrm{i}\cos\theta_\mathrm{i} + n_\mathrm{t}\cos\theta_\mathrm{t}}. \end{aligned} \tag{2.3.7}$$

The coefficients r_s and t_s are amplitude coefficients for the reflected and transmitted light. They are the perpendicular components of r and t for s-polarized light. Using Snell's law, that is $n_\mathrm{i}\sin\theta_\mathrm{i} = n_\mathrm{t}\sin\theta_\mathrm{t}$, the relations for r_s and t_s are simplified as follows:

$$\begin{aligned} r_s &= -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}, \\ t_s &= -\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)}. \end{aligned} \tag{2.3.8}$$

A similar argument for the p-polarized light leads to the calculation of the parallel components r_p and t_p of r and t. In case $\boldsymbol{\mathcal{E}}$ is parallel to the plane of incidence the amplitude coefficients are:

$$\begin{aligned} r_p &= \frac{n_t\cos\theta_i - n_i\cos\theta_t}{n_i\cos\theta_t + n_t\cos\theta_i}, \\ t_p &= \frac{2n_i\cos\theta_i}{n_i\cos\theta_t + n_t\cos\theta_i}, \end{aligned} \tag{2.3.9}$$

and their simplified relations are:

$$\begin{aligned} r_p &= \frac{\tan(\theta_i - \theta_t)}{\theta_i + \theta_t}, \\ t_p &= \frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}. \end{aligned} \tag{2.3.10}$$

Furthermore, it can be checked that

$$\begin{aligned} t_s - r_s &= 1, \\ t_p + r_p &= 1. \end{aligned} \tag{2.3.11}$$

The amplitude coefficients are shown in Fig. 2.9 for the case in which light travels from a less dense to a more dense medium ($n_i < n_t$), that is external reflection. In

Fig. 2.10 the reflection coefficients are shown for the case in which $n_i > n_t$, that is internal reflection. Note from Fig. 2.9 that r_p approaches to 0 when θ_i approaches to θ_p and it gradually decreases reaching -1 for an incident angle $\theta_i = 90°$. The angle θ_p is called Brewster's angle or polarization angle as only the component perpendicular to the incident plane is reflected at that angle and therefore light is perfectly polarized. Similarly, Fig. 2.10 shows that $r_p = 0$ for $\theta_i = \theta_{p\prime}$. It can be show that $\theta_p + \theta_{p\prime} = 90°$. Both r_p and r_s reach 1 when $\theta_i = \theta_c$. θ_c is called the critical angle. Light that hits the incident plane with an incident angle equal to or greater than the critical angle is totally reflected back and no transmitted light is observed. This phenomenon is called total internal reflection.

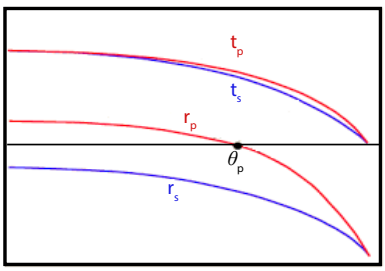


Figure 2.9: Amplitude coefficients of reflection and transmission as a function of the incident angle θ_i in the case of external reflection, i.e. $n_t < n_i$ ($n_t = 1$ and $n_i = 1.5$). θ_p is the polarization angle, [8].

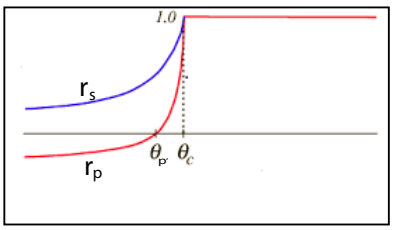


Figure 2.10: Reflection coefficients as a function of the incident angle θ_i in the case of internal reflection, i.e. $n_t > n_i$ ($n_t = 1.5$ and $n_i = 1$). $\theta_{p\prime}$ is the polarization angle and θ_c is the critical angle, [8].

The we introduce the Poynting vector $\boldsymbol{P}$ that defines the energy flux of an electromagnetic field. It is measured in [W/m^2], and it is given by:

$$\boldsymbol{P} = \frac{1}{\mu}\Big(\boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{B}}\Big), \tag{2.3.12}$$

where $\mu = \frac{1}{\varepsilon v^2}$ is the permeability and ε the permittivity of the medium. In the following, the parameters for vacuum are indicated with the subscript 0. All quantities defined in the media of the incident, reflective and transmitted light are indicated with the subscripts i, r and t, respectively. Optical rays are perpendicular to the wave front of an electromagnetic wave and parallel to the Poynting vector, [10]. The irradiance E is defined as the average energy that crosses in unit time a unit area A perpedicular to the direction of the energy flow. Therefore, defining the average of the vector $\boldsymbol{P}$ over the time as:

$$\langle \boldsymbol{P} \rangle_T = \frac{1}{T}\int_0^T \boldsymbol{P} \mathrm{d}T \tag{2.3.13}$$

we can write the irradiance E as:

$$\boldsymbol{E} = \langle \boldsymbol{P} \rangle_\tau = v\varepsilon |\boldsymbol{\mathcal{E}}|^2 . \tag{2.3.14}$$

Considering a beam of light that hits a surface such that an area A is illuminated, the incident, reflected and transmitted beams are $\boldsymbol{E}_{\mathrm{i}} A\cos\theta_{\mathrm{i}}$ $\boldsymbol{E}_{\mathrm{r}} A\cos\theta_{\mathrm{r}}$ and $\boldsymbol{E}_{\mathrm{t}} A\cos\theta_{\mathrm{t}}$,

respectively. The reflectance $\mathcal{R}$ is the ratio of the reflected power to the incident power:

$$\mathcal{R} = \frac{|\boldsymbol{E}_\mathrm{r}| \cos\theta_r}{|\boldsymbol{E}_\mathrm{i}| \cos\theta_\mathrm{i}} = \frac{|\boldsymbol{\mathcal{E}}_{0\mathrm{r}}|^2}{|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}|^2} = r^2 \tag{2.3.15}$$

where the second equality holds because $v_\mathrm{i} = v_\mathrm{t}$, $\varepsilon_\mathrm{i} = \varepsilon_\mathrm{t}$ and $\theta_\mathrm{i} = \theta_\mathrm{t}$. Similarly, the transmittance $\mathcal{T}$ is the ratio between the transmitted to the incident power:

$$\mathcal{T} = \frac{|\boldsymbol{E}_\mathrm{t}| \cos\theta_\mathrm{t}}{|\boldsymbol{E}_\mathrm{r}| \cos\theta_\mathrm{r}} = \frac{n_\mathrm{t} \cos\theta_\mathrm{t}}{n_\mathrm{t} \cos\theta_\mathrm{i}} \frac{|\boldsymbol{\mathcal{E}}_{0\mathrm{t}}|^2}{|\boldsymbol{\mathcal{E}}_{0\mathrm{i}}|^2} = \frac{n_\mathrm{t} \cos\theta_\mathrm{t}}{n_\mathrm{t} \cos\theta_\mathrm{i}} t^2 \,. \tag{2.3.16}$$

Employing total energy conservation, that is:

$$\boldsymbol{E}_\mathrm{i} A \cos\theta_\mathrm{i} = \boldsymbol{E}_\mathrm{r} A \cos\theta_\mathrm{r} + \boldsymbol{E}_\mathrm{t} A \cos\theta_\mathrm{t}, \tag{2.3.17}$$

we can easily prove that:

$$\mathcal{R} + \mathcal{T} = 1 \,. \tag{2.3.18}$$

The parallel and perpendicular components of $\mathcal{R}$ and $\mathcal{T}$ are:

$$\begin{aligned} \mathcal{R}_p &= r_p^2, \\ \mathcal{T}_p &= \frac{n_t \cos\theta_t}{n_t \cos\theta_i} t_p^2, \\ \mathcal{R}_s &= r_s^2, \\ \mathcal{T}_s &= \frac{n_t \cos\theta_t}{n_t \cos\theta_i} t_s^2. \end{aligned} \tag{2.3.19}$$

it can be show that

$$\begin{aligned} \mathcal{R}_s + \mathcal{R}_p &= 1, \\ \mathcal{T}_s + \mathcal{T}_p &= 1 \,. \end{aligned} \tag{2.3.20}$$

For normal incidence, i.e. $\theta_i = 0$, there is no polarization and Eqs. (2.3.19) lead to:

$$\begin{aligned} \mathcal{R} = \mathcal{R}_p = \mathcal{R}_s &= \left(\frac{n_i - n_t}{n_t + n_i} \right)^2, \\ \mathcal{T} = \mathcal{T}_p = \mathcal{T}_s &= \frac{4 n_i n_t}{(n_t + n_i)^2} \,. \end{aligned} \tag{2.3.21}$$

Many common light sources such as sunlight, halogen lighting, LED spotlights, and incandescent bulbs produce unpolarized light. In case of unpolarized light the amount of reflected and transmitted light is given by the average of reflectance $\mathcal{R}$ and transmittance $\mathcal{T}$ calculated considering first p-polarized light and then s-polarization, that is:

$$\begin{aligned} \mathcal{R} &= \frac{\mathcal{R}_p + \mathcal{R}_s}{2}, \\ \mathcal{T} &= \frac{\mathcal{T}_p + \mathcal{T}_s}{2}, \end{aligned} \tag{2.3.22}$$

where $\mathcal{R}_p$, $\mathcal{R}_s$, $\mathcal{T}_p$ and $\mathcal{T}_s$ are given in Eqs. (2.3.19).

With this overview we conclude this chapter. The notions given in Section 2.1 will be used in the entire thesis as our goal is to study the distribution of light at the target of some optical systems. In particular we will focus on the computation of the output intensity distribution. The reflection and refraction laws explained in Section 2.2 are needed to determine how the optical system changes the ray's direction every time that it hits a surfaces (or a line in the two-dimensional case). In Chapters 3, 4, **??**, 7 and 8 only systems where the reflection and refraction laws play a role are considered. Systems with Fresnel reflection are treated in the last chapter. The amount of reflected and transmitted light is calculated using the Fresnel's equation (introduced in the last paragraph of this chapter). Since, we restrict ourselves to two-dimensional systems, the value of reflectance and transmittance will be computed using Eqs. (2.3.22).

Chapter 3

Ray tracing

Optical ray tracing is a tool to calculate the transport of light within optical systems. Given an optical system and a set of rays at the source, ray tracing relates the emitted light with its output distribution. The influence of diffraction on the transport of a ray is neglected.

Although the method can be implemented for two or more dimensions and for any optical system, here we consider the two-dimensional case only. From now on, we will thus refer to optical lines instead of optical surfaces. The two-dimensional case has limitations. For example, it may not identify skew rays that are turned back by the system, with the consequence that a 2D analysis cannot guarantee a proper treatment of non meridional rays in 3D. Nevertheless, the two-dimensional case is particularly relevant because it is a good test case to demonstrate the performance of new methods. Optical designers often start with 2D systems, where only the meridional plane is taken into account because it gives a good prediction of the target distribution of the rays (see [11], chapter 4, p.$50-65$).

3.1 Ray tracing for two-dimensional optical systems

Light rays are straight lines and they are reflected or refracted by the optical components. Every ray emitted from the source is followed until it reaches the target. The ray tracing procedure is constructed such that the position and the direction of the rays are calculated on every optical line that they hit.

Given a Cartesian coordinate system (x, z), a two-dimensional optical system symmetric with respect to the z-axis is defined. Hence, usually the optical axis coincides with the z-axis. The optical system is formed by a source S, a target T and some optical components labeled with indexes j where $j \in \{2, \cdots, \mathrm{Nl}-1\}$ and Nl indicates the number of lines that form the system. S and T are indicated with the indexes 1 and Nl, respectively. The index of refraction of the medium in which line j is located is indicated with n_j. Every ray emitted by S (line 1) can hit some optical components $j \in \{2, \cdots, \mathrm{Nl}-1\}$ before reaching T (line Nl). The intersection point of the rays with line j are $(x_j, z_j)_{j=1,\cdots,\mathrm{Nl}}$ and, $\boldsymbol{s}_j = (-\sin t_j, \cos t_j)$ indicates the direction vector of the rays that leave j, with t_j the angle that the ray forms with respect to the optical axis measured counterclockwise. As we consider only forward rays, the angles

$t_j \in (-\pi/2, \pi/2)$. Therefore, a ray segment between (x_j, z_j) and (x_k, z_k) with $k \neq j$ is parameterized in real space by:

$$\boldsymbol{r}(s) = \begin{pmatrix} x_j - s\sin t_j \\ z_j + s\cos t_j \end{pmatrix} \qquad 0 < s \leq s_{\max}\,, \tag{3.1.1}$$

where s denotes the arc-length and $s_{\max}$ is the maximum value that it can assume. Fig. 3.1 shows an example where a single ray is traced inside a very simple optical system, the so-called two-faceted cup. The light source $\mathsf{S}= [-a, a]$ (line 1) and the

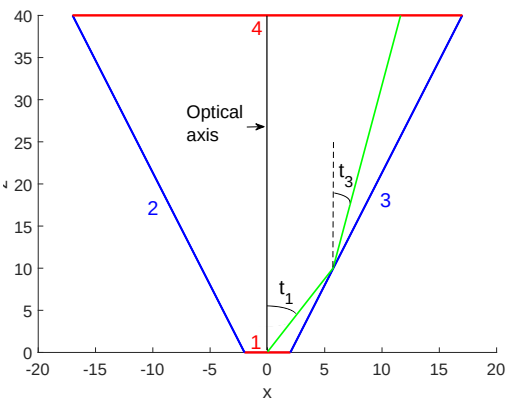


Figure 3.1: Shape of the two-faceted cup. Each line of the system is labeled with a number. The source $\mathsf{S}= [-2, 2]$ (line number 1) is located on the x-axis. The target $\mathsf{T}= [-17, 17]$ (line 4) is parallel to the source and is located at a height $z = 40$. The left and right reflectors (line 2 and 3) connect the source with the target.

target $\mathsf{T} = [-b, b]$ (line 4) are two segments normal to the z-axis, where $a = 2$ and $b = 17$. The left and right reflectors (line 2 and 3) are oblique segments that connect the source and the target. All the optical lines $j \in \{1, \cdots, 4\}$ are located in air, thus the refractive index is $n_j = 1$ for every j.

In order to compute the target photometric variables, we need to know how the optical system influences the direction of the rays when they hit an optical line. Ray tracing relates the position coordinates (x_1, z_1) and the direction vector $\boldsymbol{s}_1$ of every ray at the source S with the corresponding position $(x_{\mathrm{Nl}}, z_{\mathrm{Nl}})$ and direction $\boldsymbol{s}_{\mathrm{Nl}}$ at the target T. In the following we will often use the target coordinates of the rays thus, to simplify the notation, we do not write the subscript Nl for the target coordinates. Hence, we write (x, z) instead of $(x_{\mathrm{Nl}}, z_{\mathrm{Nl}})$, t instead of t_{Nl} and $\boldsymbol{s}$ instead of $\boldsymbol{s}_{\mathrm{Nl}}$ for the target coordinates. The ray tracing algorithm can be outlined as follows:

1. Given a ray that leaves S with initial position (x_1, z_1) and initial direction $\boldsymbol{s}_1 = (-\sin t_1, \cos t_1)$, use Eq. (3.1.1) to implement the ray parametrization $\boldsymbol{r}(s_1)$;

2. Compute the coordinates $(x_k, z_k)_{k=1\cdots,\mathrm{Nl}}$ of the intersection point of the parameterized ray $\boldsymbol{r}(s)$ with all the lines that it crosses

   a) if the shape of the lines is described by an analytical equation, the intersection points are determined analytically;

   b) if there is no analytic description for the optical lines, the intersections need to be determined using iterative methods;

3. Determine the closest line j that the forward ray encounters;
4. If $j = \mathrm{Nl}$ stop the procedure, the target ray's coordinates (x, z) and $\boldsymbol{s}$ are found.
5. Calculate the normal $\boldsymbol{\nu}_i$ to line i at the point (x_i, z_i);
6. Compute the new ray direction $\boldsymbol{s}_j$ of the ray that leaves line j at the point (x_i, z_i):
   a) if the incident line is a reflective line, $\boldsymbol{s}_j$ is given by Eq. (2.2.1);
   b) if the incident line is a refractive line, $\boldsymbol{s}_j$ is given by Eq. (2.2.3);
7. Restart the procedure from 1. for the ray that leaves line j instead of S. Consider as initial ray position coordinates (x_i, z_i) instead of (x_1, z_1) and as initial ray direction $\boldsymbol{s}_j = (-\sin t_j, \cos t_j)$ instead of $\boldsymbol{s}_1$.

The procedure explained above is repeated for every ray traced through the system, [12]. Once the target position and the direction of every ray traced are computed, the target photometric variables can be calculated using the definitions explained in the previous chapter, see section 2.1.

There are different ways to implement the ray tracing procedure. The efficiency of the ray tracing can be related with the distribution of the rays at the source. If the initial position and direction of the rays are chosen randomly we have Monte Carlo (MC) ray tracing. This is a very common method in non-imaging optics as it is very powerful and easy to implement. MC ray tracing will be explained in details in the next paragraph. If the rays are chosen from a so-called low discrepancy sequence we have the Quasi-Monte Carlo (QMC) ray tracing. This approach is discussed in Section 3.3.

3.2 Monte Carlo ray tracing

Before explain MC ray tracing we give a general introduction to the MC methods approximate computation of integrals. Given an interval $D = [\boldsymbol{a}, \boldsymbol{b}]$ with $\boldsymbol{a} = (a_1, \cdots, a_d)$ and $\boldsymbol{b} = (b_1, \cdots, b_d)$ elements of $\mathbb{R}^d$ such that $[\boldsymbol{a}, \boldsymbol{b}] = [a_1, b_1] \times \cdots \times [a_d, b_d]$, a function $f : [\boldsymbol{a}, \boldsymbol{b}] \subset \mathbb{R}^d \mapsto \mathbb{R}$ and a random variable $\boldsymbol{y} \in D$ with probability density function $\rho(\boldsymbol{y})$, the expected value of f with respect of ρ is

$$\mathbb{E}[f] = \int_D f(\boldsymbol{y})\rho(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{3.2.1}$$

If ρ is a uniform probability density function,

$$\mathbb{E}[f] = \int_D f(\boldsymbol{y})\rho(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \frac{1}{(\boldsymbol{b} - \boldsymbol{a})}\int_D f(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{3.2.2}$$

Monte Carlo approximates Eq. (3.2.2) by

$$S_{\mathrm{N}}(f) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i) \tag{3.2.3}$$

$\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}} \in D$ are indipendent samples of the density function ρ, [13]. According to the strong law of large numbers,

$$\Pr\Big(\lim_{\mathrm{N}\to\infty} \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i) = \mathbb{E}[f(\boldsymbol{y})]\Big) = 1. \tag{3.2.4}$$

Therefore,

$$\mathbb{E}[f] = \int_D f(\boldsymbol{y})\rho(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \approx \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i). \tag{3.2.5}$$

From the linearity of the expected values, it follows the obvious relation

$$\mathbb{E}[S_{\mathrm{N}}(f)] = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathbb{E}[f] = \mathbb{E}[f], \tag{3.2.6}$$

while the Bienaymé formula leads to

$$\mathrm{Var}(S_{\mathrm{N}}) = \mathrm{Var}\left(\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i)\right) = \frac{1}{\mathrm{N}^2} \sum_{i=1}^{\mathrm{N}} \mathrm{Var}\big(f(\boldsymbol{y}_i)\big) \tag{3.2.7}$$

which can be applied because the variables $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}}$ are independent, [14], Chap. 6. Suppose that f has variance $\mathrm{Var}[f] = \mathbb{E}[(f - \mathbb{E}(f))]^2 = \sigma^2[f]$, Eqs. (3.2.6) and (3.2.7) give

$$\mathrm{Var}[S_{\mathrm{N}}(f)] = \mathbb{E}[(S_{\mathrm{N}}(f) - \mathbb{E}[S_{\mathrm{N}}(f)])^2] = \mathbb{E}[(S_{\mathrm{N}}(f) - \mathbb{E}[f])^2] = \sigma[f]^2/\mathrm{N}. \tag{3.2.8}$$

Let us denote the integration error with:

$$\mathrm{err}(f, S_{\mathrm{N}}) = \int_D f(\boldsymbol{y})\rho(\boldsymbol{y})\mathrm{d}\boldsymbol{y} - S_{\mathrm{N}}(f) = \mathbb{E}[f] - S_{\mathrm{N}}(f), \tag{3.2.9}$$

then

$$\mathbb{E}\big[|\mathrm{err}(f, S_{\mathrm{N}})|\big] \leq \sqrt{\mathbb{E}\big[\mathrm{err}(f, S_{\mathrm{N}})^2\big]} = \frac{\sigma[f]}{\sqrt{\mathrm{N}}}, \tag{3.2.10}$$

where the inequality is true because

$$\begin{aligned} \mathbb{E}\big[|\mathrm{err}(f, S_{\mathrm{N}})|\big] &= \frac{1}{\mathrm{N}}\sqrt{\left(\sum_{1=1}^{\mathrm{N}} |\mathrm{err}(f, S_{\mathrm{N}})|\right)^2} \leq \frac{1}{\mathrm{N}}\sqrt{\mathrm{N}\sum_{1=1}^{\mathrm{N}} \big(\mathrm{err}(f, S_{\mathrm{N}})\big)^2} \\ &= \sqrt{\frac{1}{\mathrm{N}}\sum_{1=1}^{\mathrm{N}} \big(\mathrm{err}(f, S_{\mathrm{N}})\big)^2} = \sqrt{\mathbb{E}\big[\mathrm{err}(f, S_{\mathrm{N}})^2\big]} \end{aligned} \tag{3.2.11}$$

and the equality follows from Eqs (3.2.8) and (3.2.9). Hence, the absolute value of the integration error is, on average, bounded by $\sigma[f]/\sqrt{\mathrm{N}}$, where $\sigma[f]$ is the standard deviation of f, [15]. It is very important to note that $\mathrm{err}(f, S_{\mathrm{N}})$ does not depend on the dimension d of f.

MC technique can be combined with the ray tracing procedure in order to compute

the light distribution at the target of an optical system. In MC ray tracing the position and the direction of every ray at the source are chosen randomly. In the two-dimensional case (d=2), for every ray we need to choose one position coordinate x_1 at the source and one angular coordinate t_1 at the target, while the z_1 coordinate of every ray at the source is always given (for instance, for the two-faceted cup in Fig. 3.1, $z_1 = 0$ for every ray). Therefore, given a set of random variables $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N\} \in [\boldsymbol{a}, \boldsymbol{b}] \subset \mathbb{R}^2$, the initial position coordinate x_1 of the k-th ray corresponds to the first component of the k-th random variable $\boldsymbol{y}_k$ and, the starting angular coordinate t_1 of the k-th ray corresponds to the second component of the k-th random variable $\boldsymbol{y}_k$. Next, rays with those random coordinates at S are traced from S to T and, a probabilistic interpretation of the output photometric variables is provided. In particular, we are interested in the total target intensity I which is computed as a function of the angular coordinate t. The MC intensity is calculated dividing the target into intervals of equal length, the so-called bins. A partitioning $P_1 : -\pi/2 = t_0 < t_1 < \cdots < t_{\mathrm{Nb}} = \pi/2$ of the interval $[-\pi/2, \pi/2]$ is defined where Nb is the number of bins in P_1. We remark that, with a slight abuse of notation, we indicated the angular coordinates of the rays at the target (line Nl) with t_j instead of $t_{\mathrm{Nl},j}$ for every $j \in \{0, \cdots, \mathrm{Nb}\}$.

The normalized approximated intensity $I_{\mathrm{MC}}(t)$ is a piecewise constant function, whose value over the j-th bin is the ratio between the number of rays that fall into that bin $\mathrm{Nr}[t_{j-1}, t_j)$ and the total number of rays traced $\mathrm{Nr}[-\pi/2, \pi/2)$. Hence, I_{MC} is defined by

$$I_{\mathrm{MC}}(t) = \frac{\mathrm{Nr}[t_{j-1}, t_j)}{\mathrm{Nr}[-\pi/2, \pi/2)} \qquad \text{for } t \in [t_{j-1}, t_j). \tag{3.2.12}$$

The output intensity is computed from the value of the intensity $I_{\mathrm{MC}}(t_{j-1/2})$ along the direction $t_{j-1/2} = (t_{j-1} + t_j)/2$ for every bin $[t_{j-1}, t_j)_{j=1,\cdots,\mathrm{Nb}}$. The intensity $I_{\mathrm{MC}}(t_{j-1/2})$ gives an estimate of the probability that a ray reaches the target with an angle in the j-th interval $[t_{j-1}, t_j)$ of the partitioning P_1. This probability $\mathrm{P}_{j,\Delta t}$ is given by

$$\mathrm{P}_{j,\Delta t} = \Pr(t_{j-1} \le t < t_j) = \frac{\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t}{\int_{-\pi/2}^{\pi/2} I(t)\mathrm{d}t}\,, \tag{3.2.13}$$

where $I(t)$ is the output intensity (not normalized). Note that $\sum_{j=1}^{\mathrm{Nb}} \mathrm{P}_{j,\Delta t} = 1$. From the mean value theorem for the function $I(t)$, continuous in $[t_{j-1}, t_j]$, there exists a value $t_k \in [t_{j-1}, t_j]$ for which the integral at the numerator of the previous equation can be written as

$$\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t = \Delta t\; I(t_k). \tag{3.2.14}$$

Hence, $\mathrm{P}_{j,\Delta t}$ is proportional to the size $\Delta t = (t_{\mathrm{Nb}} - t_0)/\mathrm{Nb}$ of the intervals and to $I(t_k)$. Although t_k does depend on the number of bins Nb, $I(t_k)$ is constant as it is the value of the intensity on a given direction, so Eq. (3.2.14) proves that $\mathrm{P}_{j,\Delta t}$ is inversely proportional to the number of bins Nb of the partitioning P_1. Indicating with $\Phi = \int_{-\pi/2}^{\pi/2} I(t)\mathrm{d}t$ the total flux (measured in lumen [lm]), the error between the

intensity $I(t_{j-1/2})$ and the averaged MC intensity $\Phi I_{\mathrm{MC}}(t_{j-1/2})/\Delta t$ is given by

$$\begin{aligned}\left|I(t_{j-1/2}) - \frac{\Phi}{\Delta t} I_{\mathrm{MC}}(t_{j-1/2})\right| \leq \\ \left|I(t_{j-1/2}) - \frac{1}{\Delta t}\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t\right| + \\ \frac{1}{\Delta t}\left|\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t - \Phi\, I_{\mathrm{MC}}(t_{j-1/2})\right|.\end{aligned} \tag{3.2.15}$$

The first term of the right hand side of inequality (3.2.15) gives an estimate of how much the averaged intensity $\frac{1}{\Delta t}\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t$ differs from the exact intensity $I(t_{j-1/2})$. This term is due to the discretization of the target and therefore it depends on the number of bins Nb considered. Substituting $I(t)$ with its Taylor expansion around the point $t_{j-1/2}$ we obtain that this term is proportional to the square of the size of the bins. Therefore,

$$\left|I(t_{j-1/2}) - \frac{1}{\Delta t}\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t\right| = \mathrm{C}_1/\mathrm{Nb}^2 \tag{3.2.16}$$

with $C_1 > 0$ a certain constant.

The second part of the right hand side of inequality (3.2.15) gives an estimate of the MC error and therefore it depends also on the number of rays traced. In order to show how this term decreases as a function of the number of rays traced, we define the random variable $X_j(t)$ as the variable that is equal to 1 if the ray with angular coordinate t is inside the interval $[t_{j-1}, t_j)$ and equal to 0 otherwise:

$$X_j(t) = \begin{cases} 1 & \text{if} \quad t \in [t_{j-1}, t_j), \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.17}$$

The Bernoulli trial X_j follows a binomial distribution $B(1, \mathrm{P}_{j,\Delta t})$. Considering a sample of Nr rays, the variable $Y_j = \sum_{k=1}^{\mathrm{Nr}} X_j(t_k)$ follows a binomial distribution $B(\mathrm{Nr}, \mathrm{P}_{j,\Delta t})$, where t_k is the angle that the k-th ray forms with the optical axis. Then, using the de Moivre-Laplace theorem, we conclude that the variable Y_j is approximated by a normal distribution with mean value $E[Y_j] = \mathrm{NrP}_{j,\Delta t}$ and variance $\sigma^2[Y_j] = \mathrm{NrP}_{j,\Delta t}(1 - \mathrm{P}_{j,\Delta t})$ when a large number of rays is considered, see [16, 17]. Thus, the normalized intensity along the direction $t_{j-1/2}$ is

$$I_{\mathrm{MC}}(t_{j-1/2}) = \sum_{k=1}^{\mathrm{Nr}} X_j(t_k)/\mathrm{Nr}. \tag{3.2.18}$$

The mean value $E[I_{\mathrm{MC}}(t_{j-1/2})] = \mathrm{P}_{j,\Delta t}$ and the variance $\sigma^2[I_{\mathrm{MC}}(t_{j-1/2})] = \mathrm{P}_{j,\Delta t}(1 - \mathrm{P}_{j,\Delta t})/\mathrm{Nr}$. Note that the standard deviation $\sigma_j := \sigma[I_{\mathrm{MC}}(t_{j-1/2})]$ is approximated by

$$\sigma_j = \sqrt{\mathrm{P}_{j,\Delta t}(1 - \mathrm{P}_{j,\Delta t})/\mathrm{Nr}} \approx \frac{\mathrm{C}_2}{\sqrt{\mathrm{NbNr}}}, \tag{3.2.19}$$

for some $\mathrm{C}_2 > 0$. σ_j can be used to give an estimate of the difference between the intensity $I_{\mathrm{MC}}(t_{j-1/2})$ and its mean value $\mathrm{P}_{j,\Delta t}$. Therefore, the second term of the

right hand side of relation (3.2.15) becomes

$$\begin{aligned}\frac{1}{\Delta t}\Big|\int_{t_{j-1}}^{t_j} I(t)\mathrm{d}t - \Phi\, I_{\mathrm{MC}}(t_{j-1/2})\Big| = \\ \frac{\Phi}{\Delta t}\Big|\mathrm{P}_{j,\Delta t} - I_{\mathrm{MC}}(t_{j-1/2})\Big| \approx \\ \frac{\Phi}{\Delta t}\sigma_j[I_{\mathrm{MC}}(t_{j-1/2})] \approx \mathrm{C}_3\frac{\mathrm{Nb}}{\sqrt{\mathrm{NbNr}}} = \mathrm{C}_3\sqrt{\frac{\mathrm{Nb}}{\mathrm{Nr}}}\,,\end{aligned} \tag{3.2.20}$$

for some $\mathrm{C}_3 > 0$, where the approximation holds because σ_j gives a measure for the error between $I_{\mathrm{MC}}(t_{j-1/2})$ and the probability $\mathrm{P}_{j,\Delta t}$, [18]. The second approximation follows from (3.2.19). The MC error over the j-th bin is estimated by

$$\Big|I(t_{j-1/2}) - \frac{\Phi}{\Delta t}I_{\mathrm{MC}}(t_{j-1/2})\Big| = \frac{\mathrm{C}_1}{\mathrm{Nb}^2} + \mathrm{C}_4\sqrt{\frac{\mathrm{Nb}}{\mathrm{Nr}}}, \tag{3.2.21}$$

for $\mathrm{C}_4 > 0$. Considering a fixed number of bins, we obtain that the minimal error is reached when $\mathrm{Nr} \approx \mathrm{Nb}^5$. Hence, if we double the number of bins we need to trace 2^5 times more rays.

We conclude this chapter implementing MC ray tracing for the two-faceted cup the profile of which is depicted in Fig. 3.1. Considering a set of $\mathrm{Nr} = 10^3$ random rays at the source, we obtain an example of the rays distribution on the (x,t)-plane shown in Fig. 4.9a. Since the rays are chosen randomly, the distribution at the source could be different from the one shown in that figure.

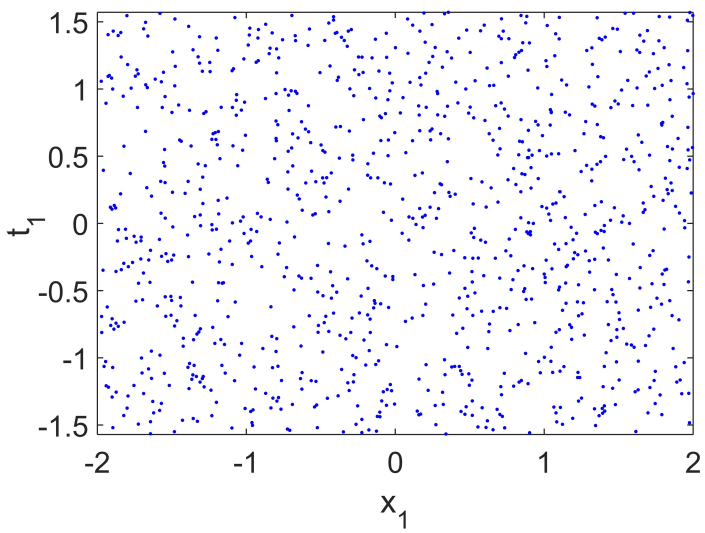


Figure 3.2: Rays at the source of the two-faceted cup with random position coordinate x and random angular coordinates t. 10^3 rays are depicted in this figure.

Then, every sample ray is traced inside the system using the ray tracing procedure.

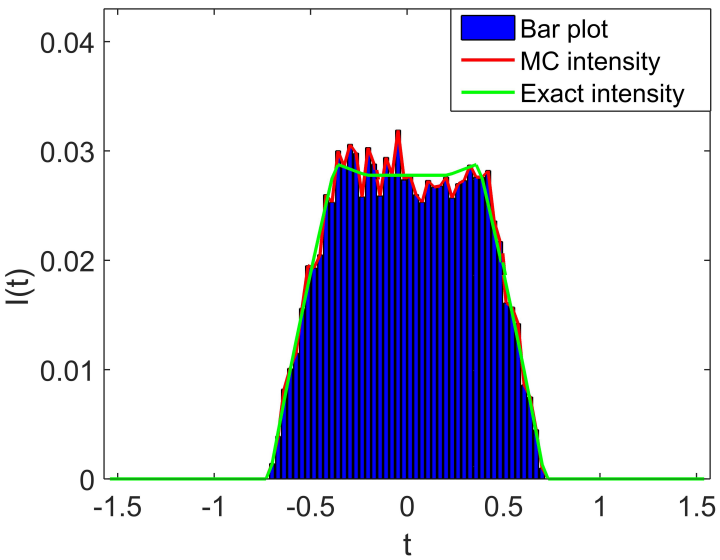


Figure 3.3: Comparison between the averaged normalized MC intensity and the normalized exact intensity.

The target $\mathsf{T} = [-b, b]$ is divided into Nb $= 100$ bins. Using Eq. (3.2.12), the normalized intensity I_{MC} is computed. I_{MC} is a piecewise constant function, therefore the averaged normalized intensity $\hat{I}_{\mathrm{MC}}(t_{j-1/2})$ is given considering the values that the intensity I_{MC} assumes on the middle point $(t_{j-1/2})_{j=0,\cdots,\mathrm{Nb}}$ of every bin of the partitioning P_1. The profile of $\hat{I}_{\mathrm{MC}}$ is depicted in Fig. 3.3 with the red line. The exact intensity (analytic intensity) is shown with the green line in the same figure. MC ray tracing has the advantages of being very easy to implement and it does not require too much regularity of the function that has to be approximate. Furthermore, the error convergence does not depend on the dimension of the domain in which the function is defined. On the other hand, MC method is time consuming as the error, for a fixed number of bins, has a speed of convergence of order $O(1/\sqrt{\mathrm{Nr}})$. Thus, to decrease the error of a factor 10 we need to increase the number of rays of a factor 100. As, MC ray tracing is a binning procedure, the error depends also on the number of bins in which the target is divided. It is a statistical procedure and the error bound is only a *probabilistic* error as shown in Eq. (3.2.10). This means that, to calculate the value of the error, several simulations have to be repeated and the average of the errors obtained in every simulation has to be calculated.

Instead of considering random variables, the sample of rays can be defined in such a way that they have a regular distribution on the domain $D \subseteq \mathbb{R}^d$ of the function f of which we want to compute the integral. Methods based on this deterministic approach are called Quasi Monte Carlo (QMC) methods. They can be seen as an improvement of MC method.

3.3 Quasi-Monte Carlo ray tracing

Quasi-Monte Carlo (QMC) methods were proposed for the first time in the 1950s in order to speed up MC. Likewise MC methods, QMC procedures can be used to approximate the integral of a function.

This chapter provides basic notions about uniform distributed theory, it follows Chapter 2 of [15]. It is useful to restrict ourselves to intervals of the form $[\boldsymbol{a}, \boldsymbol{b}) \subseteq [0,1)^d$ and introduce the concept of sequences uniformly distributed modulo one.

Definition 3.3.1. An infinite sequence $\{y_n\}_{n\in\mathbb{N}_0} \in [0,1)^d$ is said to be *uniformly distributed modulo one* (or equidistributed), if for every interval $[\boldsymbol{a}, \boldsymbol{b}) \subseteq [0,1)^d$ it holds

$$\lim_{\mathrm{N}\to\infty} \frac{\mathrm{card}(A([\boldsymbol{a}, \boldsymbol{b}), \mathrm{N}))}{\mathrm{N}} = \lambda_d([\boldsymbol{a}, \boldsymbol{b})) \tag{3.3.1}$$

where $\mathrm{card}(A([\boldsymbol{a}, \boldsymbol{b}), \mathrm{N}))$ is the cardinality of the following set

$$A([\boldsymbol{a}, \boldsymbol{b}), N) = \{n \in \mathbb{N}_0 : 0 \leq n \leq \mathrm{N} - 1 \text{ and } y_n \in [\boldsymbol{a}, \boldsymbol{b})\}, \tag{3.3.2}$$

and $\lambda_d([\boldsymbol{a}, \boldsymbol{b})) = \prod_{j=1}^{d}(b_j - a_j)$ is the d-dimensional Lesbegue measure of the interval $[\boldsymbol{a}, \boldsymbol{b})$.

Given a sequence $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}} \in [0,1)^d$ uniformly distributed modulo one and a Riemann integrable function $f : [0,1]^d \mapsto \mathbb{R}$, the integral of f can be approximate as the average of the values that f assumes on $\{\boldsymbol{y}_i\}$ for every $j = \{1, \cdots, \mathrm{N}\}$, that is:

$$\lim_{\mathrm{N}\to\infty} \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} f(y_i) = \int_{[0,1]^d} f(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}. \tag{3.3.3}$$

The idea of QMC methods is to generate the set of points in $[\boldsymbol{a}, \boldsymbol{b}]$ such that they are not randomly distributed but also not exactly uniformly distributed. To measure how much the distribution of these points differs from a uniform distribution, the concept of discrepancy was introduced. Intuitively, discrepancy measures how much the samples differ from a uniform distribution. Therefore, random sequences have a very high discrepancy, while uniform distributed sequences have zero discrepancy. The definition of discrepancy in more mathematical terms is provided below.

Definition 3.3.2. Given a set $\mathcal{S} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_\mathrm{N}\}$ of N points in $[0,1)^d$. The discrepancy $D_\mathrm{N}(\mathcal{S})$ of $\mathcal{S}$ is defined as

$$D_\mathrm{N}(\mathcal{S}) = \sup_{\boldsymbol{a},\boldsymbol{b}\in[0,1)^d} \left| \frac{\mathrm{card}(A([\boldsymbol{a}, \boldsymbol{b}), \mathrm{N}))}{\mathrm{N}} - \lambda_d([\boldsymbol{a}, \boldsymbol{b})) \right| \tag{3.3.4}$$

where $\lambda_d([\boldsymbol{a}, \boldsymbol{b})) = \prod_{j=1}^{d}(b_j - a_j)$ is the d-dimensional Lesbegue measure of the interval $[\boldsymbol{a}, \boldsymbol{b})$.

Often, it is enough to consider the discrepancy in the intervals $[0, \boldsymbol{a}) \subseteq [0,1)^d$, in that case we talk about star discrepancy.

Definition 3.3.3. Let $\mathcal{S} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{\mathrm{N}}\}$ be a set of N points in $[0,1)^d$. The star discrepancy $D^*_{\mathrm{N}}(\mathcal{S})$ of $\mathcal{S}$ is defined as:

$$D^*_{\mathrm{N}}(\mathcal{S}) = \sup_{\boldsymbol{a}\in[0,1)^d} \left| \frac{\mathrm{card}(A([0,\boldsymbol{b}),\mathrm{N}))}{\mathrm{N}} - \lambda_d([0,\boldsymbol{a})) \right| \tag{3.3.5}$$

Sequences constructed such that the corresponding star discrepancy has an order of $O(\log(\mathrm{N})^d/\mathrm{N})$ are called *low-discrepancy sequences*, [13]. An important results shows that, using a low-discrepancy sequence $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}}$, the absolute error of a QMC algorithm:

$$\mathrm{err}(f, S_{\mathrm{N}}) = \left| \int_{[0,1)^d} f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i) \right| \tag{3.3.6}$$

can be bounded by the product of a term that depends on f and another term that depends on the discrepancy of the set $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}}$. This is the result provided by the Koksma-Hlawka inequality which gives the following estimation of the error:

$$\left| \int_{[0,1)^d} f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} - \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} f(\boldsymbol{y}_i) \right| \leq V(f) D^*_{\mathrm{N}}(\mathcal{S}) \tag{3.3.7}$$

where $V(f)$ is the so-called variation function of f in the sense of Hardy-Krause (see [19, 20] for details). From the definition of low-discrepancy sequences and from the Koksma-Hlawka inequality we can state that:

$$\mathrm{err}(f, S_{\mathrm{N}}) < C\frac{\log(\mathrm{N})^d}{\mathrm{N}}. \tag{3.3.8}$$

For small dimensions d, QMC performs much better than MC methods, while for large dimension d the factor $\log(\mathrm{N})$ could be very big. The convergence of QMC method depends on the of low-discrepancy sequence that is used.

There are many ways to generate low-discrepancy sequences. The most common QMC approach uses the so-called Sobol' sequence. The algorithm for generating Sobol' sequences is widely explained in the literature, (see for instance , [21]). In appendix A we give an overview of how these kind of sequences can be constructed.

Based on QMC methods, QMC ray tracing considers as position and angular coordinates of the rays at the source, the coordinates of the corresponding points of a low-discrepancy sequence. Therefore, to implement QMC ray tracing in two-dimensions we need to construct a low-discrepancy sequence in two-dimensions. Given, for instance, a Sobol' sequence $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}}$ with $\boldsymbol{y}_i \in [0,1)^2$ for every $i = 1, \cdots, \mathrm{N}$, the two dimensional QMC ray tracing consider the position coordinate of the i-th ray at the source equal to the first component of the i-th point $\boldsymbol{y}_i$ of the Sobol' sequence $\{\boldsymbol{y}_i\}_{i=1,\cdots,\mathrm{N}}$ and, the direction coordinate of the i-th ray at the source equal to the second component of the i-th point $\boldsymbol{y}_i$ of the same sequence. A set of $\mathrm{Nr} = \mathrm{N}$ rays with these initial coordinates is traced within the system and, once the target coordinates of all the rays traced are computed, the output intensity is calculated using the same approach used for MC ray tracing, see Eqs 3.2.12 and 3.2.15. The difference between MC and QMC ray tracing consists only on the choice of the initial ray set.

In Fig. 4.9b we show the distribution of the position and direction coordinates of

the rays at the source of the two-faceted cup in Fig. 3.1. A set of 10^3 rays generated from a $2D$ Sobol sequence is considered, the coordinates (x_1, t_1) of every ray at the source are depicted with blue dots. We note that the rays have a regular distribution on the (x, t)-plane. We need to remark that, for the system in Fig. 3.1, $x_1 \in [-2, 2]$ and the angular coordinates $t_1 \in [-\pi/2, \pi/2]$. Since Sobol' sequences are defined inside intervals of the $[\boldsymbol{a}, \boldsymbol{b}) \subseteq [0, 1)^2$, we scaled the points of the sequence $\boldsymbol{y}_i$ in order to take all the possible positions and directions that the rays tan assume at the source.

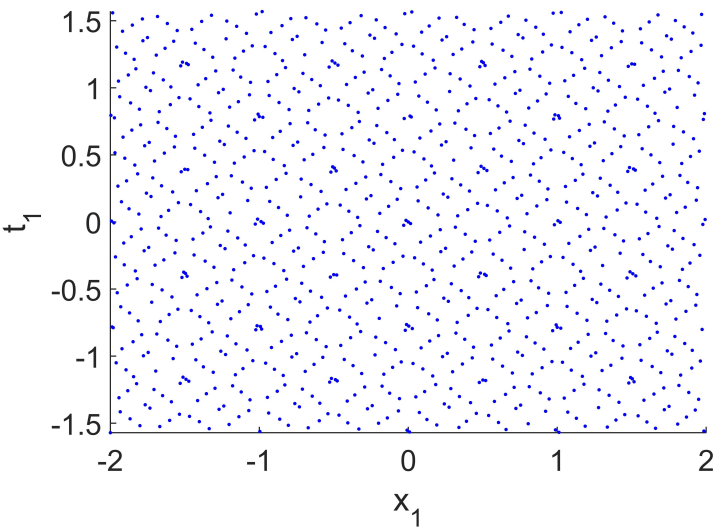


Figure 3.4: 10^3 rays at the source of the two-faceted cup with position x_1 and angular t_1 coordinates with a regular distributions. They are distributed as the points of a Sobol' sequence in two-dimensions.

Dividing the target into Nb = 100 bins, we computed the target intensity. In Fig. 3.5 we show the profile of the output intensity at the target of the two-faceted cup computed using QMC ray tracing with 10^4 rays. The QMC intensity is depicted with the red line. It is compared to the analytic intensity shown in the same figure with the green dotted line. A comparison between Fig. 3.3 and 3.5 gives the insight that for the two-faceted cup and for a set of Nr = 10^4 rays, QMC ray tracing performs better than MC ray tracing. In order to show the accuracy obtained using MC and QMC methods, we calculate the target intensity gradually increasing the number of rays traced inside the two-faceted cup. The error between the approximates intensity and the analytic intensity is calculated for every sample of rays. The speed of convergence for MC is shown in Fig. 3.6 with the red, while the behavior of QMC ray tracing is depicted in the same picture with the blue line. The results shown for a simple optical system are indeed consistent with what we expected from the theoretical analysis. Although QMC ray tracing is an improvement of MC ray tracing for small dimensions, it has two main disadvantages. First, its convergence is strongly related with the dimension in which it is implemented. Second, likewise MC ray tracing, QMC ray

tracing is a binning procedure, therefore the error still depends on the number of bins in which the target is divided and only the averaged value of the intensity over every bin is provided.

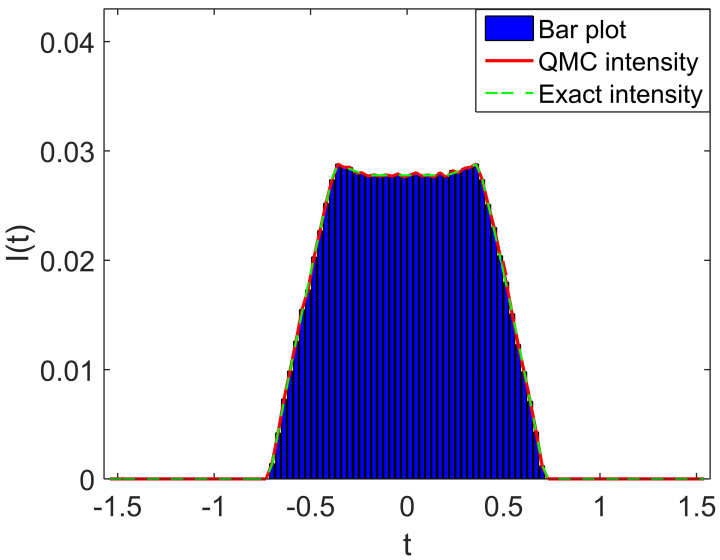


Figure 3.5: QMC intensity for the two-faceted cup obtained tracing Nr = 10^4 rays and dividing the target into Nb = 100 bins.

From the results provided in this chapter we can conclude that the choice of the initial ray set can make a big impact on the performance of the ray tracing procedure. Based on the idea of taking a smart choice of the initial ray set, we develop a new ray tracing method which is based on phase space. The phase space (PS) concept will be introduced in the next chapter. The new ray tracing method employs the PS of the source and the target of the optical systems. We will show in this thesis that phase space ray tracing allows to trace only few rays inside the system to obtain the desired accuracy of the target intensity.

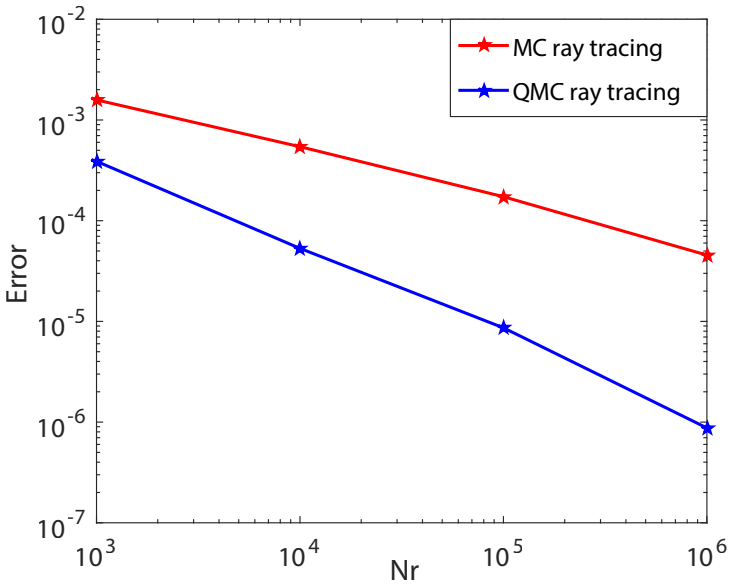


Figure 3.6: Error as function of the number of rays traced in a logarithmic scale for fixed number of bins Nb = 100. MC ray tracing convergence is of the order $O(1/\sqrt{\mathrm{Nr}})$ and it is shown with the red line. QMC ray tracing convergence is of the order $O(1/\mathrm{Nr})$ and it is depicted with the blue line.

Chapter 4

Ray tracing on phase space

Ray tracing on phase space is a method which employs the phase space (PS) of the source and the target of the optical systems. Moreover, it takes into account of the trajectory that every ray follows during its propagation. Before explaining the method, we introduce the PS concept.

4.1 Phase space concept

The PS of a three-dimensional systems is a four-dimensional space, indeed every ray is described by two position coordinates and two direction coordinates. The two position coordinates are given by two of the coordinates of the intersection point of the ray with the surface, while the two direction coordinates are the momentum coordinates of the vector tangent to the ray projected on the optical surface, [22].

For two-dimensional systems every ray in PS is given by a point in a plane. Given an optical line j, the ray position coordinate on PS is the x-coordinate of the intersection point between the ray and line j. The direction coordinate is the sine of the angle that the ray forms with respect to the normal of line j multiplied by the index of refraction of the medium in which the ray is located. We indicate the PS with $S{=}Q{\times}P$, where Q is the set of the position coordinates q and P is the set of the direction coordinates $p = n\sin\theta$, with θ the angle between the ray segment inside the system and the normal $\boldsymbol{\nu}$ of the line which we chhose inwards. Angle θ is measured conterclokwise and $\theta \in [-\pi/2, \pi/2]$. The index of refraction of the medium in which the line is located is indicated with n. The normal $\boldsymbol{\nu}$ is always directed inside the same medium in which the incident ray travels and, the angle θ between the ray and $\boldsymbol{\nu}$ is measured counterclockwise. In the following, the phase space is considered only for the source S and the target T and for no other line of the optical system. The coordinates of every ray on S and T are indicated with (q_1, p_1) and (q, p), respectively.

As an example, in Figures 4.1 and 4.2 we show the source and target PS of the two-faceted cup (in Figure 3.1), respectively. A sample of 10^4 random rays randomly are traced within the system. The coordinates of every point in Figure 4.1 correspond to the position and direction coordinates of a ray at the source, while the coordinates of every point in Figure 4.2 correspond to the position and direction coordinates of a ray at the target, which are calculated using the ray tracing procedure. Furthermore,

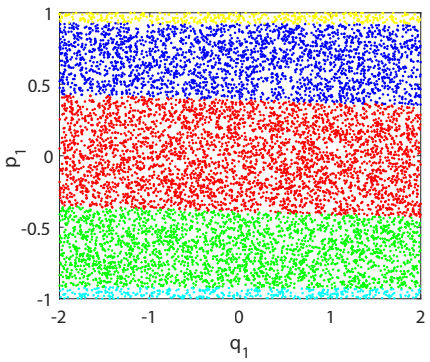


Figure 4.1: Source PS of the two-faceted cup. Five different paths can occur.

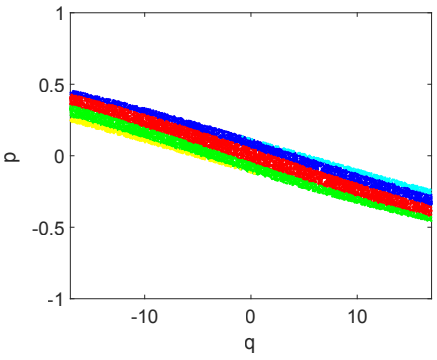


Figure 4.2: Target PS of the two-faceted cup. Five different paths can occur

we store the path that every ray follows, where we refer to a path as the sequence of the lines encountered by the ray. In Figures 4.1 and 4.2 a color is associated to every path, hence all the rays that follow the same path are depicted with the same color. We note that the source and target phase spaces are partitioned into different regions according to the path Π followed by the rays. Given a path Π, the corresponding regions are indicated with $R_1(\Pi)$ and $R(\Pi)$ at the source and the target PS, respectively. Rays that propagate through the two-faceted cup can follow 5 different paths. Some rays are emitted from the source and arrive to the target without hit any other line, they follow path $\Pi_1 = (1, 4)$. These rays are depicted in red in the PS pictures. Some other rays can hit the left or the right reflector (line 2 and 3, respectively) once, their corresponding paths are $\Pi_2 = (1, 2, 4)$ and $\Pi_3 = (1, 3, 4)$, respectively. These rays are the blue and green dots in PS, respectively. Finally, there is the possibility that the rays have two reflections before hit the target. They follow either path $\Pi_4 = (1, 2, 3, 4)$ or path $\Pi_5 = (1, 3, 2, 4)$ and they are depicted with the yellow and cyan points, respectively.

For the two-faceted cup all light emitted by the source arrives at the target. In order to derive the photometric variables at the target we need to understand where the light ends up, i.e. which parts of the target PS are illuminated by the source. Indeed, while the source PS is completely covered by rays, some parts of the target PS are not reached by any ray, that is

$$\begin{aligned} S &= \bigcup_{\Pi} R_1(\Pi), \\ T &\supset \bigcup_{\Pi} R(\Pi), \end{aligned} \tag{4.1.1}$$

where the union is over all the possible paths. This means that, while the luminance at the source PS is positive for any possible position and direction, the luminance at the target PS is positive only inside the regions $R(\Pi)$, for every path Π, and it is equal to 0 outside those regions. For this reason, from now on we will refer to $R(\Pi)$ as the regions with positive luminance.

It is very important to remark that, although S and T have a different rays

distribution, the area covered by the rays is conserved. This is due to the fact that energy is conserved along a ray. Indeed, from (2.1.19) we rewrite the two-dimensional étendue in PS as:

$$U = \int_Q \int_P \mathrm{d}q\,\mathrm{d}p. \tag{4.1.2}$$

Therefore, in two dimensions, étendue can be seen as an area in PS. Étendue conservation leads to the conservation of the areas of regions with positive luminance.

From this follows a fundamental principle in non-imaging optics which is referred to as "the edge-ray principle". A literature overview of this principle is given in the next paragraph.

4.2 The edge-ray principle

The goal in non-imaging optics is to obtain the optimal transport of light from source to target. Several methods to design ideal optical systems are based on the edge-ray principle. Basically it states that all the light rays exiting the edges of the source will end up to the edges of the target. This guarantees that all light emitted from the source will arrive to the reflector, see Figure 4.3. Therefore, the edge-ray principle constitutes a tool for designing ideal optical systems, [23, 24].

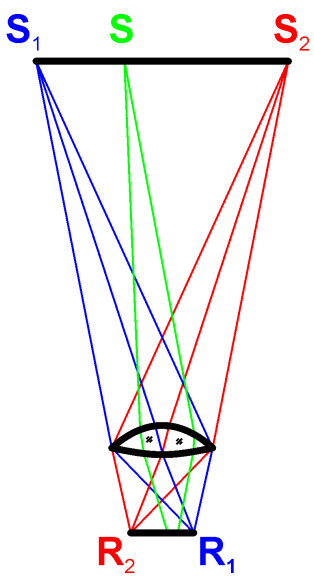


Figure 4.3: A lens that receives light from a source S_1S_2 and redirects it to the receiver R_1R_2. Rays that leave the edges of the source hit the edges of the target (blue and right rays). Rays coming from the interior of the source will end up to the interior of the target (green rays), [25].

In 1985 Miñano proved the principle by using the phase space of the source and the target of the optical systems, [26]. Suppose that R_1 and R are the regions of the rays

at the source and the target PS, respectively. Indicating with $T(P)$ the trajectory of a point P, Miñano showed that if $T(\partial R_1) = T(\partial R)$ then $T(R_1) = T(R)$. He proved the principle for system in inhomogeneous media where the index of refraction is a continuous function, so that the mapping of the regions in phase space is a continuous function. However, for some optical systems, as for instance a Compound parabolic Concentrator (CPC), the ray mapping in phase space is not continuous. This is due to multiple reflections that rays can encounter with the reflectors. This implies that some rays at the edge of the source could not be mapped into rays at the edges of the target.

In 1994 Ries and Rabl reformulated the edge-ray principle such that it is valid also for systems where the mapping in phase space is not continuous, [27]. They showed that for a given path Π. the boundaries $\partial R_1(\Pi)$ at the source are mapped into the boundaries $\partial R(\Pi)$ at the target. Then, to map one region from S to T it is sufficient to map the boundary of this region. Note that no ordered transformation of the rays from $\partial R_1(\Pi)$ to $\partial R(\Pi)$ is required. It is sufficient that the rays of $\partial R_1(\Pi)$ are transformed to the rays of $\partial R(\Pi)$, [24].

Using the PS concept and the edge-ray principle we develop a new ray tracing method. A non uniform distribution of the rays is provided by developing a triangulation refinement at the source PS which is explained in the next section. The triangulation refinement provides more rays close to the boundaries of the regions $R_1(\Pi)$ each of them is formed by the rays that follow the same path Π. Then, the boundaries $\partial R_1(\Pi)$ are approximated by using two different approaches explained in Chapter 5. For every path Π, the boundaries at the target $\partial R(\Pi)$ are obtained by mapping their corresponding boundaries ∂R_1 at the source.

4.3 Phase space ray tracing

PS ray tracing takes advantage of the fact that there exists an optical map $\mathrm{M} : S \mapsto T$ such that

$$\mathrm{M}(q_1, p_1) = (q, p), \tag{4.3.1}$$

for every $(q_1, p_1) \in S$. For very simple systems, like the two-faceted cup, it is possible to determine an analytic expression for M (as explained in Appendix B). This is not the case of most of the optical systems we deal with. In that case it is necessary to implement ray tracing to calculate how light is distributed at the target. As mentioned in the previous paragraph, for some optical systems M is not even continuous. Nevertheless, given a path Π, the map $\mathrm{M}(\Pi)$: $R_1(\Pi) \mapsto R(\Pi)$ which maps the regions $R_1(\Pi)$ in S into the regions $R(\Pi)$ in T is a continuous and bijective map. The edge ray principle guarantees that $\mathrm{M}(\Pi)$ maps $R_1(\Pi)$ into $R(\Pi)$ preserving topological features. In particular, the boundaries $\partial R_1(\Pi)$ are mapped into the boundaries $\partial R(\Pi)$. Employing the maps $\mathrm{M}(\Pi)$ for all the possible paths Π, the output light distribution is determined. Therefore, the photometric variables at the target can be calculated.

The luminance $L(q, p)$ at the target PS is given by:

$$\begin{aligned} &L(q, p) > 0 \quad \text{for } (q, p) \in R(\Pi), \\ &L(q, p) = 0 \quad \text{otherwise}, \end{aligned} \tag{4.3.2}$$

for some path Π. Since the luminance is conserved along a ray and a Lambertian source is considered, it is constant inside the regions $R(\Pi)$. The target intensity along a given direction $p = \text{const}$ is computed through an integration of the target luminance $L(q, p)$ and it is defined in T by:

$$I_{PS}(p) = \int_Q L(q, p) \mathrm{d}q. \tag{4.3.3}$$

The previous equation implies that, assuming a Lambertian source, the problem of computing the target intensity is reduced to the problem of calculating the boundaries $\partial R(\Pi)$ for all the possible paths Π. Indicating with $q^{\min}(\Pi, p)$ and $q^{\max}(\Pi, p)$ the minimum and the maximum position coordinates of the rays located on the boundary $R(\Pi)$ and using Eq. (4.3.2), Eq. (4.3.3) reduces to:

$$I_{PS}(p) = \sum_{\Pi} \int_{q^{\min}(\Pi,p)}^{q^{\max}(\Pi,p)} L(q, p) \mathrm{d}q = \sum_{\Pi} \left(q^{\max}(\Pi, p) - q^{\min}(\Pi, p) \right), \tag{4.3.4}$$

where the sum is over all the possible paths and the second equation holds as we assume $L = 1$ in $R(\Pi)$. Note that for every single ray only one path is possible as we are assuming that all the lines are reflective lines. Because of this, all the regions $R(\Pi)$ do not overlap each other, i.e.

$$\bigcap_{\Pi} R(\Pi) = \emptyset, \tag{4.3.5}$$

where the intersection is over all the possible paths.

From Eq. (4.3.3) we note that, using the PS structure, we could trace less rays inside the system to obtain the target intensity profile. The aim is to construct a ray tracing procedure that allow us tracing more rays close to the discontinuity of the luminance, i.e. close to the boundaries $\partial R(\Pi)$. To this purpose, a triangulation refinement at S is defined as explained in the following.

The regions $R(\Pi)$ can be defined only when some rays are traced. The procedure starts with coordinates $(q_1^k, p_1^k)_{k=1,\cdots,4}$ of the four points located exactly at the corners of S. For each of them, the paths $(\Pi^k)_{k=1,\cdots,4}$ are calculated. Next, for some $k \in \{1, \cdots, 4\}$, the grid is divided into two equal triangles joining two opposite vertices. For each triangle the rays located at its corners are traced. If not all the paths corresponding to those rays are equal, one or more boundaries $\partial R(\Pi)$ are expected to cross the triangle. In that case, the middle points $(q_1^k, p_1^k)_{k=5,6,7}$ of each side of the triangle are added and the three corresponding rays are traced (unless they were already traced in the previous steps). Each refinement step leads to four new triangles (see Figure 4.4).

When all the rays corresponding to the corners of each triangle have the same path, it is not necessary to refine the triangles anymore. The triangles very close to the boundaries have always vertices whose corresponding rays follow different paths. Indeed, since they are crossed by at least one boundaries, a minimum of two different paths are found for the rays at the vertices of those triangles. Because of this, the procedure could continue infinitely, therefore, two parameters $\varepsilon_q^{\min}$ and $\varepsilon_p^{\min}$ are introduced to defined a stopping criterion. The algorithm stops when the length of

the sides of the triangles is smaller than $\varepsilon_q^{\min}$ and $\varepsilon_p^{\min}$. Indicating all the possible paths with $(\Pi_j)_{j=1,\cdots,\mathrm{Np}}$ where Np is the maximum number of paths[1] (Np= 5 for the two-faceted cup). If the size of the triangles is too big, it can happen that a region formed by rays that follow a path Π_j is located completely inside a triangle whose vertices are related to the same path Π_i with $j \neq i$, see Figure 4.5. To avoid this, two parameters $\varepsilon_q^{\max}$ and $\varepsilon_p^{\max}$ are defined for the q_1-axis and the p_1-axis, respectively. When the length of the sides of the triangle are greater than these parameters, a new triangle is defined even if its vertices correspond to the same path. The values of the parameters $\varepsilon_q^{\min}$, $\varepsilon_p^{\min}$, $\varepsilon_q^{\max}$ and $\varepsilon_p^{\max}$ determine the number of rays traced. Thus, on the one hand, decreasing $\varepsilon_q^{\min}$ and $\varepsilon_p^{\min}$ more rays close to the boundaries are obtained; on the other hand, decreasing the values of $\varepsilon_q^{\max}$ and $\varepsilon_p^{\max}$ more rays in the interior of the regions are traced.

The triangulation refinement is provided by Algorithm 1 which uses the two recursive functions Left Triangle and Right Triangle. The function Left Triangle is defined in Algorithm 2 (see Figure 4.6). A similar procedure gives the function Right Triangle (see Figure 4.7).

Algorithm 1 Triangulation refinement algorithm

Initialize $\varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}$ and, $\varepsilon_p^{\max}$, Ray = [empty];
▷ $\varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}$ and, $\varepsilon_p^{\max}$ are fixed parameters needed to stop the procedure
▷ Ray: structure that contains all the information about the rays traced.

1: $(q_1^1, p_1^1) \leftarrow$ left and bottom corner of source PS: (-a, -1)
2: $(q_1^2, p_1^2) \leftarrow$ right and bottom corner of source PS: (a, -1)
3: $(q_1^3, p_1^3) \leftarrow$ right and upper corner of source PS: (a, 1)
4: $(q_1^4, p_1^4) \leftarrow$ left and upper corner of source PS: (-a, 1)
5: **for** $k = 1 \rightarrow 4$ **do**
6: Trace the ray with initial coordinates (q_1^k, p_1^k) in S;
7: Calculate the corresponding path Π^k;
8: Ray.$q \leftarrow$ [Ray.q, q_1^k];
9: Ray.$p \leftarrow$ [Ray.p, p_1^k];
10: Store the corresponding path Π^k.
11: Ray.$\Pi \leftarrow$ [Ray.Π, Π^k];
12: **end for**
13: VL $\leftarrow [1, 2, 4]$ ▷ VL = vertices of the left triangle
14: VR $\leftarrow [2, 3, 4]$ ▷ VR = vertices of the right triangle
15: Left Triangle(VL, Ray, $\varepsilon_{q_1}^{\min}, \varepsilon_{q_1}^{\max}, \varepsilon_{p_1}^{\min}, \varepsilon_{p_1}^{\max}$) ▷ Refine the left triangle
16: Right Triangle(VR, Ray, $\varepsilon_{q_1}^{\min}, \varepsilon_{q_1}^{\max}, \varepsilon_{p_1}^{\min}, \varepsilon_{p_1}^{\max}$) ▷ Refine the right triangle
17: **return** Ray;

Figure 4.8 shows an example of a triangulation refinement at the source PS of the two-faceted cup in Figure 3.1. For this optical system, the width of the q_1-axis in source phase space is two times the width of the p_1-axis. Thus, our choice is $\varepsilon_p^{\min} = \frac{1}{2}\varepsilon_q^{\min}$ and $\varepsilon_p^{\max} = \frac{1}{2}\varepsilon_q^{\max}$. with $\varepsilon_q^{\min} = 0.1$ and $\varepsilon_q^{\max} = 1$. Using the

[1] We use the apex to indicate the path Π^k followed by rays corresponding to the coordinates (q_1^k, p_1^k), can happen $\Pi^k = \Pi^h$ for $k \neq h$. The subscript refers to the index of one of the possible paths $(\Pi_j)_{j=1,\cdots,\mathrm{Np}}$, therefore $\Pi_i \neq \Pi_j$ if $i \neq j$.

triangulation refinement, all the possible paths $(\Pi_j)_{j=1,\cdots,\mathrm{Np}}$ are found and their corresponding regions $R_1(\Pi_j)_{j=1,\cdots,\mathrm{Np}}$ are determined. Using the edge-ray principle, we conclude that also the regions $R(\Pi_j)_{j=1,\cdots,\mathrm{Np}}$ at the target are determined and only the rays close to the boundaries ∂R_1 need to be considered to obtain the target rays distribution.

4.4 Conclusions

In this chapter we introduced the phase space concept. We explained a new ray tracing method based on the source and the target PS representation. In PS every point corresponds to a unique ray. The coordinates of every point correspond to the initial ray position q_1 coordinate and the initial ray direction $p_1 = \sin(\theta_1)$ coordinate (expressed with respect to the normal of the source). The method also takes into account of the paths followed by every ray traced. Considering only reflection law, every single ray follows only one path and, therefore, the PS regions do not overlap to each others.

As an example, we provide the source and the target PS representation of a very simple optical system which is the two-faceted cup. The edge-ray principle guarantees that all the rays that follow the same path are located in the same regions in PS. If we know these regions at the source we can determine the corresponding regions at the target. It is sufficient to map the boundaries at the source $\partial R_1(\Pi)$ to obtain their corresponding target boundaries $\partial R(\Pi)$.

The boundaries $\partial R(\Pi)$ are particularly relevant because there the discontinuity of the luminance occurs. Assuming a lambertian source, only the rays on the boundaries are needed to compute the target intensity. Based on this idea, a triangulation in S is construct such that the rays closest to $\partial R_1(\Pi)$ are selected and more rays in their vicinity are created to get progressively better estimates of the boundaries.

In Figures 4.9 we show three different rays distributions on the source PS of the two-faceted cup. In Figure 4.9a, 10^3 random points are shown. MC ray tracing is based on this random distribution of the initial rays set. In Figure 4.9b, 10^3 points of a two-dimensional Sobol' sequence are shown. Since Sobol's sequences are defined in a unit square, we scaled it such that all the source PS $S = [-2, 2] \times [-1, 1]$ is covered by rays. QMC ray tracing considers as initial set, rays distributed as the points of a Sobol sequence. Such regular distribution can lead to several advantages for the computation of the target intensity, see Section 3.3 of this thesis. Finally, Figure 4.9c shows a non-uniform distribution of rays at the source PS. Such distribution is obtained from the triangulation refinement explained in the previous section. The procedure allows tracing more rays close to the boundaries $\partial R_1(\Pi)$ of the regions in source PS, each of them is formed by all the rays that follow the same path. From the edge ray-principle, we obtain that these rays will be located close to the boundaries $\partial R(\Pi)$ of the regions at the target PS. PS ray tracing is based on this kind of initial rays distribution at the source.

The target PS intensity is calculated using only the rays that are located on the boundaries $\partial R(\Pi)$. Thus, in order to obtain the intensity profile at the target, the boundaries $\partial R(\Pi)$ need to be determined.

In the next chapter we provide two different approaches to find the boundaries $\partial R(\Pi)$ using a set of rays given by the triangulation refinement.

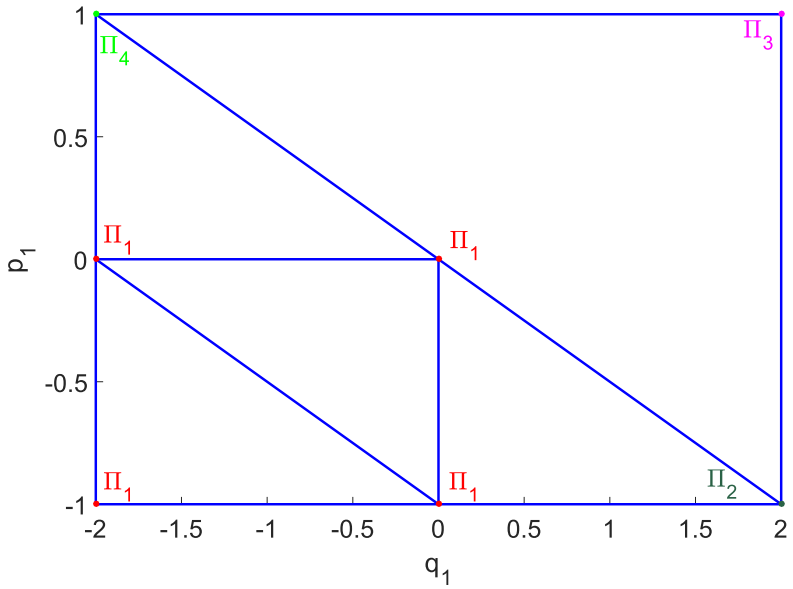


Figure 4.4: Triangulation refinement: when the rays related to the vertices of the triangles follow a different path a new refinement step is required. Each refinement step leads to four new triangles.

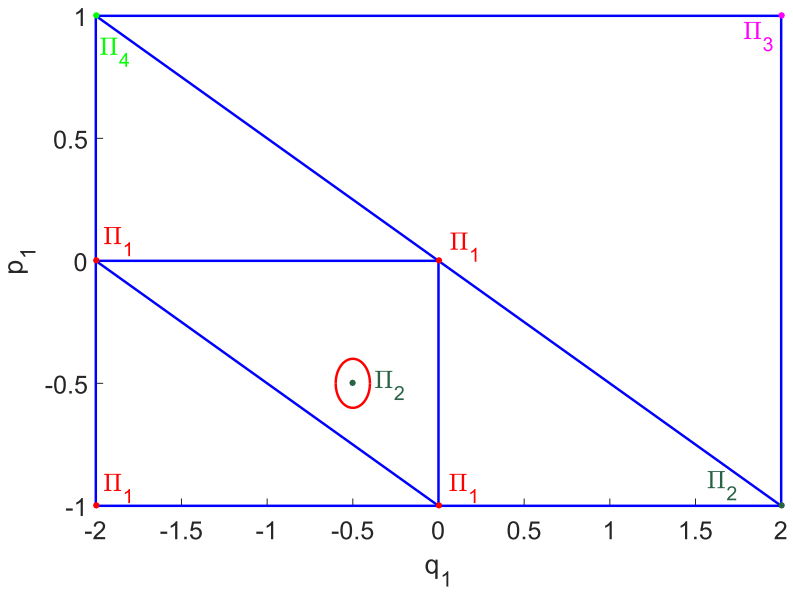


Figure 4.5: The red line encloses a region of rays that follow the path Π_2 and is completely located inside a triangle. The algorithm is not able to detect that region and, a further refinement is required.

Algorithm 2 Algorithm for the refinement of the left triangles

procedure LEFT TRIANGLE(VL, Ray, $\varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}, \varepsilon_p^{\max}$)
  VL $\leftarrow [1, 2, 4]$
  $q_1^1 \leftarrow Ray.q(\text{VL}(1))$, $p_1^1 \leftarrow Ray.p(\text{VL}(1))$
  $q_1^2 \leftarrow Ray.q(\text{VL}(2))$, $p_1^2 \leftarrow Ray.p(\text{VL}(2))$
  $q_1^3 \leftarrow Ray.q(\text{VL}(3))$, $p_1^3 \leftarrow Ray.p(\text{VL}(4))$
  $\text{dist}_q \leftarrow |q_1^2 - q_1^1|$
  $\text{dist}_p \leftarrow |p_1^3 - p_1^1|$
  RefineTriangle $\leftarrow$ false;
  DifferentPath $\leftarrow$ false;
  if $\text{dist}_q > \varepsilon_q^{\max}$ or $\text{dist}_p > \varepsilon_p^{\max}$ **then**
    RefineTriangle $\leftarrow$ true;
  end if
  for $k = 1 \rightarrow 2$ **do**
    if $\Pi^k \neq \Pi^{k+1}$ **then**
      DifferentPath $\leftarrow$ true;
    end if
  end for
  if $\text{dist}_q > \varepsilon_q^{\min}$ or $\text{dist}_p > \varepsilon_p^{\min}$ **then**
    RefineTriangle $\leftarrow$ DifferentPath;
  else
    if (DifferentPath is true) **then**
      Ray(VL).boundary $\leftarrow$ true; ▷ A boundary crosses the triangle
    end if
  end if
  if (RefineTriangle is true) **then**
    Define the points at the middle of each side of the triangle
    $(q_1^5, p_1^5) = ((q_1^1 + q_1^2)/2, p_1^1)$
    $(q_1^6, p_1^6) = (q_1^5, (p_1^1 + p_1^2)/2)$
    $(q_1^7, p_1^7) = (q_1^1, p_1^6)$
    for $k = 5 \rightarrow 7$ **do**
      if The ray with coordinates (q_1^k, p_1^k) is not traced yet **then**
        Trace the ray with initial coordinates: (q_1^k, p_1^k) in PS;
        Calculate the path Π^k;
        Compute the corresponding path Π^k;
        Store the ray's coordinates Ray.$q \leftarrow [\text{Ray}.q, q_1^k]$;
        Store the ray path Ray.$\Pi \leftarrow [\text{Ray}.\Pi, \Pi^k]$;
      end if
    end for
    return LEFT TRIANGLE$([\text{VL}(1), 5, 7], \text{Ray}, \varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}, \varepsilon_p^{\max})$;
    return LEFT TRIANGLE$([5, \text{VL}(2), 6], \text{Ray}, \varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}, \varepsilon_p^{\max})$;
    return LEFT TRIANGLE$([7, 6, \text{VL}(3)], \text{Ray}, \varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}, \varepsilon_p^{\max})$;
    return RIGHT TRIANGLE$([5, 6, 7], \text{Ray}, \varepsilon_q^{\min}, \varepsilon_q^{\max}, \varepsilon_p^{\min}, \varepsilon_p^{\max})$;
  end if
return Ray;
end procedure

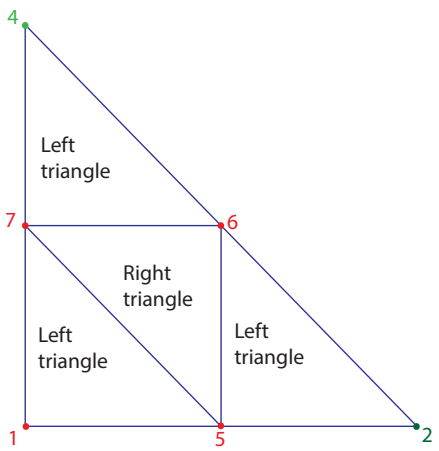


Figure 4.6: Left triangulation refinement algorithm (recursive function LEFT TRIANGLE).

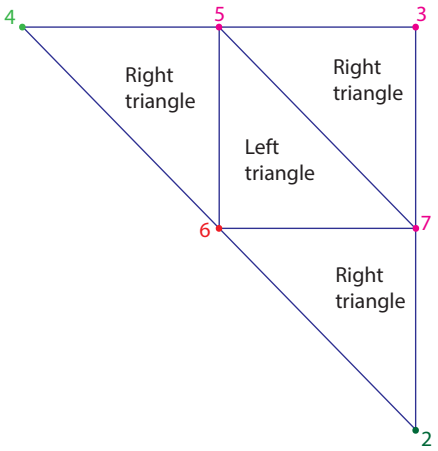


Figure 4.7: Right triangulation refinement algorithm (recursive function RIGHT TRIANGLE).

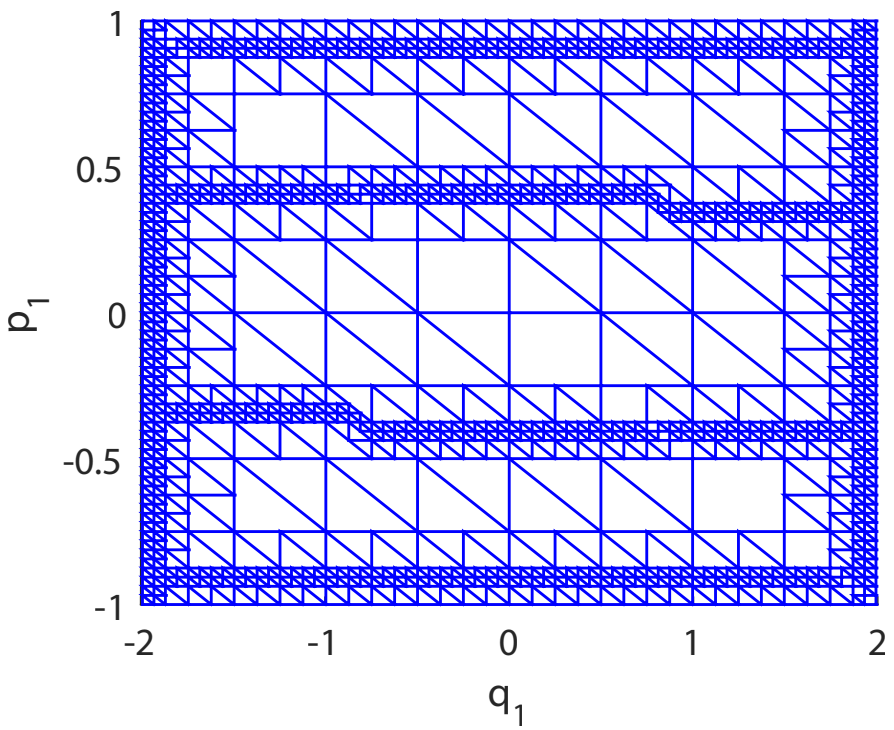
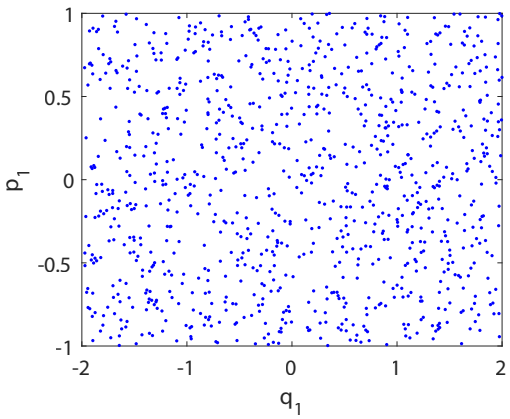
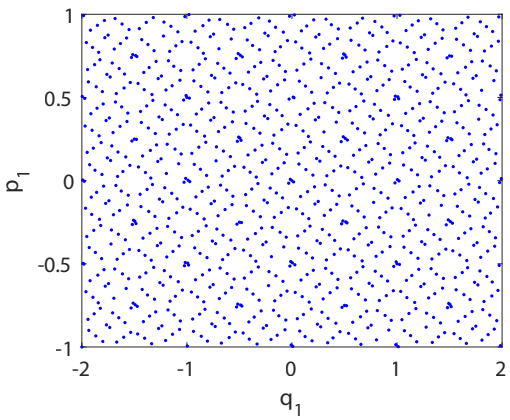


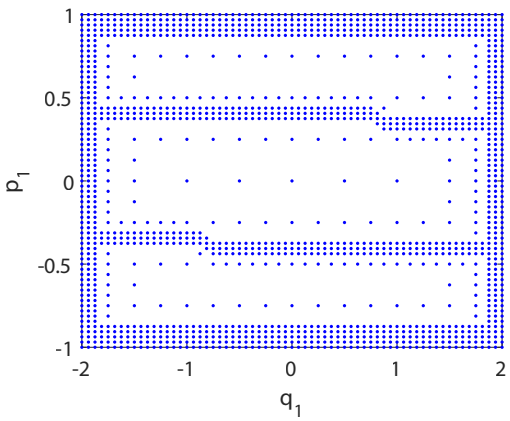
Figure 4.8: Triangulation refinement of source phase space: near the boundaries more rays are traced. The values of the parameters are $\varepsilon_{q_{min}} = 0.1$ and $\varepsilon_{q_{max}} = 1$.



(a) 10^3 random rays at the source PS (MC ray tracing).



(b) 10^3 rays at the source PS distributed as the point of a Sobol' sequence (QMC ray tracing).



(c) $1,5 \cdot 10^3$ rays distributed using the triangulation refinement (PS ray tracing).

Figure 4.9: Three different rays distribution at the source of the two-faceted cup.

Chapter 5

The α-shapes approach to compute the boundaries in target phase space

The α-shape method which is widely used for reconstruction an unknown surface from a set of data points. We develop a technique based on α-shape to obtain an approximation of the boundaries of the set of points obtained from the triangulation refinement, [28]. The aim is to give a criterion to determine the value of the parameter α that gives the better approximation of the boundaries. The results are given in Section 5.3 for two different kind of Total Internal reflection (TIR) collimator.

5.1 The α-shapes approach

Given a finite set V of points α-shapes are geometrical objects that give us an approximation of the "shape"[1] formed by the points in the set.

Before giving a formal definition, we explain an intuitive and nice interpretation of α-shapes. We can think to a mass of a stracciatella[2] ice-cream. If we desire to know the shape formed by the chocolate pieces, we can start eat the ice cream using a spoon with a spherical shape trying to do not remove any piece of chocolate. We will obtain a shape formed by arcs and points, see Figure 5.1 for the two-dimensional case. Straightening the arcs to line segments we obtain broken lines which constitute the boundaries the so-called α-shape of V. In the ice-cream example, the chocolates peaces are the points of the point set and, the parameter α determines the radius of the carving spoon (the spherical spoon in two-dimension is simply a circle). A very small spoon will allow us to eat the entire ice cream without eating any piece of chocolate, while with a huge spoon we are not able to eat any part of the ice cream because we will always take away at least one chocolate peace.

The previous example might give a better understanding of the definition of α-shape first given by Edelsbrunner, Kirkpatrick and Seidel in 1983, [30]. In their paper

[1] It will become clear through this chapter what we intend with the word *shape*.

[2] Stracciatella ice cream is made with milk-based ice-cream and fine peaces of chocolate, [29].

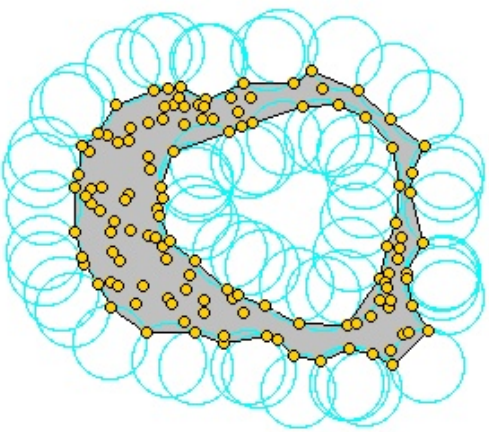


Figure 5.1: Construction of α-shape given a set of points in $\mathbb{R}^2$

they describe the α-shape has a generalization of the convex hull of a finite set of point in the plane. Let be α a not negative number $0 \leq \alpha < \infty$. If α is equal to 0 the shape degenerates to the point set V. On the other hand, when $\alpha \to \infty$ the α-shape is simply the convex hull. Finally, if $0 < \alpha < \infty$ the α-shape is a polytope of the point set, [31]. The α-shapes are closely related to Delaunay triangulation of V,[32].

Given a set V of not all aligned points, let us consider the set E of all the straight line segments whose endpoints are in V. A triangulation T of V is the maximum subset of E such that all the line segments of T intersect only at their endpoints, [33].

Delaunay triangulation can be seen as the dual of Voronoi diagram, [34].

We give here the definition Voronoi diagram only in two dimensions, (see [35] for the general case). In every finite set of point $V = \{v_1, \cdots, v_n\} \subset \mathbb{R}^2$ and for *almost*[3] every point $x \in \mathbb{R}^2$, there is a point which is the closest point to x. The Voronoi cell of a point $v_i \in V$ contains all points in $\mathbb{R}^2$ which are closer to v_i, see Figure fig:Voronoi. The Voronoi diagram is the set of all Voronoi cells, [36]. A formal definition of the Voronoi diagram is given in the following.

Definition 5.1.1. Let X be a metric space with a distance d and $V = \{v_1, \cdots, v_n\}$ a set of point in X. The Voronoi cell V_i associated with the point v_i with $v_i \in \{1, \cdots, n\}$ is defined as:

$$V_i = \{x \in X \mid \mathrm{d}(x, v_i) \leq \mathrm{d}(x, v_j) \quad \forall j \neq i\}\,, \tag{5.1.1}$$

The Voronoi diagram is defined as the union $U = \bigcup_{i=1}^{n} V_i$ where $V_i \cap V_j = \emptyset$ for $i \neq j$.

Figure 5.2 The Delaunay triangulation of the points set V has the property (also called Delaunay property) that the circle circumscribed by any triangle of T does not contain any point of V. A very common used algorithm to construct such triangulation is explained in the following. A Delaunay triangulation is constructed by modifying a general triangulation T such that every point satisfies the Delaunay

[3]It is needed to specify the word *almost* because some points can have the same distance with two or more points of V.

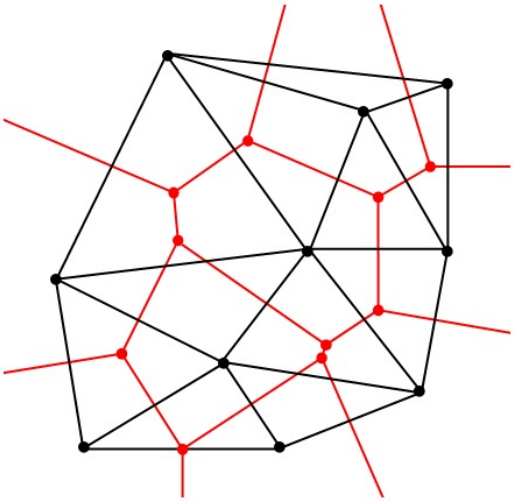


Figure 5.2: The Delaunay triangulation in black is the dual of the Voronoi diagram in red, [37].

property. Therefore, every triangle (or tetrahedron in three dimensions) that does not satisfy such property is flipped such that the new edge is part of the triangulation. More precisely, given an arbitrary triangulation T in two-dimension, for each edge $\bar{ab}$ in T which is not on the boundary of the convex hull the two triangles Δ_{abc} and Δ_{abd} with the common edge $\bar{ab}$ are found. Then, if either the circumcircle of triangle Δ_{abc} contains point d or the circumcircle of triangle Δ_{abd} contains point c the edge $\bar{ab}$ cannot be included in the Delaunay triangulation and, therefore, it is flipped such that the other two possible triangles Δ_{acd} and Δ_{bcd} are found. The new edge $\bar{cd}$ satisfy the Delaunay property locally and the triangles are added to the Delaunay triangulation, see Figure 5.3.

Several others algorithm have been developed to construct a Delaunay triangulation, see for example [38, 39]. This triangulation is unique for a given set of points V. It has the property to have the largest minimum angle among all possible triangulation of a point set V, [40]. Such triangulation triangulates the convex hull of V and, therefore it does not constitutes a suitable method for reconstruct the surface formed by a point cloud.

An important method in surface reconstruction is the α-shapes method, [41, 28]. Starting from the Delaunay triangulation T' of a point set V, the corresponding α-shape of V is formed by the only triangles of T' that satisfy the so-called "α-test". For each triangle we calculate the radius of the circumcircle. If the radius is larger that α the triangle is removed from the shape. The rule of the parameter α is highly significant in this procedure. Hence we have to choose it in such a way to get a better approximation. The choice of the parameter α is closely related to the radius of the circumcircles. A possible strategy is to find the radius of the greater empty

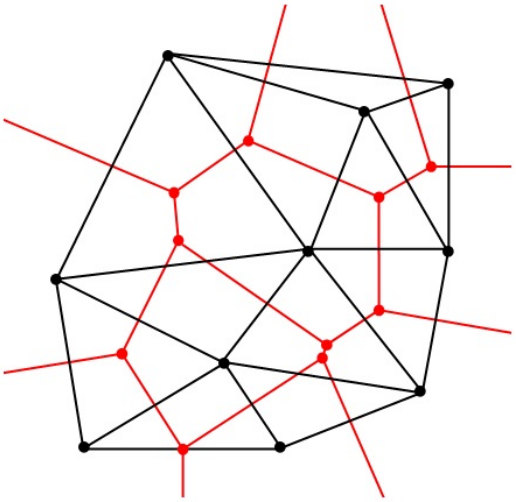


Figure 5.3: The Delaunay triangulation in black is the dual of the Voronoi diagram in red, [37].

circumcircle. Thus α is related to the density of the points. In particular we have:

$$\alpha = C\frac{1}{\Delta}\,, \tag{5.1.2}$$

with C a constant that can be determined by a simulation and Δ the density of the point set V defined as:

$$\Delta = \frac{N}{\text{surface area}}\,, \tag{5.1.3}$$

where N is the number of points in V and the surface area is the area inside the boundaries of the region formed by the points cloud. Hence Δ is a constant.

Even if α-shapes are a powerful tool to reconstruct surfaces, some simulations show that there exist surfaces that are not described well by α-shapes. Indeed for some particular surface there exist no value of α that includes all desired triangles and deletes all undesired triangles. For instance, since the parameter α depends on the density of the point cloud, is intuitively clear that using α-shapes for a non-uniform points set we won't get a good approximation of the surface. Furthermore, the α-shape method doesn't work well when there is a sharp turn or a joint. In this case α-shapes often give a "webbed-foot" appearance at such joints since they improperly connect the adjacent surfaces. Hence a generalization of "classical" α-shapes is required. In the next section a method to solve the "density problem" for two separated and close objects is described. In [42] Teichmann and Capps present "Density-scaled α-shapes".

5.2 Determination of α using étendue conservation

5.3 Results for a TIR collimator

Chapter 6

The triangulation refinement approach

Chapter 7

The inverse ray mapping method: analytic approach

Chapter 8

The extended ray mapping method

Chapter 9

Extended ray mapping method to systems with Fresnel reflection

Chapter 10

Discussion and conclusions

Appendix A

Implementation of Sobol' sequences

A.1 Van der Corput sequences

In the following we show a particular construction of a low-discrepancy sequence for $d = 1$ that was introduced the first time by Van der Corput in 1935. This kind of sequences, called *van der Corput* sequences, are particular interesting not only because they give an intuition of how to construct low discrepancy sequences but also because many other kind of sequences in higher dimensions are based on this one-dimensional case. Before introducing these sequences we need to give the concept of radical inverse function. Let $\mathrm{b} \geq 2$ be an integer base. Any natural number $n \in \mathbb{N}_0$ can be decomposed in base b as follows:

$$n = \sum_{i=0}^{\infty} d_i \mathrm{b}^i \tag{A.1.1}$$

where $d_i \in \{0, 1, \cdots, \mathrm{b}-1\}$ are the digit numbers. The radical inverse function $\phi_\mathrm{b} : \mathbb{N}_0 \mapsto [0, 1)$ in base b is defined as:

$$\phi_\mathrm{b}(n) = \sum_{i=1}^{\infty} \frac{d_{i-1}}{\mathrm{b}^i}. \tag{A.1.2}$$

As an example we provide in the following the radical inverse function $\phi_\mathrm{b}(5)$ in base $\mathrm{b} = 2$. The digit expansion in base b of $n = 5$ is:

$$5 = 1 \cdot 2^0 + 1 \cdot 2^2. \tag{A.1.3}$$

Therefore, $d_0 = 1$, $d_1 = 0$ and $d_2 = 1$. The radical inverse function $\phi_2(5)$ is:

$$\phi_2(5) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}. \tag{A.1.4}$$

Definition A.1.1. The Van der Corput sequence in base b is defined as $\{\phi_\mathrm{b}(n)\}_{n \in \mathbb{N}_0}$.

For example, suppose we have the finite sequence of numbers $n \in \{0, 1, \cdots, 8\}$ the corresponding Van der Corput sequence $\{\phi_{\rm b}(n)\}_{n\in\{0,1,\cdots,8\}}$ in base $b = 2$ is:

$$\left\{\phi_2(n)\right\}_{n\in\{0,1,\cdots,8\}} = \left\{0, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}\right\}. \tag{A.1.5}$$

It can be proved that the Van der Corput sequence in base b is uniformly distributed modulo one, [15]. The van der Corput sequence has been extended to higher dimensions. The most common QMC approach uses Sobol sequence which can be seen as an extended Van der Corput sequence in base $b = 2$. Sobol' sequence uses the same base $b = 2$ for all the dimensions $d \geq 2$.

A.2 Sobol' sequences

The aim is to generate a low-discrepancy sequence in the ipercube $[0, 1]^d$. Let us start from the simplest case of one dimension, i.e. $d = 1$. First, we need to chose a primitive polynomial P_j of degree s_j of the form

$$P_j : x^{\mathrm{s}_j} + a_{1,j}x^{\mathrm{s}_j-1} + \cdots + a_{\mathrm{s}_j-1}x + 1 \tag{A.2.1}$$

where the coefficients $\{a_{i,j}\}_{i=1,\cdots,\mathrm{s}_j-1}$ are either 0 or 1. Then a sequence $\{m_1, m_2, \cdots\}$ is defined such that:

$$m_{k,j} := 2a_{1,j}m_{k-1,j} \oplus 2^2 a_{2,j}m_{k-2,j} \oplus \cdots \oplus 2^{\mathrm{s}-1}a_{k-1,j}m_{k-\mathrm{s}+1,j} \oplus 2^{\mathrm{s}}m_{k-1,j} \oplus m_{k-\mathrm{s},j}, \tag{A.2.2}$$

where we have indicated with $\oplus$ the bit by bit exclusive or operator which operates on two bit patterns and operates on each pair of the corresponding bins giving as result 1 if one of the two bits is 1 and 0 if both bits are equal either to 0 or 1. The values $m_{k,j}$, $1 \leq k \leq d$, are chosen such that they are odd and positive numbers less than 2^k. Now, the so-called direction numbers are defined by:

$$v_{k,j} = \frac{m_{k,j}}{2^k}. \tag{A.2.3}$$

Then, the sequence $\{x_{i,j}\}$ is given by

$$x_{i,j} = i_1 v_1 \oplus i_2 v_2 \oplus \cdots \tag{A.2.4}$$

for every i, where i_k is the k-th digit from the right when i is written in binary $i = (\cdots i_3 i_2 i_1)_2$, [43]. We provide in the following an example.

Given the primitive polynomial $x^3 + x^2 + 1$ of degree $\mathrm{s}_j = 3$, the first three coefficients $m_{1,j} = 1$, $m_{2,j} = 3$, and $m_{3,j} = 7$ lead to the following direction numbers

$$v_{1,j} = \frac{1}{2}, \qquad v_{2,j} = \frac{3}{4}, \qquad v_{3,j} = \frac{7}{8}, \tag{A.2.5}$$

that in binary notation are:

$$v_{1,j} = (0.1)_2 \qquad v_{1,j} = (0.11)_2, \qquad v_{1,j} = (0.111)_2. \tag{A.2.6}$$

From Eq. (A.2.2) we can derive the others coefficients $m_{4,j} = 5$, $m_{5,j} = 7$, etc. with he corresponding direction vectors:

$$v_{4,j} = \frac{5}{16} = (0.0101)_2 \qquad v_{5,j} = \frac{7}{32} = (0.00111)_2 \tag{A.2.7}$$

From Eq. (A.2.4) we finally find the sequence

(A.2.8)

The generalization of Sobol's sequence to higher dimensions $d > 1$ is calculated considering a sequence where the i-th point has the form:

$$q_i = (x_{i,1}, x_{i,2} \cdots, x_{i,d}), \tag{A.2.9}$$

where the second index of the variables $x_{i,j}$ it refers to the polynomial P_j (with corresponding degree s_j) which is considered to calculate the direction numbers. Therefore, d different sets of direction numbers are generated from a given polynomial P_j using Eq. A.2.3 and each component $x_{i,j}$ is computed using the corresponding direction vector.

Appendix B

Calculation of the boundaries at the target PS

B.1 Analytical method to find the boundaries of the different regions in phase space

In this section, we present an analytical method to find the boundaries of the regions formed by rays that follow the same path. Furthermore, we will represent those regions on source and target phase space.

It is possible to determine the maximum number of times that a ray reflects into the two-faceted cup as follows. Rotating the entire cup we can think of the path as a straight line that hits one of the rotated targets. The idea to rotate the cup comes from the fact that in this way we consider the paths as straight lines, hence it is sufficient to find only one intersection point between the ray and one line segment (also in the case where we have more than one reflection) and finally rotate back the intersection point to find the point on the target. Next we want to explain this procedure in more detail. Our optical system is defined as in the previous section, see Figure 3.1. Let B be defined by:

$$\begin{aligned} B &= \left(h + \tfrac{a}{\tan(\gamma)}\right)\tfrac{1}{\cos(\gamma)} - \tfrac{a}{\tan(\gamma)} \\ &= \tfrac{h}{\cos(\gamma)} + a\tan\left(\tfrac{1}{2}\gamma\right), \end{aligned} \tag{B.1.1}$$

and $P : (0, B)$ is the rotation point. We define B_k as the clockwise ($k < 0$) or counterclockwise ($k \geq 0$) rotations of the point $P : (0, B)$ over an angle $\alpha_k = (2k+1)\gamma$, with γ the angle that the normal to the source forms with the reflectors of the cup and $k \in \mathbb{Z}$. The x and z-coordinates of B_k are indicated with $b_{k,x}$ and $b_{k,z}$, respectively, Figure $B.1$ is illustrative. The position vector for the points B_k is given by $\mathbf{b}_k = \begin{pmatrix} b_{k,x} \\ b_{k,z} \end{pmatrix}$ where

$$\mathbf{b}_k + \begin{pmatrix} 0 \\ \frac{a}{\tan(\gamma)} \end{pmatrix} = \begin{pmatrix} \cos(\alpha_k) & -\sin(\alpha_k) \\ \sin(\alpha_k) & \cos(\alpha_k) \end{pmatrix} \begin{pmatrix} 0 \\ B + \frac{a}{\tan(\gamma)} \end{pmatrix}. \tag{B.1.2}$$

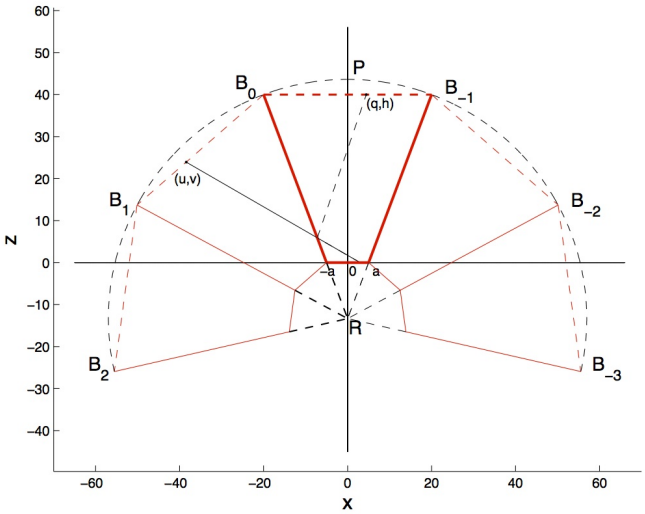


Figure B.1: The two-faceted cup rotated to both sides. The line segment $B_{k-1}B_k$ is the $|k|$ times rotated target. The point (u, v) of the intersection between a ray and the segment B_0B_1 corresponds to the point (q, h) on the target $B_{-1}B_0$. $P(0, B)$ is the point to rotate around the point $R = \left(0, -\frac{a}{\tan\gamma}\right)$. The length of the segments RB_k is equal to the radius of the dashed circle.

Then the maximum number of reflections r is:

$$r = \max\{k \in \mathbb{N} \mid b_{k-1,z} \geq 0\}\,. \tag{B.1.3}$$

This method of rotating the cup instead of reflecting the ray inside the system can also be applied to find the boundaries of the regions $M_{\mathrm{s},k}$ and $M_{\mathrm{t},k}$. In the following sections we will illustrate how this is done.

B.1.1 Source phase space

We observe that the set of rays that form the boundary of the regions $M_{\mathrm{s},k}$ only consists of rays that either leave the extremes of the source or hit one of the points B_k. In Figure $B.1$ is shown a ray that on the target phase space is located inside the region $M_{\mathrm{t},1}$, it does not constitute a point on any boundary. Furthermore, we note that the rays emitted from the corner points of the source form vertical lines in $\mathcal{P}_{\mathrm{s}}$, since $x = const$. On the other hand, rays that hit B_k form vertical lines in $\mathcal{P}_{\mathrm{t}}$, since $q = const$. Hence for the representation on the source phase space we have to choose rays that hit B_k, their directions are given by the relation

$$\tan t = \frac{x - b_{k,x}}{b_{k,z}}\,. \tag{B.1.4}$$

This is exactly what we did in the algorithm named *'Source'* (see Appendix **??** for details).

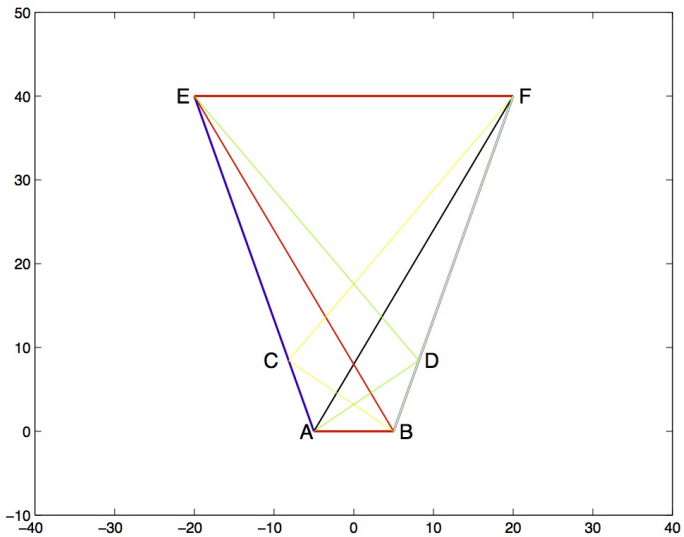


Figure B.2: Rays that leave the corner points of the source. The rays AF, BE, ACE, BDF are rays that do not hit the reflectors of the system. They constitute rays on the boundaries of the regions $M_{\mathrm{s},0}$, $M_{\mathrm{s},1}$ and $M_{\mathrm{s},-1}$. The rays ADE and BCF are rays that hit once the reflectors of the system. They constitute rays on the boundaries of the regions $M_{\mathrm{s},-1}$, $M_{\mathrm{s},-2}$, and $M_{\mathrm{s},1}$ or $M_{\mathrm{s},2}$, respectively.

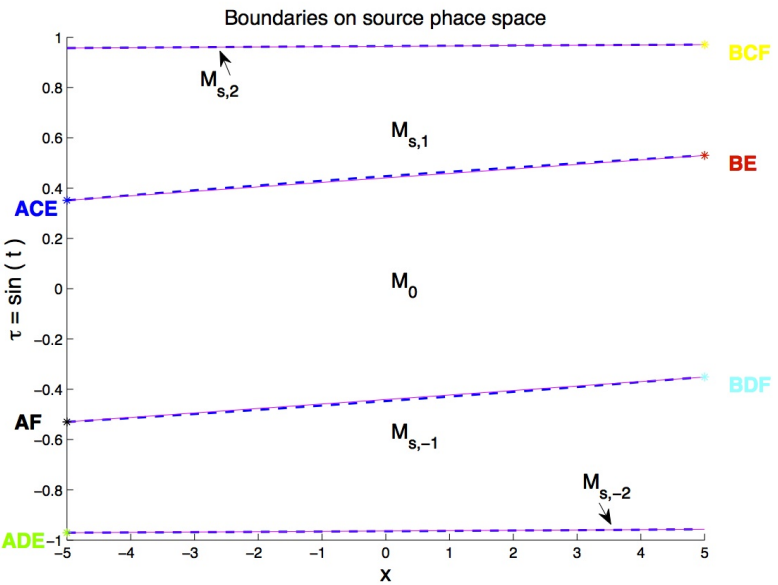


Figure B.3: Regions $M_{s,k}$ of rays that reflect $|k|$ times, with $(x,\tau) \in \mathcal{P}_s$. The parameter values are: $a = 5$, $b = 20$ and $h = 40$. The continuous lines are the boundaries of the regions M_s calculated considering rays that leave the source and hit the points B_k at the target. The dashed blue lines are the boundaries calculated using $(B.1.4)$

In Figure $B.2$ are shown some rays that compose the boundaries of $M_{s,k}$ which coordinates are:

$$ADE = \left(-a, \arctan\left(\frac{-a+b_{-1,x}}{b_{-1,z}}\right)\right),\ ACE = \big(-a, \sin(\gamma)\big),\ AF = \big(-a, -\sin(\delta)\big),$$
$$BCF = \left(a, \arctan\left(\frac{a-b_{1,x}}{b_{1,z}}\right)\right), BDF = \big(a, -\sin(\gamma)\big) \text{ and } BE = \big(a, \sin(\delta)\big).$$

The rays are represented by points in phase space. So we choose a proper number of rays that leave the source to obtain an accurate representation of the boundaries of $M_{s,k}$ regions. The final result is shown in Figure $B.3$. In addition, we derive the exact equation for the map $\mathcal{M}$. From equation $(B.1.4)$ we find the value of the angle for each ray at the source (depending on the ray position). Thus the boundaries are simply straight lines in the $(x, \tan(t))$-plane. The subdivision of phase space into regions is shown in Figure $B.3$, where we can also see the comparison between the two different methods to calculate the boundaries. Note that in this specific case the boundaries appear straight lines also in the $(x, sin(t))$-plane.

B.1.2 Target phase space

In this section we derive an exact expression for the map $\mathcal{M}$ in such a way that it is possible to determine the boundaries of the regions $M_{t,k}$ simply by finding the images

of some points on $\partial M_{\mathrm{s},k}$. Given a ray parameterization we are able to calculate the intersections point (u, v) between the ray and the line segment $B_{k-1}B_k$ as we did in *'Target'* (See Appendix **??** for the procedure). The corresponding point (q, h) on the target can be found by rotating or reflecting the point (u, v) back for k even or odd, respectively. Therefore we have the following expression for the point (q, h) on the target:

$$\begin{pmatrix} q \\ h \end{pmatrix} = \begin{pmatrix} (-1)^k & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos(-2k\gamma) & -\sin(-2k\gamma) \\ \sin(-2k\gamma) & \cos(-2k\gamma) \end{pmatrix} \begin{pmatrix} u \\ v + \frac{a}{\tan(\gamma)} \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{a}{\tan(\gamma)} \end{pmatrix}. \tag{B.1.5}$$

We observe that the sign depends on the parity of k. When $k = 0$, i.e. the ray does not reflect, the first and the second matrices become the identity matrix and the cup is not rotated nor reflected. When k is even, the determinant of the product between the first and the second matrixes at the right hand of equation $(B.1.5)$ is equal to 1 and we obtained a rotation matrix, while when k is odd the determinant of the matrix given by the product between the first and the second matrix is equal to -1 and we have a reflection matrix. Also the angle on the target is calculated. It is an addition of an angle and a change of sign depending on k:

$$\theta = (-1)^k(t - 2k\gamma). \tag{B.1.6}$$

For every k, the mapping $(x, t) \mapsto (q, \theta)$ is now well determined and also the regions $M_{\mathrm{s},k}$ of rays that reflect k times are mapped to $M_{\mathrm{t},k}$. We observe that the lines shown if Figure $B.3$ are mapped to vertical lines in target phase space by the map $\mathcal{M}$ (see Figure B.4). Hence, to obtain the boundaries of the target,we will choose rays that are emitted from points close to the boundary of the source. According to what we said so far, the case of the target requires some good calculation to determine where a ray exits the cup. We can obtain those points analytically for a suitable number of rays, as we did in *'Target'*, and then we can drawn those points on the phase space as is shown in Figure $B.4$.
The coordinates of the rays traced in Figure $B.2$ at the target are given by:

$$ADE = \big(-b, -(t_1 + 2\gamma)\big),\ ACE = \big(-b, \sin(\gamma)\big),\ AF = \big(-b, -\sin(\delta)\big),$$
$$BCF = \big(b, -(t_2 - 2\gamma)\big), BDF = \big(b, -\sin(\gamma)\big) \text{ and } BE = \big(b, \sin(\delta)\big).$$

where $t_1 = \arctan(\frac{(}{-}a + b_{-1,x}b_{-1}))$ and $t_2 = \arctan(\frac{(}{a} - b_{-1,x}b_{-1}))$.

Figure $B.3$ and $B.4$ show also the symmetry of the regions $M_{\mathrm{s},k}$ and $M_{\mathrm{t},k}$. Finally we note that, since $k = 1$ is odd, the position of the regions $M_{\mathrm{t},1}$ and $M_{\mathrm{t},-1}$ are exchanged with respect to the position of $M_{\mathrm{s},1}$ and $M_{\mathrm{s},-1}$.

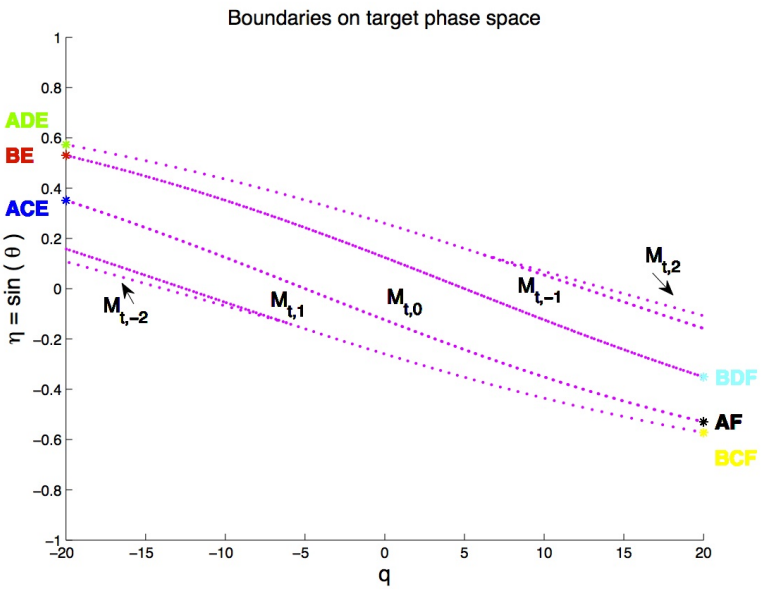


Figure B.4: Regions $M_{t,k}$ of rays that reflect $|k|$ times, for the two-faceted cup. The parameter values are: $a = 5$, $b = 20$ and $h = 40$.

Description of the research

In this thesis we studied the light propagation within optical systems. Optical engineers are interested in design systems in such a way the desired output distribution is obtained. The goal in illumination optics is to obtain the desired output distribution of light. To this purpose the ray tracing procedure is widely used. Ray tracing is a forward method where a set of rays is traced within the system from the source to the target. The propagation of light is determined computing the position and the direction of every ray for all the optical surfaces that it encounters. There are many ways to implement the ray tracing process. Monte Carlo (MC) ray tracing is often used in non-imaging optics. Rays are randomly traced from the source to the target and each time that a ray hits an optical surface the coordinates of the intersection point of the ray with the surface and the new ray direction are calculated. The output variables are computed dividing the target into intervals, the so-called bins, and counting the rays that fall into each bin. To obtain the desired accuracy, millions of rays are required, therefore the method is extremely computationally expensive and it converges as the inverse of the square root of the number of rays traced.

MC ray tracing can be improved using as sample of points a low discrepancy sequence instead of random points. Discrepancy can be interpreted as a measure of how much the sample distribution differs from a uniformly distributed sample. The discrepancy is therefore zero for uniformly distributed points. A low discrepancy sequence gives a sample of points which are regularly distributed but not exactly uniformly distributed. Quasi Monte Carlo (QMC) method considers these kind of sequences as sample of points. Therefore, QMC ray tracing is implemented tracing a set of rays whose position and direction are given by the coordinates of a low discrepancy sequence of points. The main advantage of QMC method is its rate of convergence, it is faster then MC for low dimensional problems. Nevertheless, it has some disadvantages. First, it is not easy to give an error estimation for QMC method. Second, for high dimensional spaces the QMC can become very slow. Third, it is still a binning procedure. Hence, the accuracy depends both on the number of rays traced and on the umber of bins.

In order to improve the existing methods, the phase space (PS) of the optical system is considered in this thesis. The PS of an optical surface gives information about the position and the direction of every ray on that surface where the direction is expressed with respect to the normal of the surface. In PS, the ray's direction is given by the sine of the angle that the ray forms with respect to the normal of the surface multiplied by the index of refraction of the medium in which the ray is located. In two dimensions, the PS is a two-dimensional space where the coordinates of every

ray are specified by one position coordinate and one angular coordinate. For three dimensional systems the PS is a four dimensional space because every ray is specified by two position and two angular coordinates. Our idea is to use the structure of PS to trace only the rays close to the discontinuities of the luminance at the target PS. Two new approaches based on PS are presented in this work. They are tested for two-dimensional systems.

The first method is called ray tracing on PS and it is based on the source and the target PS representation of the optical system. It takes into account the sequence of optical lines that each ray hits when it propagates inside the system, that is the ray path. We note that the source and target phase spaces are partitioned into different regions each of them is formed by the rays that follow the same path. The idea is to use the edge-ray principle proved by Ries and Rabl (1994) which states that the area of these regions is conserved: all rays that are neighbors at the source PS remain close to each other at the target PS. To this purpose, a nonuniform triangulation of the source PS is constructed in such a way that new triangles are added to the triangulation only where boundaries occur. Assuming constant brightness, we only need to compute the boundaries of the regions in target PS to obtain the output photometric variables. We test the method for optical systems where both reflection and refraction laws are involved. Numerical results show that ray tracing on PS is faster and more accurate compared to MC ray tracing.

The second method employs not only the source and the target PS, but also the PS of *all* the other lines that constitute the system. All lines can be modeled as detectors of the incident light and emitters of the reflected light. Moreover, we assume that the source can only emit light and the target can only receive light. Therefore, one PS is taken into account for the source and one for the target. For the other surfaces both the source and target PS are considered. Furthermore, instead of starting from the source, the new method starts tracing back rays from target PS. In order to determine the coordinates of these rays, an inverse map from the target to the source PS is constructed as a concatenation of the maps that relate the PS of two different lines. Employing this map we are able to detect the rays that in target PS are located on the boundaries of the regions with positive luminance. First, we implement the method for systems formed by straight and reflective lines. In this particular case, the boundaries of the regions that form every PS can be computed analytically. This allows us to obtain an analytic target intensity distribution. The results are shown for a two-faceted cup and a multi-faceted cup. In both cases we note significant advantages both in terms of the accuracy and the computational time. Second, the method is developed for systems formed by curved lines. In this case the boundaries cannot be determined analytically and therefore a numerical procedure is involved. In particular, we apply a bisection method on target PS. Also in this case we compare our method to MC ray tracing and we observe significant advantages using the PS method. Finally, the ray mapping method in PS is applied to systems where also Fresnel reflection is taken into account. We obtain relevant results also in the last case.

Curriculum Vitae

Carmela Filosa was born on November 28, 1985 in Torre del greco, Italy. She finished the high school in 2003 at Liceo Scientifico Statale "G. Marconi", Colleferro. She obtained a bachelor (2008) and Master (2013) degree in Mathematics at the University of Rome "La Sapienza", Italy. In March 2014, she moved in Eindhoven (the Netherlands) to start a PhD project at the Eindhoven University of Technology in the department of Mathematics and Computer Science. The PhD project was under the supervision of Wilbert IJzerman and Jan ten Thije Boonkkamp. The research conducted in her doctoral studies was funded by Technologiestichting STW and, the daily work took place at the Centre for Analysis, Scientific computing and Applications (CASA) of TU/e and at the department of Philips Lighting of the High Tech Campus in Eindhoven. The results of her research are presented in this thesis.

Acknowledgments

Bibliography

[1] E. F. Zalewski, "Radiometry and photometry," *Handbook of optics*, vol. 2, pp. 24–1, 1995.

[2] J. Chaves, *Introduction to nonimaging optics*. CRC press, 2015.

[3] "Luminous efficacy-wikipedia the free encyclopedia," *https://commons.wikimedia.org/wiki/File:CIE 1931 Luminosity.png media/File:CIE 1931 Luminosity.png*.

[4] A. V. Arecchi, R. J. Koshel, and T. Messadi, "Field guide to illumination," SPIE, 2007.

[5] H. Zhu and P. Blackborow, "Etendue and optical throughput calculations," *Energetiq Technology, Inc., Woburn, MA*, 2011.

[6] R. P. Feynman, "Feynman lectures on physics. volume 2: Mainly electromagnetism and matter," *Reading, Ma.: Addison-Wesley, 1964, edited by Feynman, Richard P.; Leighton, Robert B.; Sands, Matthew*, 1964.

[7] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[8] E. Hecht, *Optics*. Parson Addison-Wesley, 2002.

[9] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, vol. 1. Basic books, 2011.

[10] P. H. Jones, O. M. Maragò, and G. Volpe, *Optical tweezers: Principles and applications*. Cambridge University Press, 2015.

[11] R. Winston, J. C. Miñano, and P. Benitez, *Nonimaging optics*. Academic Press, 2005.

[12] H. Gross, *Handbook of the Optical Systems*, vol. 1. Wiley-VCH, 2005.

[13] A. B. Owen, "Quasi-monte carlo sampling," *Monte Carlo Ray Tracing: Siggraph*, vol. 1, pp. 69–88, 2003.

[14] C. M. Grinstead and J. L. Snell, *Introduction to probability*. American Mathematical Soc., 2012.

[15] G. Leobacher and F. Pillichshammer, *Introduction to quasi-Monte Carlo integration and applications*. Springer, 2014.

[16] V. M. Zolotarev, *Modern theory of summation of random variables*. Walter de Gruyter, 1997.

[17] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*, vol. 10. John Wiley & Sons, 2016.

[18] D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel, *OpenIntro statistics*. CreateSpace, 2012.

[19] L. Brandolini, L. Colzani, G. Gigante, and G. Travaglini, "A koksma–hlawka inequality for simplices," in *Trends in harmonic analysis*, pp. 33–46, Springer, 2013.

[20] A. B. Owen, "Multidimensional variation for quasi-monte carlo," in *International Conference on Statistics in honour of Professor Kai-Tai FangâĂŹs 65th birthday*, pp. 49–74, 2005.

[21] P. Bratley and B. L. Fox, "Algorithm 659: Implementing sobol's quasirandom sequence generator," *ACM Transactions on Mathematical Software (TOMS)*, vol. 14, no. 1, pp. 88–100, 1988.

[22] K. B. Wolf, *Geometric optics on phase space*. Springer Science & Business Media, 2004.

[23] W. Welford and R. Winston, "On the problem of ideal flux concentrators," *JOSA*, vol. 68, no. 4, pp. 531–534, 1978.

[24] J. C. Miñano and J. C. Gonzalez, "New method of design of nonimaging concentrators," *Applied optics*, vol. 31, no. 16, pp. 3051–3060, 1992.

[25] "Edge ray-principle-wikipedia the free encyclopedia," *https://en.wikipedia.org/wiki/Nonimaging optics Edge ray principle*.

[26] J. C. Miñano, "Design of three-dimensional nonimaging concentrators with inhomogeneous media," *JOSA A*, vol. 3, no. 9, pp. 1345–1353, 1986.

[27] H. Ries and A. Rabl, "Edge-ray principle of nonimaging optics," *JOSA A*, vol. 11, no. 10, pp. 2627–2632, 1994.

[28] B. Guo, J. Menon, and B. Willette, "Surface reconstruction using alpha shapes," in *Computer Graphics Forum*, vol. 16, pp. 177–190, Wiley Online Library, 1997.

[29] "Stracciatella (ice cream)," *https://en.wikipedia.org/wiki/Stracciatella (ice cream)*.

[30] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on information theory*, vol. 29, no. 4, pp. 551–559, 1983.

[31] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics (TOG)*, vol. 13, no. 1, pp. 43–72, 1994.

[32] E. P. Mücke, "Shapes and implementations in three-dimensional geometry," 1993.

[33] E. L. Lloyd, "On triangulations of a set of points in the plane," in *Foundations of Computer Science, 1977., 18th Annual Symposium on*, pp. 228–240, IEEE, 1977.

[34] S. Fortune, "Voronoi diagrams and delaunay triangulations," *Computing in Euclidean geometry*, vol. 1, no. 193-233, p. 2, 1992.

[35] K. Q. Brown, "Voronoi diagrams from convex hulls," *Information Processing Letters*, vol. 9, no. 5, pp. 223–228, 1979.

[36] F. Cazals, J. Giesen, M. Pauly, and A. Zomorodian, "Conformal alpha shapes," in *Point-Based Graphics, 2005. Eurographics/IEEE VGTC Symposium Proceedings*, pp. 55–61, IEEE, 2005.

[37] "Delaunay voronoi.png," *https://commons.wikimedia.org/wiki/File:Delaunay Voronoi.png*.

[38] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, 1980.

[39] R. J. Renka, "Algorithm 772: Stripack: Delaunay triangulation and voronoi diagram on the surface of a sphere," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 3, pp. 416–434, 1997.

[40] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[41] H. Edelsbrunner, "Alpha shapesâĂŤa survey," *Tessellations in the Sciences*, vol. 27, pp. 1–25, 2010.

[42] M. Teichmann and M. Capps, "Surface reconstruction with anisotropic density-scaled alpha shapes," in *Visualization'98. Proceedings*, pp. 67–72, IEEE, 1998.

[43] S. Joe and F. Y. Kuo, "Notes on generating sobol sequences," *ACM Transactions on Mathematical Software (TOMS), 29 (1), 49*, vol. 57, 2008.