Journal of Scientific Computing (2021) 89:27
https://doi.org/10.1007/s10915-021-01612-x

An Energy Conservative *hp*-method for Liouville's Equation of Geometrical Optics

R. A. M. van Gestel[1] · M. J. H. Anthonissen[1] · J. H. M. ten Thije Boonkkamp[1] · W. L. IJzerman[1,2]

Received: 28 May 2020 / Revised: 9 July 2021 / Accepted: 11 July 2021

Abstract
Liouville's equation on phase space in geometrical optics describes the evolution of an energy distribution through an optical system, which is discontinuous across optical interfaces. The discontinuous Galerkin spectral element method is conservative and can achieve higher order of convergence locally, making it a suitable method for this equation. When dealing with optical interfaces in phase space, non-local boundary conditions arise. Besides being a difficulty in itself, these non-local boundary conditions must also satisfy energy conservation constraints. To this end, we introduce an energy conservative treatment of optical interfaces. Numerical experiments are performed to prove that the method obeys energy conservation. Furthermore, the method is compared to the industry standard ray tracing. The numerical experiments show that the discontinuous Galerkin spectral element method outperforms ray tracing by reducing the computation time by up to three orders of magnitude for an error of 10^{-6}.

Keywords Liouville's equation · Geometrical optics · Discontinuous Galerkin · Energy conservation · Phase space

Mathematics Subject Classification 35L65 · 65M70 · 78A05

1 Introduction

Illumination optics deals with the design of optical components for various applications, like LED lighting [28] and automotive headlamps [10,36]. The design of these optical components requires a different approach than used in imaging optics, as imaging effects are highly undesirable [9]. Light propagation through an optical system is usually computed using ray tracing [14]. This means computing the evolution of many rays through an optical

✉ R. A. M. van Gestel
r.a.m.v.gestel@tue.nl

1 Eindhoven University of Technology, PO Box 513, 5600, MB Eindhoven, the Netherlands

2 Signify - High Tech Campus 7, 5656, AE Eindhoven, the Netherlands

Published online: 12 September 2021

system, where upon hitting an optical interface Snell's law of refraction or the law of specular reflection have to be applied. Typical optical interfaces are lenses or mirrors. Ray tracing is commonly employed to directly obtain the illuminance or intensity on a target. Although, forward quasi-Monte Carlo ray tracing converges rather slowly with rates close to $\mathcal{O}\left(N_{RT}^{-1}\right)$ where N_{RT} denotes the number of rays, it is the industry standard. For details on quasi-Monte Carlo integration, see e.g. [24].

The analysis of light propagation using a phase space description provides a new approach to understanding optical systems [17,29,35]. Phase space, being defined as the collection of all positions and direction coordinates of rays, provides a complete description of the spatial and angular distribution of light. A point in phase space evolves according to a Hamiltonian, describing the evolution of one single light ray, whenever the refractive index is smooth. When a ray hits an optical interface the laws for refraction or reflection have to be applied. In [11–13] new ray tracing methods are presented based on the phase space description. These methods allow for tracing of fewer rays to achieve the same accuracy as classical ray tracing.

An alternative approach to ray tracing is based on directly obtaining an energy distribution on phase space, rather than its integrated quantities such as the illuminance or luminous intensity. The propagation of an energy distribution, related to the luminance, through an optical system is governed by Liouville's equation for geometrical optics on phase space. Recently, numerical schemes for Liouville's equation were developed that incorporate the optical interfaces. In [33] van Lith et al. derived a first-order upwind finite difference scheme and in [32] van Lith et al. introduced a third-order active flux finite volume scheme on moving meshes. The third-order active flux scheme was proved to be faster and more accurate compared to classical ray tracing for obtaining an energy distribution on phase space. Additionally in [31] van Lith made use of the discontinuous Galerkin spectral element method to solve Liouville's equation.

The discontinuous Galerkin spectral element method (DGSEM) discussed by Kopriva in [21] is a collocation scheme for the semi-discretisation of the spatial domain for conservation laws, leaving the time-like variable continuous. In terms of Liouville's equation this entails discretising phase space. The phase space domain is partitioned into elements with each element having interior nodes placed at collocation points. The solution is approximated using a polynomial, where the polynomial degree determines the number of interior nodes used for each element. Consequently the method has an extraordinary flexibility as it is an hp-method, where h refers to the mesh size and p to the polynomial degree, i.e. the accuracy can be increased by decreasing the mesh size or by increasing the polynomial degree. Additionally, the method does not enforce continuity across the boundary of each element, making it particularly suitable for the discontinuous solutions across optical interfaces.

At an optical interface Liouville's equation is not valid. Instead, Snell's law of refraction or the law of specular reflection describe the discontinuous change in the direction coordinate, i.e., a jump in phase space. In what follows, we will see that this results in non-local boundary conditions for the energy distribution in phase space. Our contribution consists of describing the treatment of these optical interfaces so that they obey energy conservation. In the DGSEM the elements communicate using numerical fluxes. Snell's law and the law of specular reflection are incorporated in these numerical fluxes at an optical interface. In addition to the discontinuous change in the direction coordinate described by these laws, a single element before the optical interface might contribute to multiple elements after the optical interface. This connection to multiple elements is similar to fully non-conforming geometries when using subdomain refinement [3,4]. Kopriva et al. outlined such a strategy for the DGSEM in [23]. In [5] an analysis of this method is presented by Bui-Thanh and

Ghattas. Across an optical interface the numerical fluxes are discontinuous and therefore we have to take a different approach. Inspired by [23], we present a method that directly incorporates the laws of optics and obeys energy conservation.

The article is outlined as follows: in Sect. 2 we discuss the conserved quantities in an optical system and Liouville's equation, and in Sect. 3 we discuss the DGSEM. In Sect. 4 we discuss the energy conservative treatment of the optical interfaces, and in Sect. 5 we present numerical experiments proving energy conservation for two examples. The first example features a smooth refractive index field, while the second example is a test case featuring a discontinuity in the refractive index described by van Lith et al. in [33]. In the latter example, we compare the DGSEM for solving Liouville's equation to quasi-Monte Carlo ray tracing for obtaining the illuminance. Finally we present our conclusions in Sect. 6.

2 Conserved Quantities and Liouville's Equation

In optics we consider the transfer of luminous flux between surfaces. A source emits a beam of radiation or light, carrying a finite amount of luminous flux denoted by Φ. In the absence of losses by absorption or scattering in an optical system, the total flux Φ is conserved, i.e., energy throughout the optical system is conserved. A related quantity is the luminance denoted by ρ^*, which is defined as [9,25]

$$\rho^* := \frac{\mathrm{d}\Phi}{\mathrm{d}A\cos\theta\,\mathrm{d}\omega}, \tag{1}$$

where $\mathrm{d}\Phi$ is an infinitesimal amount of flux carried by an infinitesimal beam, $\mathrm{d}\omega$ is an element of solid angle subtended at the center of the source by the area at the detector $\mathrm{d}A$, and $\mathrm{d}A\cos\theta$ the projected area perpendicular to the beam, i.e., $\mathrm{d}A$ is an element of the surface area of the detector and θ describes the angle between the normal of the detector and the beam. The solid angle describes a cone on the unit sphere with the center of the source as its vertex and $\mathrm{d}\omega$ the area on the unit sphere subtended by the cone.

Another important quantity that is also conserved in an optical system is étendue, which is defined by [9]

$$\mathrm{d}\mathcal{U} := n^2\mathrm{d}A\cos\theta\,\mathrm{d}\omega, \tag{2}$$

where n is the refractive index in which the beam is immersed. This allows us to write the luminance as

$$\rho^* = n^2\frac{\mathrm{d}\Phi}{\mathrm{d}\mathcal{U}}. \tag{3}$$

When a beam is propagating through a homogeneous medium the luminance ρ^* is conserved, as is implied by conservation of energy and conservation of étendue. When a beam of light strikes an optical interface, e.g., a lens or a mirror, the beam is subject to Snell's law of refraction or the law of specular reflection. In the case where the beam is refracted, e.g., a transition from a medium with refractive index n_1 to a medium with refractive index n_2, the luminance is not conserved. Assuming no Fresnel reflections, and applying conservation of energy and étendue, we obtain [25,26]

$$\frac{\rho_\mathrm{i}^*}{n_1^2} = \frac{\rho_\mathrm{t}^*}{n_2^2}, \tag{4}$$

where ρ_i^* and ρ_t^* describe the incident and transmitted luminance, respectively. The quantity ρ^*/n^2, known as basic luminance is conserved for refractions, cf. (4). A similar result can be

derived for reflections, where the refractive indices are equal. Relation (4) will be referred to as basic luminance invariance. For a complete derivation including Fresnel reflections see [9,25,26].

The definitions of luminance, étendue and basic luminance invariance described above hold for three-dimensional optics, whereas in two-dimensional optics the definitions are slightly altered. For more details, see [9]. In summary, we denote the basic luminance for both two- and three-dimensional optics by ρ, which is defined by

$$\rho := \frac{\mathrm{d}\Phi}{\mathrm{d}\mathcal{U}}, \tag{5}$$

where the étendue $\mathrm{d}\mathcal{U}$ for two- and three-dimensional systems reads [9]

$$\mathrm{d}\mathcal{U} = \begin{cases} n^2 \mathrm{d}A \cos\theta \, \mathrm{d}\omega & \text{for 3D optics,} \\ n \, \mathrm{d}l \cos\theta \, \mathrm{d}\theta & \text{for 2D optics,} \end{cases} \tag{6}$$

where for 2D optics θ denotes the angle between the normal of the detector and the beam, and $\mathrm{d}\theta$ an element of angle subtended at the center of the source by the infinitesimal line segment at the detector $\mathrm{d}l$. The basic luminance is related to the luminance by $\rho = \rho^*/n^2$ for three-dimensional optics, whereas for two-dimensional optics $\rho = \rho^*/n$.

2.1 Liouville's Equation

In geometrical optics the evolution of light rays in a beam of light can be cast in a Hamiltonian system, where we denote with $\boldsymbol{q} \in \mathbb{R}^d$ the position and $\boldsymbol{p} \in \mathbb{R}^d$ the momentum coordinates [35]. For two-dimensional optics $d = 1$, while for three-dimensional optics $d = 2$. Both terms together describe a point $(\boldsymbol{q}, \boldsymbol{p})$ in phase space, where the phase space $\mathcal{P}$ is defined as the collection of all positions $\boldsymbol{q}$ and momenta $\boldsymbol{p}$ at a certain position along the optical axis denoted by the z-coordinate. A point in phase space evolves when we move along the optical axis.

The momentum $\vec{p} = (\boldsymbol{p}, p_z) \in \mathbb{R}^3$ is restricted to Descartes' sphere $|\vec{p}| = n(z, \boldsymbol{q})$ where n is the refractive index field as a function of the three-dimensional position coordinates $\vec{q} = (\boldsymbol{q}, z)$ [35]. This restriction invites us to use spherical coordinates to represent the momentum vector $\vec{\mathbf{p}}$ as

$$\vec{p} = (\boldsymbol{p}, p_z) = n \, (\sin\theta \cos\varphi, \sin\theta \sin\varphi, \cos\theta) \, , \tag{7}$$

where θ represents the polar angle, describing the angle between the z-axis and $\vec{p}$ measured from the z-axis, and φ the azimuthal angle for describing the direction in the $\boldsymbol{q}$-plane. Therefore, at a given position z_0 along the optical axis, one can visualise the phase space coordinates on the screen that is perpendicular to the z-axis and intersects the z-axis at z_0, where $\boldsymbol{q}$ is the position on the screen and $\boldsymbol{p}$ describes the projection of $\vec{p}$ on the screen [32]. The restriction of the momentum for physical rays $\vec{p}$ to Descartes' sphere also implies that the two-dimensional momentum vector is restricted by $|\boldsymbol{p}| \leq n$, describing a region known as Descartes' disc [35].

The phase space coordinates of a light ray evolve as a function of the distance along the optical axis according to Hamilton's equations, which read

$$\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}z} = \frac{\partial h}{\partial \boldsymbol{p}}, \tag{8a}$$

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}z} = -\frac{\partial h}{\partial \boldsymbol{q}}, \tag{8b}$$

with $h = h(z, \boldsymbol{q}, \boldsymbol{p})$ the optical Hamiltonian given by

$$h(z, \boldsymbol{q}, \boldsymbol{p}) = -\sigma\sqrt{n(z, \boldsymbol{q})^2 - |\boldsymbol{p}|^2}. \tag{9}$$

Here, $\sigma \in \{-1, 0, +1\}$ denotes the direction of the light ray travelling along the optical axis, with $\sigma = 0$ being marginal rays that travel perpendicular to the optical axis [35]. For simplicity, we assume that all rays travel in the positive z-direction given by $\sigma = +1$.

Hamilton's Eqs. (8) hold for a single light ray, however, this may be generalised to a beam of light carrying a finite amount of energy in terms of luminous flux. The flow generated by Hamilton's equations describes canonical transformations, otherwise known as symplectic transformations, on phase space. These transformations preserve the symplectic structure of phase space [1]. In other words the phase space volume element $\mathrm{d}q_1\mathrm{d}q_2\mathrm{d}p_1\mathrm{d}p_2$ is constant. In the context of optics this has the equivalent meaning of étendue conservation. In fact $\mathrm{d}\mathcal{U} = \mathrm{d}q_1\mathrm{d}q_2\mathrm{d}p_1\mathrm{d}p_2$, which can be obtained from the first two components of the three-dimensional momentum vector described by expression (7). The Jacobian determinant of $\boldsymbol{p}$ with respect to the polar and azimuthal angles θ and φ, can be computed as

$$\mathrm{d}p_1\mathrm{d}p_2 = \det\left(\frac{\partial\,(p_1, p_2)}{\partial(\theta, \varphi)}\right)\mathrm{d}\theta\mathrm{d}\varphi = n^2\cos\theta\mathrm{d}\theta\,\sin\theta\mathrm{d}\varphi = n^2\cos\theta\mathrm{d}\omega, \tag{10}$$

with $\mathrm{d}\omega = \sin\theta\mathrm{d}\theta\mathrm{d}\varphi$ an element of solid angle. Noting that the differential area on a screen can be written as $\mathrm{d}A = \mathrm{d}q_1\mathrm{d}q_2$ and substituting (10) into relation (6) for 3D optics, we obtain

$$\mathrm{d}\mathcal{U} = \mathrm{d}q_1\mathrm{d}q_2\mathrm{d}p_1\mathrm{d}p_2. \tag{11}$$

The basic luminance invariance (4) implies that ρ remains constant if we move an arbitrary distance Δz along the optical axis, i.e.,

$$\rho\,(z + \Delta z, \boldsymbol{q}(z + \Delta z), \boldsymbol{p}(z + \Delta z)) = \rho(z, \boldsymbol{q}(z), \boldsymbol{p}(z)). \tag{12}$$

Note that this relation also holds when a beam of light is reflected or refracted. If the solution is sufficiently smooth, one can derive Liouville's equation by subtracting the right-hand side of (12) from its left-hand side and dividing by Δz and subsequently taking the limit $\Delta z \to 0$, resulting in

$$\frac{\partial\rho}{\partial z} + \frac{\partial h}{\partial \boldsymbol{p}}\cdot\frac{\partial\rho}{\partial \boldsymbol{q}} - \frac{\partial h}{\partial \boldsymbol{q}}\cdot\frac{\partial\rho}{\partial \boldsymbol{p}} = 0. \tag{13}$$

Here, we have already applied Hamilton's equations (8). The advective form of Liouville's equation (13) may be written in conservative form by assuming that h is twice differentiable, upon which we obtain

$$\frac{\partial\rho}{\partial z} + \nabla\cdot\boldsymbol{f} = 0, \tag{14a}$$

with $\nabla = (\frac{\partial}{\partial \boldsymbol{q}}, \frac{\partial}{\partial \boldsymbol{p}})^{\mathrm{T}}$ and the flux vector $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{p})$ defined as

$$\boldsymbol{f} := \rho\boldsymbol{u} = \rho\begin{pmatrix}\frac{\partial h}{\partial \boldsymbol{p}} \\ -\frac{\partial h}{\partial \boldsymbol{q}}\end{pmatrix}, \tag{14b}$$

where we have used that the velocity field $\boldsymbol{u}$ is divergence-free, and the superscript T denotes transpose. Note that an optical interface causes the Hamiltonian to be discontinuous. Therefore, at an optical interface, Liouville's equation does not hold and we must apply (12) together with Snell's law and/or the law of specular reflection, in the limit $\Delta z \to 0$.

Solving Liouville's equation on phase space at any point z along the optical axis tells us how the basic luminance changes when light propagates through an optical system, allowing us to compute at each z-coordinate the related integral quantities such as luminous flux on the screen. The total luminous flux Φ in the optical system at $z = \text{const}$ reads

$$\Phi(z) = \int_{\mathcal{P}(z)} \rho(z, \boldsymbol{q}, \boldsymbol{p}) \, \mathrm{d}\mathcal{U}. \tag{15}$$

Here, the phase space dependence on the z-coordinate is denoted explicitly, since the momentum domain is restricted according to Descartes' disc. Assuming the optical system is lossless the total luminous flux should be constant, i.e., $\Phi(z) = \Phi(0)$.

An infinitesimal element of illuminance E is defined by

$$\mathrm{d}E := \frac{\mathrm{d}\Phi}{\mathrm{d}A}. \tag{16}$$

Applying definition (5) for the basic luminance and relation (11), $\mathrm{d}E$ (16) can be rewritten as

$$\mathrm{d}E = \rho \, \mathrm{d}p_1 \mathrm{d}p_2.$$

Next, integrating over momentum space results in the illuminance $E(z, \boldsymbol{q})$, i.e.,

$$E(z, \boldsymbol{q}) = \int_{P(z)} \rho(z, \boldsymbol{q}, \boldsymbol{p}) \, \mathrm{d}p_1 \mathrm{d}p_2, \tag{17}$$

where $\boldsymbol{p} = (p_1, p_2) \in P(z)$ in which $P(z)$ denotes the two-dimensional momentum space at a certain position z along the optical axis. Alternatively, an infinitesimal element of luminous intensity I is defined by

$$\mathrm{d}I := \frac{\mathrm{d}\Phi}{\mathrm{d}\omega}. \tag{18}$$

Applying again definition (5) for the basic luminance and definition (6) for the étendue, $\mathrm{d}I$ (18) can be written as

$$\mathrm{d}I = \rho \, n^2 \cos\theta \, \mathrm{d}A.$$

Subsequently using the relation for p_z defined in (7) and $\mathrm{d}A = \mathrm{d}q_1 \mathrm{d}q_2$, followed by integration over the position coordinates on the screen, denoted by $\boldsymbol{q} = (q_1, q_2) \in Q(z)$, we obtain

$$I(z, \boldsymbol{p}) = \int_{Q(z)} \rho(z, \boldsymbol{q}, \boldsymbol{p}) p_z(z, \boldsymbol{q}, \boldsymbol{p}) n(z, \boldsymbol{q}) \, \mathrm{d}q_1 \mathrm{d}q_2. \tag{19}$$

With these definitions, the main quantities of interest in optics can thus be easily computed from the basic luminance, satisfying Liouville's equation. In the next section, we explore a method for solving Liouville's equation.

3 Derivation of DGSEM

In what follows, we restrict ourselves to two-dimensional optical systems, hence reducing the complexity from a four-dimensional phase space to a two-dimensional phase space with position coordinate q and momentum coordinate p, and the distance along the optical axis denoted by z. Note that the position and momentum are now scalar quantities, therefore we omit the bold-face notation for these quantities. Next, we outline a spatial semi-discretisation of Liouville's equation, leaving only z continuous. For the semi-discretisation we apply the

discontinuous Galerkin spectral element method (DGSEM) described by Kopriva in [21], to the two-dimensional Liouville equation for $\rho = \rho(z, q, p)$ in conservative form

$$\frac{\partial \rho}{\partial z} + \nabla \cdot \boldsymbol{f} = 0, \tag{20a}$$

where $\nabla = \left(\frac{\partial}{\partial q}, \frac{\partial}{\partial p}\right)^{\mathrm{T}}$ and the flux vector $\boldsymbol{f}$ now reads

$$\boldsymbol{f} = \rho \boldsymbol{u} = \rho \begin{pmatrix} \frac{\partial h}{\partial p} \\ -\frac{\partial h}{\partial q} \end{pmatrix}. \tag{20b}$$

The Hamiltonian h for two-dimensional optics reduces to

$$h(z, q, p) = -\sqrt{n(z, q)^2 - p^2}, \tag{21}$$

and consequently the velocity $\boldsymbol{u}$ reads

$$\boldsymbol{u} = \frac{1}{\sqrt{n^2 - p^2}} \begin{pmatrix} p \\ n \frac{\partial n}{\partial q} \end{pmatrix}. \tag{22}$$

For phase space discretisation, the two-dimensional phase space domain $\mathcal{P}$ is covered with straight-sided quadrilaterals $\Omega^k \subset \mathcal{P}$ with k the index of the element. In a more general discretisation, the boundaries of quadrilaterals are allowed to be curved, such that curved boundaries from physical constraints can be modelled appropriately. In fact, when the refractive index field changes continuously as a function of $\boldsymbol{q}$, then the maximum allowed momentum varies as a function of q due to the restriction of $\boldsymbol{p}$ to Descartes' sphere. This restriction can be accommodated by curved boundaries when solving Liouville's equation, see [31]. For a discussion on DGSEM with curved quadrilateral elements, see for example [8,18,21,22]. In this paper we only consider straight-sided quadrilaterals.

Each quadrilateral Ω^k has four vertices $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$ labelled in counter-clockwise direction where $\boldsymbol{x} = (q, p)^{\mathrm{T}}$ and we have omitted the element index (superscript k), see Fig. 1. For ease of computation, the reference square $\chi = [-1, 1]^2$ is mapped to each quadrilateral Ω^k, transforming a point in the reference domain $(\xi, \eta) \in \chi$ to a point in physical space $\boldsymbol{x}(\xi, \eta) \in \mathcal{P}$ using the following bilinear transformation

$$\begin{aligned} \boldsymbol{x}(\xi, \eta) = \frac{1}{4} [&(1 - \xi)(1 - \eta)\boldsymbol{x}_1 + (1 + \xi)(1 - \eta)\boldsymbol{x}_2 \\ &+ (1 + \xi)(1 + \eta)\boldsymbol{x}_3 + (1 - \xi)(1 + \eta)\boldsymbol{x}_4]. \end{aligned} \tag{23}$$

The Jacobian of the transformation is given by $\frac{\partial(q,p)}{\partial(\xi,\eta)} = \left(\frac{\partial \boldsymbol{x}}{\partial \xi}, \frac{\partial \boldsymbol{x}}{\partial \eta}\right)$, where the columns read

$$\frac{\partial \boldsymbol{x}}{\partial \xi} = \begin{pmatrix} \frac{\partial q}{\partial \xi} \\ \frac{\partial p}{\partial \xi} \end{pmatrix} = \frac{1}{4} [(1 - \eta)(\boldsymbol{x}_2 - \boldsymbol{x}_1) + (1 + \eta)(\boldsymbol{x}_3 - \boldsymbol{x}_4)], \tag{24a}$$

$$\frac{\partial \boldsymbol{x}}{\partial \eta} = \begin{pmatrix} \frac{\partial q}{\partial \eta} \\ \frac{\partial p}{\partial \eta} \end{pmatrix} = \frac{1}{4} [(1 - \xi)(\boldsymbol{x}_4 - \boldsymbol{x}_1) + (1 + \xi)(\boldsymbol{x}_3 - \boldsymbol{x}_2)]. \tag{24b}$$

The divergence term in (20a) can be rewritten by applying the chain rule resulting in

$$\nabla \cdot \boldsymbol{f} = \frac{1}{\mathcal{J}} \nabla_{\boldsymbol{\xi}} \cdot \widetilde{\boldsymbol{f}}, \tag{25}$$

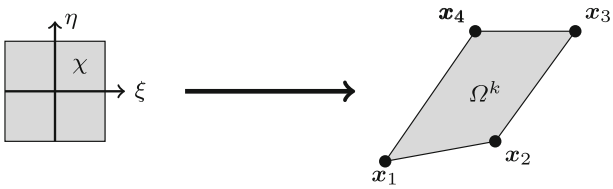


Fig. 1 Mapping from reference square χ to a quadrilateral Ω^k

where $\mathcal{J} = \frac{\partial q}{\partial \xi}\frac{\partial p}{\partial \eta} - \frac{\partial q}{\partial \eta}\frac{\partial p}{\partial \xi}$ denotes the Jacobian determinant, $\nabla_{\boldsymbol{\xi}} = \left(\frac{\partial}{\partial \xi}, \frac{\partial}{\partial \eta}\right)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{f}}$ is an auxiliary flux defined by the product of the adjoint Jacobian matrix and the flux $\boldsymbol{f}$, i.e.,

$$\widetilde{\boldsymbol{f}} := \begin{pmatrix} \frac{\partial p}{\partial \eta} & -\frac{\partial q}{\partial \eta} \\ -\frac{\partial p}{\partial \xi} & \frac{\partial q}{\partial \xi} \end{pmatrix} \boldsymbol{f}. \tag{26}$$

Applying the transformation (25) to Liouville's Eq. (20a), we obtain

$$\frac{\partial \rho}{\partial z} + \frac{1}{\mathcal{J}} \nabla_{\boldsymbol{\xi}} \cdot \widetilde{\boldsymbol{f}} = 0, \tag{27}$$

where $\rho = \rho(z, \xi, \eta)$.

The weak formulation of Liouville's equation is obtained by first multiplying the PDE (27) by the Jacobian determinant $\mathcal{J}$ and by a smooth test function ϕ, and subsequently integrating over the reference domain χ. This results in

$$\int_{\chi} \phi \mathcal{J} \frac{\partial \rho}{\partial z} \, \mathrm{d}A_{\boldsymbol{\xi}} + \int_{\chi} \phi \nabla_{\boldsymbol{\xi}} \cdot \widetilde{\boldsymbol{f}} \, \mathrm{d}A_{\boldsymbol{\xi}} = 0. \tag{28}$$

The second term is rewritten by applying the product rule and Gauss's theorem, so that

$$\begin{aligned} \int_{\chi} \phi \nabla_{\boldsymbol{\xi}} \cdot \widetilde{\boldsymbol{f}} \, \mathrm{d}A_{\boldsymbol{\xi}} &= \int_{\chi} \left(\nabla_{\boldsymbol{\xi}} \cdot \left(\phi \widetilde{\boldsymbol{f}} \right) - \left(\nabla_{\boldsymbol{\xi}} \phi \right) \cdot \widetilde{\boldsymbol{f}} \right) \mathrm{d}A_{\boldsymbol{\xi}} \\ &= \oint_{\partial \chi} \phi \widetilde{\boldsymbol{f}} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}\sigma - \int_{\chi} \left(\nabla_{\boldsymbol{\xi}} \phi \right) \cdot \widetilde{\boldsymbol{f}} \, \mathrm{d}A_{\boldsymbol{\xi}}, \end{aligned}$$

where $\hat{\boldsymbol{n}}$ is the outward unit normal on $\partial \chi$ and the orientation of the closed curve $\partial \chi$ is counter-clockwise. Using this, we obtain the weak formulation of Liouville's equation on the reference domain

$$\int_{\chi} \phi \mathcal{J} \frac{\partial \rho}{\partial z} \, \mathrm{d}A_{\boldsymbol{\xi}} + \oint_{\partial \chi} \phi \widetilde{\boldsymbol{f}} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}\sigma - \int_{\chi} \left(\nabla_{\boldsymbol{\xi}} \phi \right) \cdot \widetilde{\boldsymbol{f}} \, \mathrm{d}A_{\boldsymbol{\xi}} = 0. \tag{29}$$

Note that for strong solutions we require the flux to be differentiable, hence, $h(z, q, p)$ should be twice differentiable. However, the DGSEM uses the weak form of the solution and only requires the flux to be continuous, therefore, $h(z, q, p)$ being once continuously differentiable is sufficient. For typical optical interfaces this is not sufficient since the refractive index field is discontinuous and, therefore, $h(z, q, p)$ and also the flux are discontinuous. In particular, for these interfaces we require a special treatment of the fluxes which we will discuss in Sect. 4.

3.1 Tools for Approximating the Solution

The solution ρ in Eq. (29) is approximated by an expansion in basis functions [21]. We choose one-dimensional basis-functions φ_i, $i = 0, ..., N$, for which $\varphi_i(\xi_j) = \delta_{ij}$ holds for chosen points ξ_j. Moreover, we require that the basis-functions form an orthogonal basis with respect to the standard L^2-inner product. A suitable choice are the Lagrange polynomials defined on Gauss-Legendre nodes. In the following, we will replace ρ by the approximation

$$\rho(z, \xi, \eta) \approx \rho^{\mathrm{h}}(z, \xi, \eta) = \sum_{i,j=0}^{N} \rho_{ij}(z)\varphi_i(\xi)\varphi_j(\eta), \tag{30}$$

where $\rho_{ij}(z)$ are the expansion coefficients for the chosen basis. In this paper we will restrict ourselves to using the same basis functions in both directions with an equal number of expansion coefficients, although in general this restriction is not necessary.

The quadrature rule defined by Gauss-Legendre nodes $\{\xi_i\}_{i=0}^N$ and corresponding weights $\{w_i\}_{i=0}^N$ allows us to approximate the integral of any function g, i.e.,

$$\int_{-1}^{1} g(\xi)\,\mathrm{d}\xi \approx \sum_{i=0}^{N} g\,(\xi_i)\,w_i, \tag{31}$$

with $-1 < \xi_i < 1$ and $w_i > 0$. The quadrature rule on the reference domain has a tensor product structure, hence,

$$\int_{-1}^{1}\int_{-1}^{1} g(\xi, \eta)\,\mathrm{d}\xi\mathrm{d}\eta \approx \sum_{i,j=0}^{N} g(\xi_i, \eta_j)w_i w_j. \tag{32}$$

Thus, we place nodes inside the reference domain at the Gauss-Legendre nodes (ξ_i, η_j). Furthermore, for one-dimensional integrals the Gauss-Legendre quadrature gives exact integration for at least all polynomials of degree $2N + 1$.

Focusing on one dimension, the Lagrange polynomials on the Gauss-Legendre nodes read

$$\ell_i(\xi) = \prod_{\substack{j=0 \\ j\neq i}}^{N} \frac{\xi - \xi_j}{\xi_i - \xi_j}, \tag{33}$$

which satisfy the Kronecker property, i.e.,

$$\ell_i(\xi_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{34}$$

It can be readily verified, that the Lagrange polynomials defined on the Gauss-Legendre nodes are orthogonal with respect to the standard L^2-inner product, i.e.,

$$\int_{-1}^{1} \ell_i(\xi)\ell_j(\xi)\,\mathrm{d}\xi = \sum_{n=0}^{N} \ell_i(\xi_n)\ell_j(\xi_n)w_n = \delta_{ij}w_i, \tag{35}$$

where the integration is exact since $\ell_i(\xi)\ell_j(\xi)$ is a polynomial of degree at most $2N$.

A very useful property of Lagrange polynomials is that the polynomial interpolation of any function g is rather easy, i.e.,

$$g_N(\xi) = (\mathcal{I}_N g)(\xi) = \sum_{j=0}^{N} \ell_j(\xi) g(\xi_j), \tag{36}$$

where $\mathcal{I}_N$ is the polynomial interpolation operator using $N+1$ nodes. The approximation of the derivative of g is defined as the derivative of the interpolant, i.e.,

$$\frac{\mathrm{d}g_N}{\mathrm{d}\xi}(\xi) = \sum_{j=0}^{N} \frac{\mathrm{d}\ell_j}{\mathrm{d}\xi}(\xi) g(\xi_j). \tag{37}$$

Note that if g is a polynomial of degree N or less, both the interpolant and the derivative are exact. In what follows, we require the derivative at the nodes, i.e.,

$$\frac{\mathrm{d}g_N}{\mathrm{d}\xi}(\xi_i) = \sum_{j=0}^{N} \frac{\mathrm{d}\ell_j}{\mathrm{d}\xi}(\xi_i) g(\xi_j) = \sum_{j=0}^{N} D_{ij} g(\xi_j), \tag{38}$$

where $D_{ij} = \frac{\mathrm{d}\ell_j}{\mathrm{d}\xi}(\xi_i)$. The elements of the differentiation matrix $\boldsymbol{D} = (D_{ij})$ can be found by differentiating the Lagrange polynomial followed by evaluation at the node, and thus read

$$D_{ij} = \sum_{\substack{n=0 \\ n\neq j}}^{N} \frac{1}{\xi_j - \xi_n} \prod_{\substack{m=0 \\ m\neq n, m\neq j}}^{N} \frac{\xi_i - \xi_m}{\xi_j - \xi_m} \quad \text{for } i \neq j, \tag{39a}$$

and the diagonal elements read

$$D_{ii} = -\sum_{\substack{n=0 \\ n\neq i}}^{N} D_{in}, \tag{39b}$$

due to the fact that the derivative of a constant function vanishes.

3.2 Approximating the Solution

To derive an approximation of the solution, we expand both the solution and the flux in Lagrange polynomials. The expansions read

$$\rho(z, \xi, \eta) \approx \rho^{\mathrm{h}}(z, \xi, \eta) = \sum_{i,j=0}^{N} \rho_{ij}(z) \ell_i(\xi) \ell_j(\eta), \tag{40a}$$

$$\widetilde{\boldsymbol{f}}(z, \xi, \eta) \approx \widetilde{\boldsymbol{f}}^{\mathrm{h}}(z, \xi, \eta) = \sum_{i,j=0}^{N} \widetilde{\boldsymbol{f}}_{ij}(z) \ell_i(\xi) \ell_j(\eta). \tag{40b}$$

The coefficients, indicated by the index-subscript ij, in each expansion are related to the position of an element's interior node (q_{ij}, p_{ij}), by $\rho_{ij}(z) = \rho(z, q_{ij}, p_{ij})$ and $\widetilde{\boldsymbol{f}}_{ij}(z) = \widetilde{\boldsymbol{f}}(z, q_{ij}, p_{ij})$. The auxiliary flux coefficients $\widetilde{\boldsymbol{f}}_{ij}(z)$ are related to ρ by

$$\widetilde{\boldsymbol{f}}_{ij}(z) := \widetilde{\boldsymbol{u}}_{ij}(z) \rho_{ij}(z), \tag{41}$$

with $\widetilde{\boldsymbol{u}}_{ij}$ the transformed velocity, similarly defined to (26). Here the velocity $\widetilde{\boldsymbol{u}}_{ij}(z) = \widetilde{\boldsymbol{u}}(z, q_{ij}, p_{ij})$ depends on z if the refractive index n depends on z. In the following, we omit $\widetilde{\boldsymbol{f}}_{ij}$'s dependence on z for ease of notation.

Next, we have to approximate the integrals in Eq. (29). The test function ϕ is chosen to be in the same basis as the solution ρ, resulting in a Galerkin method. Therefore, taking

$$\phi(\xi, \eta) = \ell_i(\xi)\ell_j(\eta), \tag{42}$$

allows us to derive $(N+1)^2$ equations for the $(N+1)^2$ coefficients ρ_{ij}. Combining this together with the approximations (40a) and (40b) for ρ and $\widetilde{\boldsymbol{f}}$ we can approximate the integrals using the Gauss-Legendre quadrature rules. Therefore, substituting the approximation (40a) in the first term of (29), we obtain

$$\begin{aligned}\int_\chi \phi \mathcal{J} \frac{\partial \rho^{\mathrm{h}}}{\partial z}\,\mathrm{d}A_{\boldsymbol{\xi}} &= \int_\chi \ell_i(\xi)\ell_j(\eta)\mathcal{J}(\xi,\eta)\left(\sum_{k,l=0}^{N}\frac{\mathrm{d}\rho_{kl}(z)}{\mathrm{d}z}\ell_k(\xi)\ell_l(\eta)\right)\mathrm{d}A_{\boldsymbol{\xi}} \\ &= \sum_{n,m=0}^{N} w_n w_m \ell_i(\xi_n)\ell_j(\eta_m)\mathcal{J}(\xi_n,\eta_m)\left(\sum_{k,l=0}^{N}\frac{\mathrm{d}\rho_{kl}(z)}{\mathrm{d}z}\ell_k(\xi_n)\ell_l(\eta_m)\right).\end{aligned}$$

Applying the Kronecker property (34) of the Lagrange polynomials, the sums reduce to

$$\int_\chi \phi \mathcal{J} \frac{\partial \rho^{\mathrm{h}}}{\partial z}\,\mathrm{d}A_{\boldsymbol{\xi}} = w_i w_j \mathcal{J}_{ij}\frac{\mathrm{d}\rho_{ij}(z)}{\mathrm{d}z}, \tag{43}$$

where $\mathcal{J}_{ij} := \mathcal{J}(\xi_i, \eta_j)$. Note that the integral is exact for the given combination of a bilinear mapping $\boldsymbol{x}(\xi, \eta)$ and Lagrangian polynomials, since then the integrand is a polynomial of degree $2N+1$ in ξ and in η. The Gauss-Legendre quadrature rule is exact for this bivariate polynomial.

For the third term in (29), we substitute the approximation (40b) and denote $\widetilde{\boldsymbol{f}} = (\widetilde{f}, \widetilde{g})$, resulting in

$$\begin{aligned}\int_\chi \left(\nabla_{\boldsymbol{\xi}}\phi\right)\cdot\widetilde{\boldsymbol{f}}^{\mathrm{h}}\,\mathrm{d}A_{\boldsymbol{\xi}} &= \int_\chi \left(\ell_i'(\xi)\ell_j(\eta)\widetilde{f}(\xi,\eta) + \ell_i(\xi)\ell_j'(\eta)\widetilde{g}(\xi,\eta)\right)\mathrm{d}A_{\boldsymbol{\xi}} \\ &= \sum_{n,m=0}^{N} w_n w_m \left(\ell_i'(\xi_n)\ell_j(\eta_m)\widetilde{f}(\xi_n,\eta_m) + \ell_i(\xi_n)\ell_j'(\eta_m)\widetilde{g}(\xi_n,\eta_m)\right) \\ &= w_j \sum_{n=0}^{N} w_n D_{ni}\widetilde{f}_{nj} + w_i \sum_{m=0}^{N} w_m D_{mj}\widetilde{g}_{im},\end{aligned}$$

where we have used the definition of the differentiation matrix (39). Furthermore, we introduce the following auxiliary matrix

$$\widehat{D}_{ij} := D_{ji}\frac{w_j}{w_i}, \tag{44}$$

for ease of computation. The third term then reads

$$\int_\chi \left(\nabla_{\boldsymbol{\xi}}\phi\right)\cdot\widetilde{\boldsymbol{f}}^{\mathrm{h}}\,\mathrm{d}A_{\boldsymbol{\xi}} = w_i w_j\left(\sum_{n=0}^{N}\widehat{D}_{in}\widetilde{f}_{nj} + \sum_{m=0}^{N}\widehat{D}_{jm}\widetilde{g}_{im}\right). \tag{45}$$

In what follows, we will replace the flux appearing in the boundary integral from Eq. (29) with a numerical flux $\widetilde{\boldsymbol{F}} = \left(\widetilde{F}, \widetilde{G}\right)$. The boundary integral can be split into four parts and

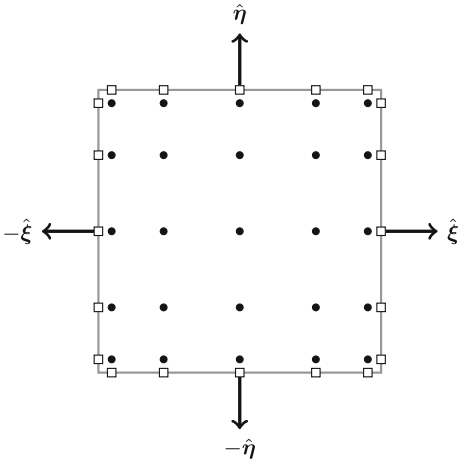


Fig. 2 Reference square with polynomial degree $N = 4$. The normals are denoted by arrows, interior nodes are denoted by filled circles and boundary points by open squares

evaluated for each boundary segment, see Fig. 2. Along each segment the numerical flux $\widetilde{\boldsymbol{F}}$ is described by an Nth degree polynomial at the boundary nodes shown in the figure. For the bottom part, with $\eta = -1$, the integral can be exactly evaluated using Gauss-Legendre quadrature, such that we obtain

$$\int_{-1}^{1} \ell_i(\xi)\ell_j(-1)\widetilde{\boldsymbol{F}}(\xi,-1)\cdot\left(-\hat{\boldsymbol{\eta}}\right)\mathrm{d}\xi = -w_i\ell_j(-1)\widetilde{G}\left(\xi_i,-1\right). \tag{46}$$

Similarly, we can compute the other components and the result for the full boundary integral reads

$$\begin{aligned}\oint_{\partial\chi}\phi\widetilde{\boldsymbol{F}}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}\sigma =& w_j\left(\ell_i(1)\widetilde{F}\left(1,\eta_j\right)-\ell_i(-1)\widetilde{F}\left(-1,\eta_j\right)\right)\\ &+ w_i\left(\ell_j(1)\widetilde{G}\left(\xi_i,1\right)-\ell_j(-1)\widetilde{G}\left(\xi_i,-1\right)\right).\end{aligned} \tag{47}$$

In the discontinuous Galerkin spectral element method the elements communicate by fluxes through the faces of each element. The solution on the boundary between two elements is allowed to be discontinuous, thus the limit towards the boundary of an element can have two values, one for each element it touches. The flux on the boundary must be replaced by a numerical flux so that the neighbouring elements can communicate. The numerical flux depends on the values of ρ just left and right of the boundary, i.e., ρ_{L} and ρ_{R}, which are computed by evaluating the interior solution (40a) at the boundary. For the numerical flux we take the upwind flux. Due to the transformation (26) of the flux to the reference domain, the physical upwind flux F is scaled at an edge by $\Delta l/2$ with Δl the length of the edge, such that the upwind flux $\widetilde{F} = \frac{1}{2}\Delta l F$ over an edge reads

$$\widetilde{F} = \frac{\Delta l}{2}\left(\boldsymbol{u}\cdot\hat{\boldsymbol{n}}\right)\begin{cases}\rho_{\mathrm{L}} \text{ if } \boldsymbol{u}\cdot\hat{\boldsymbol{n}} \geq 0,\\ \rho_{\mathrm{R}} \text{ if } \boldsymbol{u}\cdot\hat{\boldsymbol{n}} < 0,\end{cases} \tag{48}$$

where $\hat{\boldsymbol{n}}$ denotes the outward normal vector w.r.t. the element left of the boundary.

Next, we substitute expressions (43), (45) and (47) in equation (29), so that we obtain the semi-discrete ODE system for the expansion coefficients $\rho_{ij}(z)$:

$$\mathcal{J}_{ij}\frac{\mathrm{d}\rho_{ij}(z)}{\mathrm{d}z} = \sum_{n=0}^{N}\widehat{D}_{in}\widetilde{f}_{nj} + \sum_{m=0}^{N}\widehat{D}_{jm}\widetilde{g}_{im} - \left[\frac{\ell_i(1)}{w_i}\widetilde{F}(1,\eta_j) - \frac{\ell_i(-1)}{w_i}\widetilde{F}(-1,\eta_j) + \frac{\ell_j(1)}{w_j}\widetilde{G}(\xi_i,1) - \frac{\ell_j(-1)}{w_j}\widetilde{G}(\xi_i,-1)\right], \tag{49}$$

with the numerical fluxes $\widetilde{\boldsymbol{F}} = \left(\widetilde{F},\widetilde{G}\right)$ given by (48). This ODE system can be solved using any numerical time integrator, e.g., the classical fourth order Runge–Kutta method. Other popular choices in the literature are explicit low-storage Runge-Kutta methods, see [6,19,34].

The discontinuous Galerkin spectral element method approximates the exact solution by an Nth degree polynomial, so the global spatial error e for a typical mesh size Δx behaves as

$$e = \mathcal{O}(\Delta x^{N+1}). \tag{50}$$

Furthermore, the scheme is restricted by stability in terms of a CFL condition. For discontinuous Galerkin methods on quadrilaterals there is no direct known bound for the CFL condition. For triangular grids the relation between the Courant number and the shape of the triangles is studied in [7,30].

4 Optical Interfaces

In the phase space representation, the flow of ρ describes a beam of light propagating through an optical system. When the beam hits an optical interface, the momentum p changes discontinuously according to the law of specular reflection or Snell's law of refraction. Furthermore, from the discussion in Sect. 2 we know that the total luminous flux should remain constant throughout the optical system. The numerical solution should respect the actual physics, therefore, the discontinuity at optical interfaces coupled with conservation of energy should be incorporated into the DGSEM when we solve Liouville's equation.

In the DGSEM the solution is allowed to be discontinuous across the boundary connecting two or multiple elements. Therefore, the mesh in phase space is aligned such that elements adjacent to the interface have edges that coincide with the optical interface, across which the solution is discontinuous [31]. The elements in the DGSEM communicate through numerical fluxes, hence, we have to incorporate both Snell's law and the energy conservation constraint in the numerical flux when integrating (49). In particular, for a beam of light moving towards an optical interface, i.e., the velocity is directed towards the interface, we have to leave ρ free, whilst for a beam moving away from an optical interface, i.e., the velocity is directed away from the interface, we have a Dirichlet boundary condition for ρ due to Snell's law or the law of specular reflection [31].

Refraction or reflection causes the elements to be connected in a non-trivial manner at the optical interface. For example, one single element, on the side where light is moving towards the interface, can contribute to multiple elements on the other side. This occurs because both Snell's law and the law of reflection are non-linear in the momentum p. Therefore, this requires special treatment of the numerical fluxes to ensure that the scheme obeys energy conservation.

Van Lith et al. present Snell's function in [33], which is an explicit version of Snell's law and the law of specular reflection combined on phase space. Snell's function $\mathcal{S}$ relates

the momentum p of an incident ray to the outgoing momentum $\bar{p}$ of the ray. Let n_1 be the refractive index of the incident medium and n_2 the index of the transmitted medium. Then for a generic two-dimensional optical interface in the (q, z)-plane with surface unit normal $\vec{\nu} = (\nu_q, \nu_z)$ directed towards the incident medium, Snell's function reads

$$\bar{p} = \mathcal{S}(p; n_1, n_2, \vec{\nu}) := \begin{cases} p - (\psi + \operatorname{sgn}(n_2)\sqrt{\delta})\nu_q & \text{if } \delta \geq 0, \\ p - 2\psi\nu_q & \text{if } \delta < 0, \end{cases} \tag{51a}$$

with the auxiliary variables δ and ψ defined by

$$\delta := n_2^2 - n_1^2 + \psi^2, \quad \psi := \begin{pmatrix} p \\ \pm\sqrt{n_1^2 - p^2} \end{pmatrix} \cdot \begin{pmatrix} \nu_q \\ \nu_z \end{pmatrix}. \tag{51b}$$

In the expression for ψ the plus sign should be taken for rays that propagate in the positive z-direction, while the minus sign should be taken for rays that propagate in the negative z-direction. Furthermore, the sign of n_2, i.e., $\operatorname{sgn}(n_2)$, in the first case of (51a) can be used to accommodate embedded mirrors in a medium of refractive index $n_1 \geq 1$ by taking $n_2 = -n_1$, see [32]. Note that there is a so-called critical momentum p_c when $n_1 \geq n_2$ so that $\delta = 0$. For $\delta < 0$ all light will be reflected, referred to as total internal reflection (TIR), while for $\delta \geq 0$ light will be refracted. The outgoing momentum is computed as $\bar{p} = \mathcal{S}(p; n_1, n_2, \vec{\nu})$, for which we will frequently use the shorthand notation $\bar{p} = \mathcal{S}(p)$ and take the other parameters as given. Furthermore, the inverse of Snell's function will also be frequently used, for which we will use the shorthand notation $p = \mathcal{S}^{-1}(\bar{p})$. This means, find the momentum p such that $\bar{p} = \mathcal{S}(p)$. For example, for refraction the inverse reads [33]

$$p = \mathcal{S}^{-1}(\bar{p}) = -\mathcal{S}\left(-\bar{p}; n_2, n_1, -\vec{\nu}\right). \tag{52}$$

Snell's function (51) combined with (12) results in [33]

$$\rho\left(z^-, q^-, p^-\right) = \rho\left(z^+, q^+, p^+\right), \tag{53}$$

where $p^+ = \mathcal{S}(p^-; n_1, n_2, \vec{\nu})$ and the $\pm$ denote one-sided limits towards the optical interface. This relation allows us to relate the basic luminance ρ on both sides of the interface.

To elaborate the energy conservation constraint, we consider the following flat optical interface parallel to the z-axis

$$n(q) = \begin{cases} n_1 & \text{for } q \leq q_0, \\ n_2 & \text{for } q > q_0. \end{cases} \tag{54}$$

Note that the optical interface in phase space is represented by a line parallel to the p-axis. The optical interface has two sides where on one side the normal in phase space is directed towards $q < q_0$ and describes the part with refractive index n_1, whereas on the other side the normal is directed towards $q > q_0$ and describes the part with refractive index n_2. The normal in phase space is given by $\hat{\boldsymbol{n}} = (\pm 1, 0)$ with the plus sign for the direction towards $q > q_0$. Since the optical interface is represented by a line parallel to the p-axis at some constant q-value, only the q-component of the flux (20b) needs to be considered, i.e.,

$$f(z, q, p) = \rho(z, q, p)\frac{p}{\sqrt{n(z, q)^2 - p^2}}, \tag{55}$$

cf. (22).

In what follows, we assume that light is initially in the medium with refractive index n_1. We partition the optical interface, represented in phase space, into line segments for both

sides of the optical interface. The partitioning is based on whether light is moving towards or away from the optical interface.

The line segment on the side of the optical interface with velocity directed towards the optical interface is denoted L. The line segment L describes the incoming momentum of light from the medium with refractive index n_1, due to the assumption of light being initially in this medium.

The line segments just on the optical interface with velocity directed away from the optical interface is denoted R. The line segments R describe the outgoing momentum of light, and is further split into two parts, i.e., $R = R_{\mathrm{R}} \cup R_{\mathrm{T}}$. The line segment denoted R_{R} represents the momentum of light after total internal reflection, and is part of the optical interface where the refractive index is n_1, while the line segment denoted R_{T} is in the medium with refractive index n_2, and represents the momentum after transmission.

To distinguish the momentum taken from either side of the optical interface, we write $p \in L$ and $\bar{p} \in R$. First, consider the integral of the flux entering an arbitrary momentum interval $[\bar{p}_1, \bar{p}_2] \subseteq R_{\mathrm{T}}$. The integral reads

$$\int_{\bar{p}_1}^{\bar{p}_2} \rho(z^+, q_0^+, \bar{p}) \frac{\bar{p}}{\sqrt{n_2^2 - \bar{p}^2}} \,\mathrm{d}\bar{p}, \tag{56}$$

where q_0^+ denotes the limit towards the optical interface from the line segment R_{T}. Relation (53) implies

$$\int_{\bar{p}_1}^{\bar{p}_2} \rho\left(z^+, q_0^+, \bar{p}\right) \frac{\bar{p}}{\sqrt{n_2^2 - \bar{p}^2}} \,\mathrm{d}\bar{p} = \int_{\bar{p}_1}^{\bar{p}_2} \rho\left(z^-, q_0^-, \mathcal{S}^{-1}(\bar{p})\right) \frac{\bar{p}}{\sqrt{n_2^2 - \bar{p}^2}} \,\mathrm{d}\bar{p}, \tag{57}$$

where q_0^- denotes the limit towards the optical interface from the line segment L. Subsequently, we transform the integral using $\bar{p} = \mathcal{S}(p) = \mathcal{S}(p; n_1, n_2, \vec{\nu})$ resulting in

$$\int_{\bar{p}_1}^{\bar{p}_2} \rho\left(z^+, q_0^+, \bar{p}\right) \frac{\bar{p}}{\sqrt{n_2^2 - \bar{p}^2}} \,\mathrm{d}\bar{p} = \int_{p_1}^{p_2} \rho\left(z^-, q_0^-, p\right) \frac{\mathcal{S}(p)}{\sqrt{n_2^2 - \mathcal{S}(p)^2}} \frac{\mathrm{d}\mathcal{S}(p)}{\mathrm{d}p} \,\mathrm{d}p, \tag{58}$$

where $\bar{p}_i = \mathcal{S}(p_i)$ for $i = 1, 2$, and $[p_1, p_2] \subseteq L$.

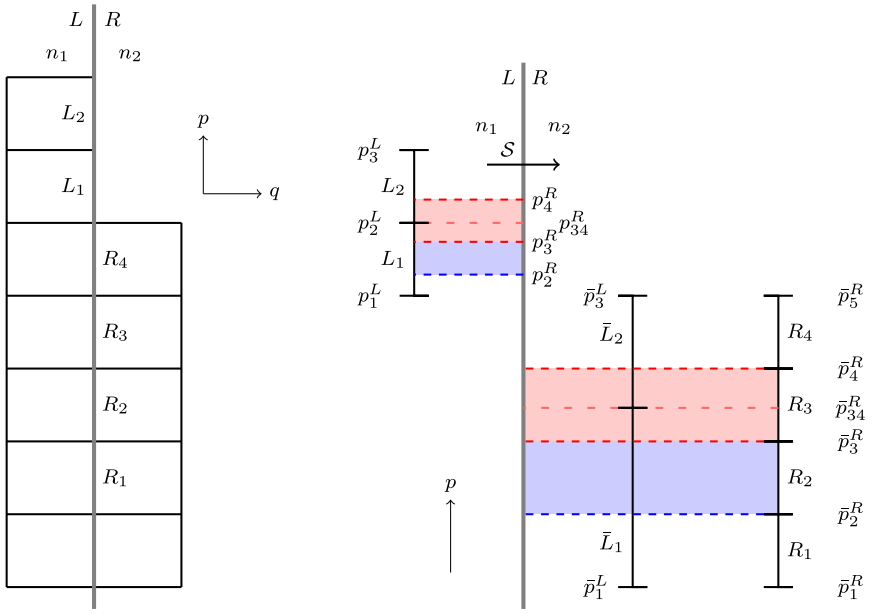
The relation for reflection can be derived similarly by considering the integral of the flux entering an arbitrary momentum interval $[\bar{p}_3, \bar{p}_4] \subseteq R_{\mathrm{R}}$. We obtain the relation

$$\int_{\bar{p}_3}^{\bar{p}_4} \rho\left(z^+, q_0^+, \bar{p}\right) \frac{\bar{p}}{\sqrt{n_1^2 - \bar{p}^2}} \,\mathrm{d}\bar{p} = \int_{p_3}^{p_4} \rho\left(z^-, q_0^-, p\right) \frac{\mathcal{S}(p)}{\sqrt{n_1^2 - \mathcal{S}(p)^2}} \frac{\mathrm{d}\mathcal{S}(p)}{\mathrm{d}p} \,\mathrm{d}p, \tag{59}$$

with $\bar{p}_i = \mathcal{S}(p_i)$ for $i = 3, 4$, and $[p_3, p_4] \subseteq L$. The relations (58) and (59) describe how the fluxes leaving L are related to the fluxes entering R_{T} or R_{R}, respectively. Henceforth, they are known as energy conservation constraints.

For the particular optical interface (54) the constraints can be simplified. Since, we assumed that light was initially in the medium with refractive index n_1, the optical interface normal on the full position space is equal to $\vec{\nu} = (\nu_q, \nu_z) = (-1, 0)$. Snell's function (51) reduces for this flat interface to

$$\bar{p} = \mathcal{S}(p; n_1, n_2, \vec{\nu}) = \begin{cases} \sqrt{n_2^2 - n_1^2 + p^2} & \text{if } p \geq p_{\mathrm{c}}, \\ -p & \text{if } p < p_{\mathrm{c}}, \end{cases} \tag{60}$$



(a) Elements in phase space connected due to Snell's function.

(b) Illustration of the geometry at the optical interface.

Fig. 3 Conservative handling of fluxes. The incident and transmitted momenta are related by $\bar{p} = \mathcal{S}(p)$

with $p_c = \sqrt{n_1^2 - n_2^2}$. Then, the energy conservation constraint for R_T, given by (58), can be simplified by noting that for refraction the following relations hold

$$\frac{\mathrm{d}\mathcal{S}(p)}{\mathrm{d}p} = \frac{p}{\mathcal{S}(p)}, \quad \sqrt{n_2^2 - \mathcal{S}(p)^2} = \sqrt{n_1^2 - p^2}.$$

Hence, we obtain

$$\int_{\bar{p}_1}^{\bar{p}_2} \rho\left(z^+, q_0^+, \bar{p}\right) \frac{\bar{p}}{\sqrt{n_2^2 - \bar{p}^2}} \,\mathrm{d}\bar{p} = \int_{p_1}^{p_2} \rho\left(z^-, q_0^-, p\right) \frac{p}{\sqrt{n_1^2 - p^2}} \,\mathrm{d}p, \tag{61a}$$

and similarly for reflection the constraint for R_R, given by (59), reduces to

$$\int_{\bar{p}_3}^{\bar{p}_4} \rho\left(z^+, q_0^+, \bar{p}\right) \frac{\bar{p}}{\sqrt{n_1^2 - \bar{p}^2}} \,\mathrm{d}\bar{p} = \int_{p_3}^{p_4} \rho\left(z^-, q_0^-, p\right) \frac{p}{\sqrt{n_1^2 - p^2}} \,\mathrm{d}p. \tag{61b}$$

The balances (61) have to be combined with relation (53) to ensure the scheme conserves energy. However, the coupling between the line segments L and R is not straightforward, which will be discussed in the next section.

4.1 Conservative Handling of Fluxes

First, consider only the refractive part of the optical interface. Elements adjacent to the optical interface in phase space are shown in Fig. 3a. These elements have edges on the optical interface and these edges are denoted by L_i $(i = 1, 2)$ and R_j $(j = 1, 2, 3, 4)$. Due

to refraction, the value of ρ in the elements that contain R_j as an edge is determined by the flow through the elements that contain L_1 and L_2. In fact, taking a closer look at how Snell's function connects the line segments from L to R in momentum space at the optical interface, we obtain for example Fig. 3b. In Fig. 3b L_i and R_j denote line segments along the optical interface. The basic luminance ρ along the line segments L_i and R_j are represented by their inner-element solution evaluated at the optical interface. To simplify notation, we denote these polynomials along the optical interface by $\rho^{L_i}(p)$ with $i = 1, 2$ and $\rho^{R_j}(p)$ with $j = 1, 2, 3, 4$. For example:

$$\rho^{L_i}(p) = \sum_{j=0}^{N} \rho_j^{L_i} \ell_j(\zeta(p)), \tag{62}$$

where $\zeta = \zeta(p)$ denotes the line segment's local reference coordinate along the interface.

In Fig. 3b also virtual line segments $\bar{L}_i$ are shown. The virtual line segment $\bar{L}_i$ is the image of L_i under $\mathcal{S}$, i.e.,

$$\bar{L}_i = \mathcal{S}(L_i). \tag{63}$$

Hence, the endpoints of these line segments are found applying Snell's function to the endpoints of L_i, i.e., $\bar{p}_i^L = \mathcal{S}(p_i^L)$. Note that due to Snell's function, the line segments L_i are stretched or compressed in the momentum direction. Computing the endpoints $\bar{p}_i^L$ allows us to determine which line segments before the optical interface contribute to a single line segment after the optical interface. From the figure we see that part of L_1 contributes to R_2 (the blue coloured region). Therefore, a relation connecting $\rho^{L_1}(p)$ and $\rho^{R_2}(p)$ on opposite sides of the optical interface must be found. Hence, as a first step applying relation (53) to a polynomial on L_i, allows us to find the corresponding ρ on $\bar{L}_i$, i.e.,

$$\rho^{L_i}(p) = \rho^{L_i}\left(\mathcal{S}^{-1}(\bar{p})\right) = \rho^{\bar{L}_i}(\bar{p}), \tag{64}$$

with $\bar{p} = \mathcal{S}(p)$.

The coupling between line segments that do not exactly match, as shown in Fig. 3b, is similar to what is known as a geometrically non-conforming mesh [20,23]. In [23] the authors describe a discontinuous Galerkin method for non-conforming meshes, applied to Maxwell's equations that form a hyperbolic system of PDEs. In their approach to treating non-conforming interfaces the solutions are first transferred to an intermediate construct called a 'mortar', and on this mortar the numerical fluxes are computed and transferred back to the corresponding elements. The transfer of the solutions and numerical fluxes is done using a least-squares matching, with integrals evaluated using Gauss-Legendre quadrature [5].

We will take a slightly different approach since in Liouville's equation for optics the flux $\boldsymbol{f}$ is discontinuous across an optical interface. Instead, relation (53) and Snell's function together describe how ρ transforms across an optical interface, cf. (64), therefore, a least-squares matching of the polynomials describing ρ along either side of the interface is used with Snell's function directly incorporated and an additional constraint is used to satisfy energy conservation.

For the reflective part of a flat interface that is parallel to the z-axis, Snell's function reduces to $\bar{p} = -p$, see (60). The conservative treatment of these types of optical interfaces is easily accommodated by choosing a mesh such that the elements and nodes are symmetric with respect to the line $p = 0$ and, therefore, the constraint (61b) is easily satisfied. Due to this choice of mesh each node $\bar{p}_j \in R_{\mathrm{R}}$ will exactly correspond to $-\bar{p}_j = p_j \in L$ and a

point-by-point transfer of ρ can be made. Henceforth, the following exposition of the method describes the method considering only refraction.

4.2 Contribution from One Element

From Fig. 3b we see that the line segment R_2 only depends on the solution in L_1. The polynomial ρ^{R_2} must thus be computed from the polynomial ρ^{L_1} with the additional constraint of energy conservation. That is, the integral of the flux within the blue interval on either side of the optical interface should be equal analogous to equation (61a). Therefore, the constrained least-squares approximation reads

$$\min_{\rho^{R_2}\in\mathbb{P}_N} \int_{\bar{p}_2^R}^{\bar{p}_3^R} \left[\rho^{R_2}(\bar{p}) - \rho^{\bar{L}_1}(\bar{p})\right]^2 \mathrm{d}\bar{p}, \tag{65a}$$

$$\text{subject to} \int_{\bar{p}_2^R}^{\bar{p}_3^R} F^{R_2}(\bar{p})\mathrm{d}\bar{p} = \int_{p_2^R}^{p_3^R} F^{L_1}(p)\,\mathrm{d}p. \tag{65b}$$

Here, $[p_2^R, p_3^R] \subseteq L_1 = [p_1^L, p_2^L]$ and the momenta on both sides are related by $p_i^R := \mathcal{S}^{-1}(\bar{p}_i^R)$, see Fig. 3b. Furthermore, the numerical fluxes are defined as expansions in the Lagrange polynomial basis on Gauss-Legendre nodes, similar to (62), with flux coefficients $F_j := u_j\rho_j$. The minimisation of the integral in (65a) requires finding a polynomial that matches in the least-squares sense, while the constraint (65b) ensures that the scheme conserves energy.

The integrals in the constrained minimisation problem (65) are transformed to reference line segments, so that we can compute the integrals using Gauss-Legendre quadrature. To be more specific, the integral on the LHS of (65b) and the integral in (65a) are transformed to the reference line segment along R_2, while the integral on the RHS of (65b) is transformed to the reference line segment along L_1. Omitting the element's subscripts, applying relation (64) and introducing an auxiliary function Ξ for ease of notation, we obtain

$$\begin{aligned} &\min_{\rho^R\in\mathbb{P}_N} \int_{-1}^{1} \left[\rho^R(\zeta) - \rho^L(\Xi(\zeta))\right]^2 \mathrm{d}\zeta, \\ &\text{subject to} \quad \Delta\bar{p}^R \int_{-1}^{1} F^R(\zeta)\mathrm{d}\zeta = \Delta p^L \int_{\sigma^L}^{\sigma^L+\lambda^L} F^L(\zeta)\,\mathrm{d}\zeta, \end{aligned} \tag{66}$$

where $\Delta\bar{p}^R := \bar{p}_3^R - \bar{p}_2^R$ and $\Delta p^L := p_2^L - p_1^L$. Furthermore, the coefficients $\sigma^L \in [-1, 1]$ and $\lambda^L \in [0, 2]$ denote the offset and scaling in L_1's reference frame, such that $p(\sigma^L) = p_2^R$ and $p(\sigma^L + \lambda^L) = p_3^R$ in L_1. Finally, the auxiliary function Ξ reads

$$\Xi\left(\zeta; p^L, \Delta p^L, \bar{p}^R, \Delta\bar{p}^R\right) = 2\frac{\mathcal{S}^{-1}\left(\bar{p}^R + \frac{1}{2}\Delta\bar{p}^R(1+\zeta)\right) - p^L}{\Delta p^L} - 1. \tag{67}$$

This function relates the reference frame coordinates for a momentum interval $[\bar{p}^R, \bar{p}^R + \Delta\bar{p}^R]$ past the optical interface to the reference frame coordinates on a momentum interval $[p^L, p^L + \Delta p^L]$ before the optical interface.

Next, we write the constrained minimisation problem (66) in terms of a Lagrange function $\mathcal{L}$ with a Lagrange multiplier μ for the energy conservation constraint, i.e.,

$$\begin{aligned}\mathcal{L} = &\frac{1}{2}\int_{-1}^{1}\left[\rho^R(\zeta) - \rho^L(\Xi(\zeta))\right]^2 \mathrm{d}\zeta \\ &+ \mu\left[\Delta\bar{p}^R\int_{-1}^{1} F^R(\zeta)\,\mathrm{d}\zeta - \Delta p^L\int_{\sigma^L}^{\sigma^L+\lambda^L} F^L(\zeta)\,\mathrm{d}\zeta\right].\end{aligned} \tag{68}$$

The coefficients ρ_j^R for the polynomial $\rho^R \in \mathbb{P}_N$ can then be computed by solving

$$\frac{\partial\mathcal{L}}{\partial\rho_i^R} = 0, \quad \text{for } i = 0, 1, ..., N, \tag{69a}$$

$$\frac{\partial\mathcal{L}}{\partial\mu} = 0. \tag{69b}$$

Recalling that both F^L and F^R are written as expansions in the Lagrange polynomial basis on Gauss-Legendre nodes, we obtain for the energy conservation constraint (69b)

$$\Delta\bar{p}^R\int_{-1}^{1}\sum_{j=0}^{N} u_j^R\rho_j^R\ell_j(\zeta)\,\mathrm{d}\zeta = \Delta p^L\int_{\sigma^L}^{\sigma^L+\lambda^L}\sum_{j=0}^{N} u_j^L\rho_j^L\ell_j(\zeta)\,\mathrm{d}\zeta, \tag{70}$$

where we have used $F_j = u_j\rho_j$.

To evaluate the second integral, we transform it to the reference interval $[-1, 1]$ using $\zeta(\xi) = \sigma^L + \frac{1}{2}\lambda^L(1+\xi)$. Next, we replace both integrals with Gauss-Legendre quadrature to find the exact values, since both integrands are at most Nth degree polynomials, therefore we obtain

$$\Delta\bar{p}^R\sum_{j=0}^{N} w_j u_j^R\rho_j^R = \Delta p^L\frac{\lambda^L}{2}\sum_{j=0}^{N} u_j^L\rho_j^L\sum_{k=0}^{N} w_k\ell_j\left(\sigma^L + \frac{1}{2}\lambda^L(1+\xi_k)\right), \tag{71}$$

with ξ_k and w_k the Gauss-Legendre nodes and weights, respectively.

Recalling again that the polynomials ρ^L and ρ^R are written as an expansion in a Lagrange polynomial basis, cf. (62), we can rewrite the equations (69a) to

$$0 = \int_{-1}^{1}\left[\rho^R(\zeta) - \rho^L(\Xi(\zeta))\right]\ell_i(\zeta)\,\mathrm{d}\zeta + \mu\Delta\bar{p}^R\int_{-1}^{1} u_i^R\ell_i(\zeta)\,\mathrm{d}\zeta, \quad \text{for } i = 0, 1, ..., N. \tag{72}$$

For the evaluation of the first integral, we introduce a generic auxiliary variable S_{ij}, given by

$$S_{ij}\left(\sigma^R, \lambda^R\right) := \int_{\sigma^R}^{\sigma^R+\lambda^R}\ell_i(\zeta)\ell_j(\Xi(\zeta))\,\mathrm{d}\zeta, \tag{73}$$

with Ξ defined in (67). The integral is evaluated by transforming to the reference interval and subsequently applying Gauss-Legendre quadrature. The integration interval $[\sigma^R, \sigma^R + \lambda^R]$ of S_{ij} depends on how large R_j is compared to $\bar{L}_i$. In this case, the entire line segment R_2 fits in $\bar{L}_1$, therefore, the integral over R_2 is transformed to a reference line segment $[-1, +1]$, corresponding to $\sigma^R = -1$ and $\lambda^R = 2$.

The integrals in (72) are evaluated using Gauss-Legendre quadrature resulting in

$$\sum_{j=0}^{N} M_{ij}\rho_j^R + \mu\Delta\bar{p}^R u_i^R w_i = \sum_{j=0}^{N} S_{ij}\,(-1,2)\,\rho_j^L, \quad \text{for } i = 0, 1, ..., N, \tag{74a}$$

with

$$M_{ij} = \int_{-1}^{1} \ell_i(\zeta)\ell_j(\zeta)\,\mathrm{d}\zeta, \tag{74b}$$

and S_{ij} given by (73). The integral M_{ij} describes the orthogonality of the Lagrange polynomials on Gauss-Legendre nodes, and therefore, is easily evaluated, cf. (35), to be

$$M_{ij} = w_i\delta_{ij}. \tag{75}$$

The coefficients M_{ij} and S_{ij} are elements of the matrices $\boldsymbol{M}, \boldsymbol{S} \in \mathbb{R}^{(N+1)\times(N+1)}$. The matrix $\boldsymbol{M}$ is simply a diagonal matrix containing the Gauss-Legendre quadrature weights, i.e., $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{w})$ with $\boldsymbol{w} = (w_0, w_1, ..., w_N)^{\mathrm{T}}$.

By defining

$$\alpha_j^R := \Delta\bar{p}^R u_j^R, \quad \beta_j^L := \Delta p^L \frac{\lambda^L}{2} u_j^L \sum_{k=0}^{N} w_k \ell_j\left(\sigma^L + \frac{1}{2}\lambda^L\,(1+\xi_k)\right), \tag{76}$$

as the components of the vectors $\boldsymbol{\alpha}^R$ and $\boldsymbol{\beta}^L$, we can write the linear system given by (74) and (71) for $\boldsymbol{\rho}^R = (\rho_0^R, \rho_1^R, ..., \rho_N^R)^{\mathrm{T}}$ and μ compactly in matrix-vector form:

$$\begin{pmatrix} \mathrm{diag}\,(\boldsymbol{w}) & \boldsymbol{\alpha}^R \circ \boldsymbol{w} \\ (\boldsymbol{\alpha}^R \circ \boldsymbol{w})^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\rho}^R \\ \mu \end{pmatrix} = \begin{pmatrix} \boldsymbol{S} \\ (\boldsymbol{\beta}^L)^{\mathrm{T}} \end{pmatrix} \boldsymbol{\rho}^L, \tag{77}$$

where we take the arguments for $\boldsymbol{S}$ as understood. Furthermore, $\circ$ denotes the Hadamard product for vectors $\boldsymbol{a} = (a_0, a_1, ..., a_N)^{\mathrm{T}}$ and $\boldsymbol{b} = (b_0, b_1, ..., b_N)^{\mathrm{T}}$, i.e., $\boldsymbol{a} \circ \boldsymbol{b} = (a_0b_0, a_1b_1, ..., a_Nb_N)^{\mathrm{T}}$. Let us denote the matrix on the LHS by $\boldsymbol{A}$:

$$\boldsymbol{A} := \begin{pmatrix} \mathrm{diag}\,(\boldsymbol{w}) & \boldsymbol{\alpha}^R \circ \boldsymbol{w} \\ (\boldsymbol{\alpha}^R \circ \boldsymbol{w})^{\mathrm{T}} & 0 \end{pmatrix}. \tag{78}$$

The determinant of this matrix reads

$$\det(\boldsymbol{A}) = -\left(\sum_{i=0}^{N} \left(\alpha_i^R\right)^2 w_i\right) \prod_{i=0}^{N} w_i, \tag{79}$$

see Appendix A for details. Therefore, the matrix is only singular if all coefficients $\alpha_i^R = 0$, or equivalent if all velocities $u_i = 0$, which would mean no flux can enter the element from that side. Hence, we can safely assume that the matrix $\boldsymbol{A}$ is regular.

An analytical inverse for the matrix $\boldsymbol{A}$ is derived in Appendix A, and reads

$$\boldsymbol{A}^{-1} = \frac{1}{r}\begin{pmatrix} \boldsymbol{B} & -\boldsymbol{\alpha}^R \\ (-\boldsymbol{\alpha}^R)^{\mathrm{T}} & 1 \end{pmatrix}, \quad r := -\sum_{i=0}^{N} \left(\alpha_i^R\right)^2 w_i, \tag{80}$$

where the coefficients of the matrix $\boldsymbol{B}$ read

$$B_{ij} := \begin{cases} \left(\alpha_i^R\right)^2 + \frac{r}{w_i} & \text{if } i = j, \\ \alpha_i^R\alpha_j^R & \text{if } i \neq j. \end{cases} \tag{81}$$

Now, we can directly obtain an expression for the Dirichlet boundary condition values $\boldsymbol{\rho}^R$ in terms of $\boldsymbol{\rho}^L$, i.e.,

$$\boldsymbol{\rho}^R = \frac{1}{r}\begin{pmatrix}\boldsymbol{B} & -\boldsymbol{\alpha}^R\end{pmatrix}\begin{pmatrix}\boldsymbol{S} \\ \left(\boldsymbol{\beta}^L\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^L =: \boldsymbol{C}\boldsymbol{\rho}^L. \tag{82}$$

Note that for problems where the refractive index n does not depend on z, the coefficient matrix $\boldsymbol{C}$ relating $\boldsymbol{\rho}^R$ and $\boldsymbol{\rho}^L$ can be pre-computed and re-used during integration along the z-axis.

4.3 Contributions from Multiple Elements

From Fig. 3b we see that the element R_3 depends on both $\bar{L}_1$ and $\bar{L}_2$. The idea remains the same, i.e., to use a least-squares matching with a constraint to ensure that the scheme is energy conservative. The constrained least-squares problem for R_3 reads

$$\min_{\rho^{R_3}\in\mathbb{P}_N} \quad \int_{\bar{p}_3^R}^{\bar{p}_4^R}\left[\rho^{R_3}(\bar{p}) - \rho^{\bar{L}}(\bar{p})\right]^2 \mathrm{d}\bar{p}, \tag{83a}$$

$$\text{subject to} \quad \int_{\bar{p}_3^R}^{\bar{p}_4^R} F^{R_3}(\bar{p})\,\mathrm{d}\bar{p} = \int_{p_3^R}^{p_{34}^R} F^{L_1}(p)\,\mathrm{d}p + \int_{p_{34}^R}^{p_4^R} F^{L_2}(p)\,\mathrm{d}p, \tag{83b}$$

where $p_{34}^R := \mathcal{S}^{-1}(\bar{p}_{34}^R)$ and $\bar{p}_{34}^R$ is the momentum value where the intervals $\bar{L}_1$ and $\bar{L}_2$ meet, see Fig. 3b. Furthermore, $\rho^{\bar{L}}$ contains the contributions from ρ^{L_1} and ρ^{L_2}, and is defined by

$$\rho^{\bar{L}}(\bar{p}) := \begin{cases}\rho^{L_1}(\mathcal{S}^{-1}(\bar{p})) & \text{for } \bar{p}_3^R \le \bar{p} \le \bar{p}_{34}^R, \\ \rho^{L_2}(\mathcal{S}^{-1}(\bar{p})) & \text{for } \bar{p}_{34}^R < \bar{p} \le \bar{p}_4^R.\end{cases} \tag{84}$$

The integrals in (83) are transformed to their respective line segments, e.g., the integral on the LHS of (83b) and the integral in (83a) are transformed to the reference interval $[-1, 1]$ along R_3, such that we obtain

$$\begin{aligned}\min_{\rho^R\in\mathbb{P}_N} \quad & \int_{-1}^{1}\left[\rho^R(\zeta) - \rho^L\left(\Xi^L(\zeta)\right)\right]^2 \mathrm{d}\zeta \\ \text{subject to} \quad & \Delta\bar{p}^R\int_{-1}^{1} F^R(\zeta)\,\mathrm{d}\zeta = \Delta p^{L_1}\int_{\sigma^{L_1}}^{\sigma^{L_1}+\lambda^{L_1}} F^{L_1}(\zeta)\,\mathrm{d}\zeta \\ & + \Delta p^{L_2}\int_{\sigma^{L_2}}^{\sigma^{L_2}+\lambda^{L_2}} F^{L_2}(\zeta)\,\mathrm{d}\zeta,\end{aligned} \tag{85a}$$

with

$$\Xi^L(\zeta) := \begin{cases}\Xi\left(\zeta;\, p_1^L,\ \Delta p^{L_1},\ \bar{p}_3^R,\ \bar{p}_{34}^R - \bar{p}_3^R\right) & \text{for } -1 \le \zeta \le \kappa, \\ \Xi\left(\zeta;\, p_2^L,\ \Delta p^{L_2},\ \bar{p}_{34}^R,\ \bar{p}_4^R - \bar{p}_{34}^R\right) & \text{for } \kappa < \zeta \le 1,\end{cases} \tag{85b}$$

where we write R instead of R_3 for brevity, and κ is defined such that $p(\kappa) = \bar{p}_{34}^R$ in R_3 and $\Delta\bar{p}^R := \bar{p}_4^R - \bar{p}_3^R$, $\Delta p^{L_1} := p_2^L - p_1^L$ and $\Delta p^{L_2} := p_3^L - p_2^L$. Note that $\sigma^{L_1} + \lambda^{L_1} = 1$ and $\sigma^{L_2} = -1$, however, for illustration purposes we will keep using the variables rather than

these values. The Lagrange function $\mathcal{L}$ for this constrained minimisation problem reads

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\int_{-1}^{1}\left[\rho^R(\zeta)-\rho^L\left(\Xi^L(\zeta)\right)\right]^2\,\mathrm{d}\zeta+\mu\left[\Delta\bar{p}^R\int_{-1}^{1}F^R(\zeta)\,\mathrm{d}\zeta\right.\\ &\left.-\Delta p^{L_1}\int_{\sigma^{L_1}}^{\sigma^{L_1}+\lambda^{L_1}}F^{L_1}(\zeta)\,\mathrm{d}\zeta-\Delta p^{L_2}\int_{\sigma^{L_2}}^{\sigma^{L_2}+\lambda^{L_2}}F^{L_2}(\zeta)\,\mathrm{d}\zeta\right].\end{aligned} \tag{86}$$

The coefficients ρ_j^R for the polynomial $\rho^R\in\mathbb{P}_N$ can be found by solving

$$\frac{\partial\mathcal{L}}{\partial\rho_i^R}=0\quad\text{for } i=0,1,...,N,\quad\frac{\partial\mathcal{L}}{\partial\mu}=0.$$

Following the same steps as in Sect. 4.2, we obtain the following system of equations

$$\sum_{j=0}^{N}M_{ij}\rho_j^R+\mu w_i\alpha_i^R=\sum_{j=0}^{N}\left[S_{ij}\left(-1,1+\kappa\right)\rho_j^{L_1}+S_{ij}\left(\kappa,1-\kappa\right)\rho_j^{L_2}\right],$$
$$\text{for } i=0,1,...,N, \tag{87a}$$

$$\sum_{j=0}^{N}w_j\alpha_j^R\rho_j^R=\sum_{j=0}^{N}\beta_j^{L_1}\rho_j^{L_1}+\sum_{j=0}^{N}\beta_j^{L_2}\rho_j^{L_2}, \tag{87b}$$

with

$$\alpha_j^R:=\Delta\bar{p}^R u_j^R,\quad \beta_j^{L_1}:=\Delta p^{L_1}\frac{\lambda^{L_1}}{2}u_j^{L_1}\sum_{k=0}^{N}w_k\ell_j\left(\sigma^{L_1}+\frac{1}{2}\lambda^{L_1}\left(1+\xi_k\right)\right), \tag{87c}$$

$$\beta_j^{L_2}:=\Delta p^{L_2}\frac{\lambda^{L_2}}{2}u_j^{L_2}\sum_{k=0}^{N}w_k\ell_j\left(\sigma^{L_2}+\frac{1}{2}\lambda^{L_2}\left(1+\xi_k\right)\right). \tag{87d}$$

The linear system described by (87) can once again be assembled into a matrix-vector form:

$$\begin{pmatrix}\mathrm{diag}\left(\boldsymbol{w}\right) & \boldsymbol{\alpha}^R\circ\boldsymbol{w}\\ \left(\boldsymbol{\alpha}^R\circ\boldsymbol{w}\right)^{\mathrm{T}} & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\rho}^R\\ \mu\end{pmatrix}=\begin{pmatrix}\boldsymbol{S}^{L_1}\\ \left(\boldsymbol{\beta}^{L_1}\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^{L_1}+\begin{pmatrix}\boldsymbol{S}^{L_2}\\ \left(\boldsymbol{\beta}^{L_2}\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^{L_2}, \tag{88}$$

where we have used the shorthand notation $\boldsymbol{S}^{L_1}=\left(S_{ij}(-1,1+\kappa)\right)$ and $\boldsymbol{S}^{L_2}=\left(S_{ij}(\kappa,1-\kappa)\right)$. Note that the matrix on the LHS is exactly the same as the matrix obtained in the previous section, except for possibly different values for α_j^R. Therefore, we can again solve the linear system explicitly for the Dirichlet boundary condition values $\boldsymbol{\rho^R}$, resulting in

$$\boldsymbol{\rho}^R=\frac{1}{r}\begin{pmatrix}\boldsymbol{B} & -\boldsymbol{\alpha}^R\end{pmatrix}\left[\begin{pmatrix}\boldsymbol{S}^{L_1}\\ \left(\boldsymbol{\beta}^{L_1}\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^{L_1}+\begin{pmatrix}\boldsymbol{S}^{L_2}\\ \left(\boldsymbol{\beta}^{L_2}\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^{L_2}\right], \tag{89}$$

cf. (82). This result can of course be generalised to K elements contributing to $\boldsymbol{\rho}^R$, resulting in

$$\boldsymbol{\rho}^R=\frac{1}{r}\begin{pmatrix}\boldsymbol{B} & -\boldsymbol{\alpha}^R\end{pmatrix}\left[\sum_{k=1}^{K}\begin{pmatrix}\boldsymbol{S}^{L_k}\\ \left(\boldsymbol{\beta}^{L_k}\right)^{\mathrm{T}}\end{pmatrix}\boldsymbol{\rho}^{L_k}\right]. \tag{90}$$

4.4 Overview

To summarise, during a z-step the numerical fluxes over the optical interface are evaluated as follows. First, the elements are identified that have an edge on the optical interface. Those elements are separated into elements with velocities directed towards the optical interface, denoted L, and elements with velocities directed away from the optical interface, denoted R. For the elements from L the solution is evaluated at edges on the optical interface. The numerical flux over the edges for the elements L can be directly computed as there is no constraint on ρ. For each element from R there is a Dirichlet boundary condition on the edge at the optical interface given by (53), that is incorporated into the numerical flux.

The value for the Dirichlet boundary condition is determined from the elements L, as follows. To determine which elements from L contribute to a single R element, Snell's function is applied to the momentum boundaries of the elements L. Subsequently, the geometric quantities relating the element sizes are computed. Next, the momenta p at the quadrature nodes, for evaluation of the integral S_{ij}, are determined. Subsequently, we apply $\mathcal{S}^{-1}$ to these nodes, and compute Ξ using (67). Hereafter, the integrals S_{ij} are evaluated and the coefficients α_j^L, α_j^R are computed. Finally, the values for the Dirichlet boundary condition can be found from their contributing L-elements by applying (90).

5 Results

Numerical experiments were performed for two examples. The first example features light propagating through a gradient-index medium. The smooth refractive index field of the medium fits naturally into the DGSEM for solving Liouville's equation. For such optical systems ray-tracers usually have to resort to difficult to obtain closed-form expressions for the trajectory of the rays [2], or use symplectic integrators to solve Hamilton's equations for every ray [27]. Solving Liouville's equation with the DGSEM provides directly the energy distribution, i.e., the basic luminance ρ for the optical system. Furthermore, the method conserves energy by design.

The second example features a single optical interface. The problem exhibits both total internal reflection and refraction. At the optical interface we apply the strategy outlined in Sect. 4. Furthermore, a comparison is made between solving Liouville's equation using the DGSEM and applying quasi-Monte Carlo ray tracing [14]. The illuminance is solved using both methods and the performance of both methods is tested.

5.1 Elliptic Waveguide

As a first example, we consider the elliptic waveguide [35] which features a smooth refractive index field given by

$$n(q) = \begin{cases} \sqrt{n_0^2 - \kappa^2 q^2} & \text{if } \kappa\,|q| \leq \sqrt{n_0^2 - 1}, \\ 1 & \text{otherwise.} \end{cases} \tag{91}$$

The parameters n_0 and κ are taken to be $n_0 = 1.4$ and $\kappa = \sqrt{n_0^2 - 1}$. The refractive index field and several rays are shown in Fig. 4. We observe that the elliptic waveguide contains light much like an optical fibre. Hamilton's equations (8) for rays inside the elliptic waveguide

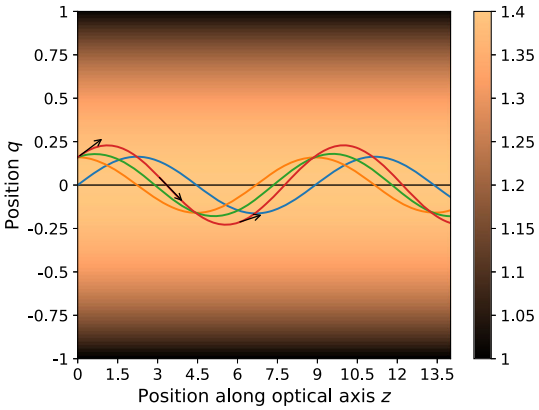


Fig. 4 Elliptic waveguide: background colour indicates the refractive index value $n(q)$, and the solid lines represent ray trajectories. Arrows indicate the direction of the ray, i.e., the momenta (p_z, p)

read

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}z} &= -\frac{p}{h}, \\ \frac{\mathrm{d}p}{\mathrm{d}z} &= \frac{\kappa^2}{h} q. \end{aligned} \tag{92}$$

Since the refractive index field does not depend on z, the Hamiltonian h remains constant for each ray. The solution of (92) reads

$$\begin{aligned} q(z) &= q_0 \cos\left(\frac{\kappa}{h} z\right) - \frac{p_0}{\kappa} \sin\left(\frac{\kappa}{h} z\right), \\ p(z) &= p_0 \cos\left(\frac{\kappa}{h} z\right) + \kappa q_0 \sin\left(\frac{\kappa}{h} z\right), \end{aligned} \tag{93}$$

where the initial conditions are given by $(q(0),\, p(0)) = (q_0,\, p_0)$. Note that from the refractive index field n and the Hamiltonian h we obtain [35]

$$\kappa^2 q^2 + p^2 = n_0^2 - h^2, \tag{94}$$

where the right-hand side is constant when we move along the z-axis. We can readily see that the trajectories follow an elliptical path in phase space, hence, the name elliptic waveguide.

Let the function φ_m be defined as

$$\varphi_m(x) := \begin{cases} \cos^{m+1}\left(\frac{\pi}{2} x^2\right) & \text{if } |x| < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{95}$$

which is a C_0^m-function, meaning its first m derivatives are continuous and has compact support. The function φ_m is plotted in Fig. 5 for $m = 7, 28$. We solve Liouville's equation (20a) with the following initial condition

$$\rho_0(q, p) = \varphi_m\left(\frac{q}{\sigma_q}\right) \varphi_m\left(\frac{p}{\sigma_p}\right), \tag{96}$$

at $z = 0$ and on the boundary of the domain we leave ρ free whenever the velocity field is pointing out of the domain, otherwise we prescribe $\rho = 0$. In (96) we take $m = 7, \sigma_q = 0.25$ and $\sigma_p = 0.1$.

The ODE system (49) is integrated using the low-storage 4th order Runge-Kutta method by Zingg and Chisholm [37]. The numerical solution is integrated from $z = 0$ to $z = Z = 3$.

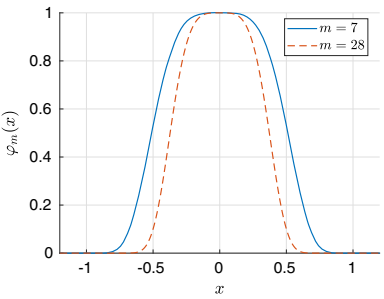
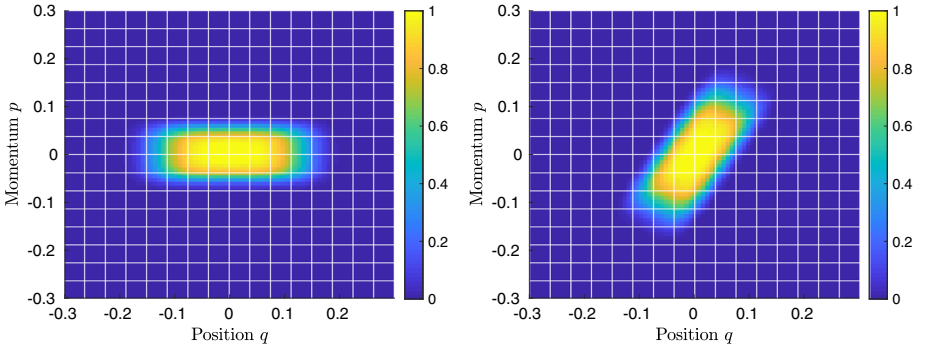


Fig. 5 Function φ_m for $m = 7$ and $m = 28$



(a) at $z = 0$ **(b)** at $z = Z$

Fig. 6 Elliptic waveguide: basic luminance distributions $\rho(z, q, p)$. Parameters are $N = 6$, $K = 256$, $Z = 3$

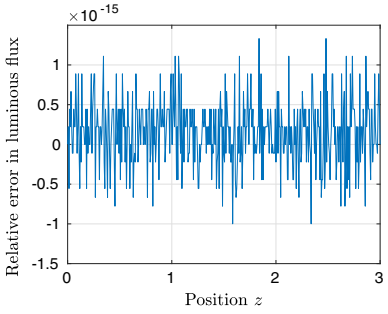


Fig. 7 Elliptic waveguide: relative error in luminous flux

The result using a 6th degree polynomial ($N = 6$) and $K = 16 \times 16 = 256$ elements is shown in Fig. 6, together with the initial condition. The numerical solution at $z = Z$ has roughly the same phase space area and is approximately a rotation of the initial condition. The scheme is also energy conservative up to machine precision, as can be seen in the plot of Fig. 7. Here, the relative error in the total luminous flux is plotted as a function of z. The total luminous flux is computed according to its definition (15) by applying a quadrature rule in agreement with the chosen polynomial degree.

Furthermore, a convergence test is performed for this example by changing the number of elements K and varying the polynomial degree from $N = 1, 2, ..., 6$. The numerical solution is compared to the exact solution, which can be found from the trajectory of the rays given by expressions (93). The expressions describe the evolution of a ray, given the initial conditions

Table 1 Elliptic waveguide: convergence data

K	e_{DG}	γ_{DG}	e_{DG}	γ_{DG}	e_{DG}	γ_{DG}
	$N=1$		$N=2$		$N=3$	
16	8.86e-03		5.47e-03		3.18e-03	
64	3.93e-03	1.17	1.71e-03	1.68	7.36e-04	2.11
256	1.62e-03	1.28	3.19e-04	2.43	3.28e-05	4.49
1024	4.27e-04	1.92	2.81e-05	3.51	2.07e-06	3.99
4096	8.32e-05	2.36	2.98e-06	3.24	1.21e-07	4.10
	$N=4$		$N=5$		$N=6$	
16	2.15e-03		1.17e-03		5.17e-04	
64	1.68e-04	3.68	5.82e-05	4.33	1.29e-05	5.32
256	6.37e-06	4.72	7.02e-07	6.37	9.89e-08	7.03
1024	1.56e-07	5.35	1.05e-08	6.06	6.51e-10	7.25
4096	4.34e-09	5.17	1.49e-10	6.14	4.67e-12	7.12

of the ray. For the analytical solution to Liouville's equation at an arbitrary z we want to know where the ray originated from, since the phase space coordinates on the mesh are known. This amounts to tracing the ray backwards starting from an arbitrary z to $z = 0$, resulting in

$$\rho(z, q, p) = \rho_0(q(-z), p(-z)), \tag{97}$$

with $q(z)$ and $p(z)$ given in (93).

Using the exact solution (97) we can evaluate the discretisation error for which we take the L^1-norm, i.e.,

$$e_{\mathrm{DG}} := \int_{\mathcal{P}} |\rho_{\mathrm{DG}}(Z, q, p) - \rho(Z, q, p)| \, \mathrm{d}q\mathrm{d}p, \tag{98}$$

where ρ_{DG} denotes the numerical solution and ρ denotes the exact solution (97). The integrals in (98) are evaluated using Gauss-Legendre quadrature with $N + 3$ nodes. The convergence order γ_{DG} is estimated from the empirical relation

$$e_{\mathrm{DG}} = C_{\mathrm{DG}} K^{-\gamma_{\mathrm{DG}}/2}, \tag{99}$$

with $C_{\mathrm{DG}} > 0$ an arbitrary constant.

The convergence data is shown in Table 1. The spatial discretisation is done using an Nth degree polynomial, and therefore the spatial order of accuracy is $N + 1$. The temporal discretisation is done using a 4th order explicit Runge-Kutta method, where we choose Δz to be the maximum allowable step such that the temporal integration is stable. Furthermore, a uniform rectangular mesh is used, where upon mesh refinement the mesh size in each direction is halved and similarly Δz is halved to ensure stability.

The global error depends on whether the spatial or temporal discretisation errors dominate. From Table 1, we observe that the spatial discretisation error dominates for the polynomial degrees $N = 1$ to $N = 6$. Choosing a smaller Δz-step in the numerical experiments did not influence the discretisation error. The results show that we obtain the expected $N + 1$ order of convergence.

5.2 Bucket of Water

To illustrate that the strategy outlined in Sect 4 for handling optical interfaces is energy conservative, we apply it to a test case. The test case 'bucket of water' introduced by van Lith et al. [32,33] is a suitable choice. The refractive index for this problem is given by

$$n(q) = \begin{cases} n_1, & \text{if } q \le 0, \\ n_2, & \text{if } q > 0, \end{cases} \tag{100}$$

where we take $n_1 = 1.4$ and $n_2 = 1$. Using an initial basic luminance ρ_0 that has support $\mathcal{D}$ where $q < 0$ and $p > 0$ for all $(q, p) \in \mathcal{D}$, the solution features both refraction and total internal reflection in two separate quadrants of phase space. The exact solution reads [33]

$$\rho(z, q, p) = \begin{cases} \rho_0\left(q - z\frac{p}{\sqrt{n_1^2 - p^2}}, p\right) & \text{if } q < 0,\ p \ge 0, \\ \rho_0\left(z\frac{p}{\sqrt{n_1^2 - p^2}} - q, -p\right) & \text{if } q < 0,\ -p_c < p < 0, \\ \rho_0\left((\delta z - z)\frac{\bar{p}}{\sqrt{n_1^2 - \bar{p}^2}}, \bar{p}\right) & \text{if } q > 0,\ p \ge 0, \\ 0 & \text{otherwise,} \end{cases} \tag{101a}$$

where $p_c = \sqrt{n_1^2 - n_2^2}$, $\bar{p} = -\mathcal{S}(-p; n_2, n_1, -\vec{\nu})$ with $\vec{\nu} = (-1, 0)$, and

$$\delta z = \frac{q}{p}\sqrt{n_2^2 - p^2}. \tag{101b}$$

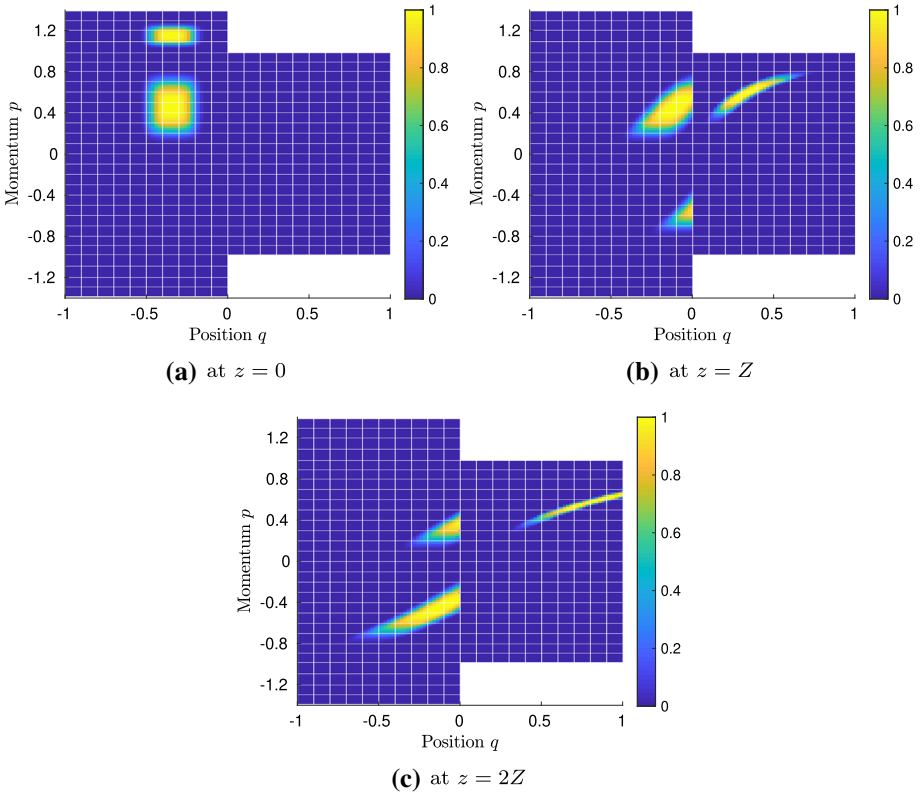
The region described by $\{q < 0, p \ge 0\}$ features propagation through the medium with refractive index n_1. The region $\{q < 0, -p_c < p < 0\}$ describes light that was reflected at the optical interface, and the region $\{q > 0, p \ge 0\}$ describes light that was refracted.

As an initial condition we use

$$\rho_0(q, p) := \varphi_m\left(\frac{q - q_0}{\sigma_q}\right)\left[\varphi_m\left(\frac{p - p_0}{\sigma_{p,0}}\right) + \varphi_m\left(\frac{p - p_1}{\sigma_{p,1}}\right)\right], \tag{102}$$

with φ_m defined in (95) and on the part of the boundary of the domain that is not on the optical interface, we prescribe $\rho = 0$ whenever the velocity field is pointing into the domain, otherwise we leave ρ free. Since the q-position is restricted to $q \in [-1, 1]$, this means that at $q = \pm 1$ we place virtual detectors that capture any luminous flux leaving the system. For the parameters in (102), we take $q_0 = -0.35$, $\sigma_q = 0.25$, $p_0 = 0.45$, $\sigma_{p,0} = 0.45$, $p_1 = \frac{1}{2}(1.3 + p_c)$ and $\sigma_{p,1} = 1.3 - p_1$. Furthermore, we take $m = 7$ unless specified otherwise.

Again, the explicit 4th order RK method from the previous example is used with a constant Δz-step as determined by the stability of the temporal integration. The numerical solution is integrated from $z = 0$ to $z = Z = 0.7$ and $z = 2Z$, and is shown in Fig. 8, together with the used initial condition. The result was obtained using a 6th degree polynomial ($N = 6$) and $K = 480$ elements. The mesh uses only rectangular elements and is almost uniform. To easily treat the optical interface, we have shifted the elements below and above the critical momentum $p_c \approx 0.98$ in the p-direction such that the critical momentum is aligned with the edges of these elements. The mesh spacings for $K = 480$ are $\Delta q = 0.1$ and $\Delta p \approx 0.1$.



(a) at $z = 0$ **(b)** at $z = Z$

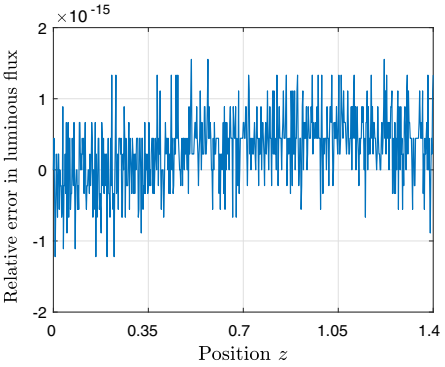
(c) at $z = 2Z$

Fig. 8 Bucket of water: basic luminance distributions $\rho(z, q, p)$. Parameters are $N = 6$, $K = 480$, $Z = 0.7$

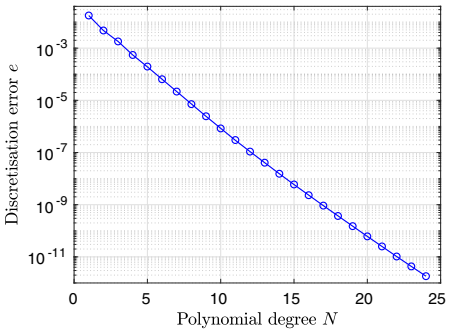
In Fig. 8 the quadrants featuring reflection and refraction can be clearly distinguished, while the solution is, as expected, perfectly discontinuous along the optical interface. Furthermore, at $z = 2Z$ some of the solution has passed $q = 1$, meaning some energy has hit the detectors. We observe that a total 7.5 % of the initial luminous flux has hit the detectors at $z = 2Z$. Taking into account the luminous flux on the detectors, we compute the relative error in the total luminous flux as a function of z which is plotted in Fig. 9a. The plot shows that the method obeys energy conservation up to machine precision.

Furthermore, to show that the optical interface treatment does not incur any penalty on the convergence order, we compute the discretisation error for this example defined in (98). The convergence data for $N = 1, 2, ..., 6$ is shown in Table 2. Also for this example, we observe that the spatial discretisation error is dominant and choosing smaller Δz-steps did not result in different discretisation errors. Moreover, the expected spatial order of convergence $N + 1$ is obtained.

Next, we verify the exponential convergence of DGSEM by increasing the polynomial degree, whilst keeping the number of elements fixed to $K = 1920$ and choosing $m = 28$ in (102). For temporal integration a fixed number of $2 \cdot 10^4$ z-steps are performed, chosen such that the temporal integration error does not interfere with the convergence test. The result is shown in Fig. 9b and exponential convergence is observed.



(a) relative error in the total luminous flux, computed with $N = 6$ and $K = 480$



(b) polynomial degree refinement with $K = 1920$ and $m = 28$

Fig. 9 Bucket of water

Table 2 Bucket of water: convergence data

K	e_{DG}	γ_{DG}	e_{DG}	γ_{DG}	e_{DG}	γ_{DG}
	$N = 1$		$N = 2$		$N = 3$	
480	4.93e-02		1.71e-02		8.70e-03	
1920	1.82e-02	1.44	4.90e-03	1.80	1.23e-03	2.83
7680	6.25e-03	1.54	6.61e-04	2.89	8.07e-05	3.92
30720	1.56e-03	2.00	5.82e-05	3.50	3.71e-06	4.44
	$N = 4$		$N = 5$		$N = 6$	
480	4.15e-03		2.03e-03		1.03e-03	
1920	3.55e-04	3.55	1.08e-04	4.24	3.36e-05	4.94
7680	1.17e-05	4.93	1.98e-06	5.77	3.79e-07	6.47
30720	3.08e-07	5.24	3.10e-08	6.00	3.37e-09	6.81

5.2.1 Comparison with Ray Tracing

We compare forward ray tracing with bin-counting to solving Liouville's equation using the DGSEM. Solving Liouville's equation already has two advantages, i.e., it conserves energy and provides a more complete picture due to solving the luminance instead of its integrated quantities, the illuminance or luminous intensity. The latter advantage also comes at a price of having to solve a two-dimensional problem in phase space followed by integration to compute these quantities. Ray tracing on the other hand can directly use bins on a one-dimensional grid to compute either the illuminance or luminous intensity.

For a fair comparison, we compute the illuminance E, defined by (17), for this test case using both quasi-Monte Carlo ray tracing and the DGSEM. For quasi-Monte Carlo ray tracing we fix the number of bins to $B = 1000$ and employ a uniform grid on $q \in [-1, 1]$, i.e.,

$$Q_j = (j-1)\Delta q - 1, \quad j = 1, ..., B+1, \tag{103}$$

with $\Delta q = \frac{2}{B}$. The jth bin is defined by $[Q_j, Q_{j+1}]$ with midpoint $q_j = \frac{1}{2}(Q_j + Q_{j+1})$. The global error for quasi-Monte Carlo integration using a 2D Sobol sequence behaves as

Table 3 Bucket of water: discretisation error using ray tracing (RT) for computing the illuminance. Number of bins is fixed to $B = 1000$

N_{RT} $(\cdot 10^6)$	e_{RT}	γ_{RT}	t_{RT}
0.04	1.49e-02		0.079 s
0.16	6.52e-03	0.59	0.295 s
0.64	2.46e-03	0.70	1.239 s
2.56	9.28e-04	0.70	3.996 s
10.24	3.58e-04	0.69	19.865 s
40.96	1.17e-04	0.81	1 min 19 s
163.84	3.50e-05	0.87	5 min 22 s
655.36	1.33e-05	0.70	21 min 36 s
2621.44	4.65e-06	0.76	1 h 26 min 6 s

$\mathcal{O}(\log(M)^2/M)$ with M the number of 2D points [11]. The 2D points are in our case the initial phase space coordinates $(q_i, p_i) \in \mathcal{P}$ of each ray. For more details on quasi-Monte Carlo integration, see [24]. In the bucket of water example $M = N_{\mathrm{RT}}$ denotes the number of rays and we use a fixed number of bins.

For the DGSEM we compute the luminance followed by integration such that we obtain the illuminance. Ray tracing defines an average illuminance on each bin, hence, for a fair comparison we also average the illuminance for the DGSEM when computing the discretisation error. For the discretisation error we take the L^1-norm and compare the numerical solution to the exact illuminance, which is computed by integrating the exact luminance (101) numerically up to machine precision.

Once again we take the initial condition (102) and (95) with $m = 7$. The illuminance computed using ray tracing with $N_{\mathrm{RT}} = 0.64 \cdot 10^6$ rays and the illuminance obtained with DGSEM on a mesh with $K = 480$ elements and $N = 4$ are shown in Fig. 10a, together with the exact solution. The ray tracing (RT) solution is noisy, which is inherent in the method due to the quasi-random Monte Carlo process, while the DGSEM solution is almost indistinguishable from the exact solution.

The discretisation error for ray tracing for increasing number of rays is shown in Table 3, while the results for the DGSEM with increasing number of elements K is shown in Table 4. In the tables e_{RT} and e_{DG} denote the errors for ray tracing and solving Liouville's equation using the DGSEM, respectively, while t_{RT} and t_{DG} denote their respective computation times using only a single core. Furthermore, γ_{RT} is estimated from the empirical relation

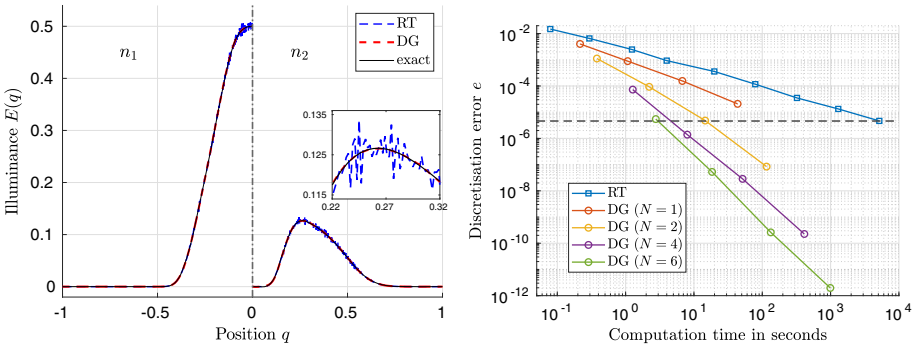
$$e_{\mathrm{RT}} = C_{\mathrm{RT}} N_{\mathrm{RT}}^{-\gamma_{\mathrm{RT}}}. \tag{104}$$

while γ_{DG} is estimated from the empirical relation (99).

From the tables we observe that ray tracing uses $2.62 \cdot 10^9$ rays taking almost an hour and a half, while the DGSEM achieves roughly the same accuracy in only 8.0 seconds when using 1920 elements. Varying the polynomial degrees results in the performance graph shown in Fig. 10b. It can be observed that the DGSEM always achieves a better accuracy for $N \geq 1$, compared to quasi-Monte Carlo ray tracing in the same amount of time. The DGSEM significantly outperforms ray tracing and, moreover, can achieve high accuracies in reasonable time.

Table 4 Bucket of water: discretisation error using the DGSEM (DG) with $N = 4$ for computing the illuminance

K	e_{DG}	γ_{DG}	t_{DG}
480	7.28e-05		1.271 s
1920	1.39e-06	5.71	7.998 s
7680	2.86e-08	5.60	51.524 s
30720	2.26e-10	6.98	6 min 52 s



(a) Illuminance $E(q)$ **(b)** Performance comparison

Fig. 10 Left: the illuminance computed using ray tracing (RT) with $N_{\mathrm{RT}} = 0.64 \cdot 10^6$ rays and the DGSEM with $K = 480$ and $N = 4$ (DG). Right: the error as a function of the computation time for both methods. Both results were computed at $z = Z = 0.7$

6 Conclusions

We have solved Liouville's equation for geometrical optics, using the discontinuous Galerkin spectral element method. For smooth refractive index fields the scheme obeys energy conservation by design. At optical interfaces Snell's law of refraction or the law of specular reflection needs to be applied. Together with the basic luminance invariance, this corresponds to non-local boundary conditions in phase space. A method was presented to treat the non-local boundary conditions along the optical interface such that the scheme remains energy conservative in the presence of optical interfaces.

Energy conservation is verified in an example. Moreover, in the same example the scheme was compared with forward ray tracing when computing the illuminance. Ray tracing uses bins on a one-dimensional grid to compute the illuminance, while the DGSEM has to solve a two-dimensional problem followed by integration. This still resulted in a better performance compared to ray tracing. In particular, for a fourth degree polynomial, the DGSEM has a computation time of 8.0 seconds, while ray tracing took 1 hour and 26 minutes to achieve almost the same accuracy.

At the moment Fresnel reflections [15,16] were not included. Therefore, an obvious next step will be to include this in the method by modifying the basic luminance invariance over an optical interface (53), see [26]. Moreover, only two-dimensional optics was considered in this paper. Hence, for future research we intend to extend the method to a three-dimensional optics settings. This requires a four-dimensional phase space together with the propagation along the z-coordinate, making it a five-dimensional problem. Despite the increased computational complexity due to the high dimensionality of the problem, the high convergence rates of

DGSEM will still be an advantage over ray tracing, where in theory the DGSEM might achieve a better performance [31].

Acknowledgements This work is part of the research programme NWO-TTW Perspectief with project number P15-36, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Funding This work is funded as a part of the research programme NWO-TTW Perspectief with project number P15-36, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Availability of Data Data will be made available on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code Availability Custom code was written in MATLAB and will not be made available online.

A Appendix

The matrix $\boldsymbol{A} \in \mathbb{R}^{(N+2)\times(N+2)}$ defined in equation (78) has a certain structure that allows for analytical computation of its determinant and inverse. Omitting the superscripts from (78) the matrix $\boldsymbol{A}$ reads

$$\boldsymbol{A} := \begin{pmatrix} \operatorname{diag}(\boldsymbol{w}) & \boldsymbol{\alpha} \circ \boldsymbol{w} \\ (\boldsymbol{\alpha} \circ \boldsymbol{w})^{\mathrm{T}} & 0 \end{pmatrix}, \tag{105}$$

which can be rewritten as

$$\boldsymbol{A} = \begin{pmatrix} \operatorname{diag}(\boldsymbol{w}) & \boldsymbol{0} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{pmatrix} \boldsymbol{Q} \tag{106a}$$

with $\boldsymbol{Q}$ defined by

$$\boldsymbol{Q} := \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\alpha} \\ (\boldsymbol{\alpha} \circ \boldsymbol{w})^{\mathrm{T}} & 0 \end{pmatrix}. \tag{106b}$$

The determinant of $\boldsymbol{A}$ is equal to product of the determinant of the diagonal matrix and the determinant of $\boldsymbol{Q}$. The determinant of $\boldsymbol{Q}$ can be found by Laplace (cofactor) expansion along the first row, i.e.,

$$\det(\boldsymbol{Q}) = \begin{vmatrix} 1 & 0 & \dots & \alpha_1 \\ 0 & 1 & \dots & \alpha_2 \\ \vdots & & \ddots & \vdots \\ \alpha_1 w_1 & \alpha_2 w_2 & \dots & 0 \end{vmatrix} + (-1)^{1+N+2} \alpha_0 \begin{vmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ \alpha_0 w_0 & \alpha_1 w_1 & \alpha_2 w_2 & \dots & \alpha_N w_N \end{vmatrix}. \tag{107}$$

The second term on the RHS can be easily evaluated using a cofactor expansion along the first column, since it has all zeros except for $\alpha_0 w_0$ and the remaining minor is the determinant of an identity matrix. Therefore, we obtain

$$(-1)^{1+N+2}\alpha_0 \begin{vmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ \alpha_0 w_0 & \alpha_1 w_1 & \alpha_2 w_2 & \dots & \alpha_N w_N \end{vmatrix} = (-1)^{1+N+2}\alpha_0(-1)^{1+N+1}\alpha_0 w_0 = -\alpha_0^2 w_0. \tag{108}$$

The first term on the RHS of (107) can again be expanded along the first row, resulting in

$$\begin{vmatrix} 1 & 0 & \dots & \alpha_1 \\ 0 & 1 & \dots & \alpha_2 \\ \vdots & & \ddots & \vdots \\ \alpha_1 w_1 & \alpha_2 w_2 & \dots & 0 \end{vmatrix} = \begin{vmatrix} 1 & 0 & \dots & \alpha_2 \\ 0 & 1 & \dots & \alpha_3 \\ \vdots & & \ddots & \vdots \\ \alpha_2 w_2 & \alpha_3 w_3 & \dots & 0 \end{vmatrix} + (-1)^{1+N+1}\alpha_1 \begin{vmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ \alpha_1 w_1 & \alpha_2 w_2 & \alpha_3 w_3 & \dots & \alpha_N w_N \end{vmatrix}$$
$$= \begin{vmatrix} 1 & 0 & \dots & \alpha_2 \\ 0 & 1 & \dots & \alpha_3 \\ \vdots & & \ddots & \vdots \\ \alpha_2 w_2 & \alpha_3 w_3 & \dots & 0 \end{vmatrix} - \alpha_1^2 w_1.$$

Repeating these steps we obtain the expression for r, defined in (80), i.e.,

$$r = \det(\boldsymbol{Q}) = -\sum_{i=0}^{N} \alpha_i^2 w_i, \tag{109}$$

so that

$$\det(\boldsymbol{A}) = -\sum_{i=0}^{N} \alpha_i^2 w_i \prod_{i=0}^{N} w_i. \tag{110}$$

The inverse of $\boldsymbol{A}$ can readily be found if the inverse of $\boldsymbol{Q}$ is known. The derivation of the inverse of $\boldsymbol{Q}$ is briefly outlined for a 3×3 matrix, as it can easily be extended to the generic case. We start with the augmented matrix

$$\left(\begin{array}{ccc|ccc} 1 & 0 & \alpha_0 & 1 & 0 & 0 \\ 0 & 1 & \alpha_1 & 0 & 1 & 0 \\ \alpha_0 w_0 & \alpha_1 w_1 & 0 & 0 & 0 & 1 \end{array}\right). \tag{111}$$

First, we subtract $\alpha_i w_i$ times the $(i+1)$th row from the last row, resulting in

$$\left(\begin{array}{ccc|ccc} 1 & 0 & \alpha_0 & 1 & 0 & 0 \\ 0 & 1 & \alpha_1 & 0 & 1 & 0 \\ 0 & 0 & r & -\alpha_0 w_0 & -\alpha_1 w_1 & 1 \end{array}\right).$$

Next, we subtract α_i / r times the $(i+1)$th row from the first two rows, and divide the last row by r, such that we obtain

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 + \alpha_0 \frac{\alpha_0 w_0}{r} & \alpha_0 \frac{\alpha_1 w_0}{r} & -\frac{\alpha_0}{r} \\ 0 & 1 & 0 & \alpha_1 \frac{\alpha_0 w_0}{r} & 1 + \alpha_1 \frac{\alpha_1 w_1}{r} & -\frac{\alpha_1}{r} \\ 0 & 0 & 1 & -\frac{\alpha_0 w_0}{r} & -\frac{\alpha_1 w_1}{r} & \frac{1}{r} \end{array}\right).$$

The inverse of $\boldsymbol{Q}$ now allows for easy evaluation of the inverse of $\boldsymbol{A} \in \mathbb{R}^{3\times 3}$, by taking inverse of $\boldsymbol{A}$ in (106a) resulting in

$$\begin{aligned}\boldsymbol{A}^{-1} &= \frac{1}{r}\begin{pmatrix} r+\alpha_0\alpha_0 w_0 & \alpha_0\alpha_1 w_1 & -\alpha_0 \\ \alpha_1\alpha_0 w_0 & r+\alpha_1\alpha_1 w_1 & -\alpha_1 \\ -\alpha_0 w_0 & -\alpha_1 w_1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{w_0} & & \\ & \frac{1}{w_1} & \\ & & 1 \end{pmatrix} \\ &= \frac{1}{r}\begin{pmatrix} \frac{r}{w_0}+\alpha_0^2 & \alpha_0\alpha_1 & -\alpha_0 \\ \alpha_1\alpha_0 & \frac{r}{w_1}+\alpha_1\alpha_1 & -\alpha_1 \\ -\alpha_0 & -\alpha_1 & 1 \end{pmatrix}.\end{aligned}$$

Similarly, for the general case we obtain

$$\boldsymbol{A}^{-1} = \frac{1}{r}\begin{pmatrix} \boldsymbol{B} & -\boldsymbol{\alpha} \\ -\boldsymbol{\alpha}^{\mathrm{T}} & 1 \end{pmatrix} \tag{112a}$$

with r defined in (109) and the coefficients of the matrix $\boldsymbol{B} = (B_{ij})$ read

$$B_{ij} := \begin{cases} \alpha_i^2 + \frac{r}{w_i} & \text{if } i = j, \\ \alpha_i\alpha_j & \text{if } i \neq j. \end{cases} \tag{112b}$$

References

1. Arnold, V.I.: Mathematical Methods of Classical Mechanics, vol. 60. Springer Science & Business Media, Berlin (2013)
2. Bahrami, M., Goncharov, A.V.: Geometry-invariant GRIN lens: finite ray tracing. Opt. Express **22**(23), 27797–27810 (2014)
3. Bernardi, C., Maday, Y., Landriani, G.S.: Noncorming matching conditions for coupling spectral and finite element methods. Appl. Numer. Math. **6**(1–2), 65–84 (1989)
4. Bernardi, C., Maday, Y., Patera, A. T.: Domain Decomposition by the Mortar Element Method. In: Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters, pages 269–286. Springer (1993)
5. Bui-Thanh, T., Ghattas, O.: Analysis of an hp-nonconforming discontinuous Galerkin spectral element method for wave propagation. SIAM J. Numer. Anal. **50**(3), 1801–1826 (2012)
6. Carpenter, M. H., Kennedy, C. A.: Fourth-order 2N-storage Runge-Kutta schemes. NASA TM 109112 (1994)
7. Chalmers, N., Krivodonova, L.: A robust CFL condition for the discontinuous Galerkin method on triangular meshes. J. Comput. Phys. **403**, 10905 (2020)
8. Chan, J., Hewett, R.J., Warburton, T.: Weight-adjusted discontinuous Galerkin methods: curvilinear meshes. SIAM J. Sci. Comput. **39**(6), A2395–A2421 (2017)
9. Chaves, J.: Introduction to Nonimaging Optics. CRC Press, Boca Raton (2017)
10. Cvetkovic, A., Dross, O., Chaves, J., Benitez, P., Miñano, J.C., Mohedano, R.: Etendue-preserving mixing and projection optics for high-luminance LEDs, applied to automotive headlamps. Opt. Express **14**(26), 13014 (2006)
11. Filosa, C.: Phase Space Ray Tracing for Illumination Optics. PhD thesis, Eindhoven University of Technology (2018)
12. Filosa, C., ten Thije Boonkkamp, J.H.M., Ijzerman, W.L.: Ray tracing method in phase space for two-dimensional optical systems. Appl. Op. **55**(13), 3599–3606 (2016)
13. Filosa, C., ten Thije Boonkkamp, J.H.M., Ijzerman, W.L.: Phase space ray tracing for a two-dimensional parabolic reflector. Math. Stat. **5**(4), 135–142 (2017)
14. Glassner, A.S.: An Introduction to Ray Tracing. Elsevier, Amsterdam (1989)
15. Griffiths, D. J.: Introduction to Electrodynamics (2005)
16. Hecht, E., et al.: Optics, vol. 4. Addison Wesley, San Francisco (2002)
17. Herkommer, A.M.: Phase space optics: an alternate approach to freeform optical systems. Opt. Eng. **53**(3), 031304 (2013)

18. Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications. Springer Science & Business Media, Berlin (2007)
19. Ketcheson, D.I.: Runge-Kutta methods with minimum storage implementations. J. Comput. Phys. **229**(5), 1763–1773 (2010)
20. Kopriva, D.A.: A conservative staggered-grid Chebyshev multidomain method for compressible flows. II. A semi-structured method. J. Comput. Phys. **128**(2), 475–488 (1996)
21. Kopriva, D.A.: Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers. Springer Science & Business Media, Berlin (2009)
22. Kopriva, D.A., Gassner, G.J.: Geometry effects in nodal discontinuous Galerkin methods on curved elements that are provably stable. Appl. Math. Comput. **272**, 274–290 (2016)
23. Kopriva, D.A., Woodruff, S.L., Hussaini, M.Y.: Computation of electromagnetic scattering with a nonconforming discontinuous spectral element method. Int. J. Numer. Meth. Eng. **53**(1), 105–122 (2002)
24. Leobacher, G., Pillichshammer, F.: Introduction to Quasi-Monte Carlo Integration and Applications. Springer, Berlin (2014)
25. McCluney, W.R.: Introduction to Radiometry and Photometry. Artech House, New York (2014)
26. Nicodemus, F.E.: Radiance. Am. J. Phys. **31**(5), 368–377 (1963)
27. Ohno, H.: Symplectic ray tracing based on Hamiltonian optics in gradient-index media. JOSA A **37**(3), 411–416 (2020)
28. Pelka, D.G., Patel, K.: An overview of LED applications for general illumination. Des. Eff. Illum. Syst. **5186**, 15–26 (2003)
29. Rausch, D., Rommel, M., Herkommer, A.M., Talpur, T.: Illumination design for extended sources based on phase space mapping. Op. Eng. **56**(6), 065103 (2017)
30. Toulorge, T., Desmet, W.: CFL conditions for Runge-Kutta discontinuous Galerkin methods on triangular grids. J. Comput. Phys. **230**(12), 4657–4678 (2011)
31. van Lith, B. S.: Principles of Computational Illumination Optics. PhD thesis, Eindhoven University of Technology (2017)
32. van Lith, B.S., ten Thije Boonkkamp, J.H.M., IJzerman, W.L.: Active flux schemes on moving meshes with applications to geometric optics. J. Comput. Phys. X **3**, 100030 (2019)
33. van Lith, B.S., ten Thije Boonkkamp, J.H.M., IJzerman, W.L., Tukker, T.W.: A novel scheme for Liouvilles equation with a discontinuous Hamiltonian and applications to geometrical optics. J. Sci. Comput. **68**(2), 739–771 (2016)
34. Williamson, J.: Low-storage Runge-Kutta schemes. J. Comput. Phys. **35**(1), 48–56 (1980)
35. Wolf, K.B.: Geometric Optics on Phase Space. Springer Science & Business Media, Berlin (2004)
36. Zhu, X., Zhu, Q., Wu, H., Chen, C.: Optical design of LED-based automotive headlamps. Op. Laser Technol. **45**, 262–266 (2013)
37. Zingg, D.W., Chisholm, T.T.: Runge-Kutta methods for linear ordinary differential equations. Appl. Numer. Math. **31**(2), 227–238 (1999)