1 Floating-point arithmetic

A computer floating-point system $\mathbf{F}$ is a subset of the real number system $\mathbf{R}$, denoted as $\mathbf{F} \subset \mathbf{R}$. A real number x in $\mathbf{F}$ can be represented as

$$x = \pm m \times \beta^{e-t}, \tag{1}$$

where m, β, e, and t are integers. β is the base (which is 2 throughout this paper), t represents the precision, and e represents the exponent.

According to IEEE-754 standard, 64-bit floats are represented with 1 sign bit, 11 bits for the exponent, and 52 bits for the significand. Since the floating-point numbers are normalized, the most significant bit is always 1 and is not stored. Therefore, the actual bits for the significand are 53. One particularly important number in scientific computation is the machine epsilon ϵ, which is defined as the distance between 1 and the next larger floating-point number, calculated using the formula 2^{1-t}. The unit roundoff u, is equal to half of the machine epsilon and is given by $u = \epsilon/2 = 2^{-53} \approx 1.11 \times 10^{-16}$. It is the most useful quantity associated with $\mathbf{F}$ and is ubiquitous in the world of rounding error analysis.

Let $\mathbf{F} \subset \mathbf{R}$ denote the set of all real numbers of the form (1). If $x \in \mathbf{R}$, then $fl(x)$ denotes the element in $\mathbf{F}$ that is closest to x. The transformation $x \to fl(x)$ is referred to as rounding, and there are theorems available for the analysis of rounding errors. These theorems serve as the important theoretical basis for deriving the error model of ray tracing in this paper.

Theorem 1 *If $x \in \mathbf{R}$ in the range of $\mathbf{F}$, then $fl(x) = x(1 \pm \delta)$, $|\delta| \leq u$.*

Theorem (1) shows that every real number x lying in the range of $\mathbf{F}$ can be approximated by an element of $\mathbf{F}$ with a relative error no larger than u.

Theorem 2 *If $x, y \in \mathbf{F}$, then $fl(x\ op\ y) = (x\ op\ y)(1 \pm \delta)$, $|\delta| \leq u$, op denotes one of the five arithmetic operators: addition, subtraction, multiplication, division, or square root.*

Theorem (2) states that for basic floating operations op, such as $+, -, \times, \div$, and $\sqrt{}$, the computed value of $fl(x\ op\ y)$ is approximated by $(x\ op\ y)$ with a relative error no larger than u.

Theorem 3 *If $\rho_i = \pm 1$ and $i = 1 : n$, when $nu < 1$ and $0 \leq \delta_i \leq u$, then $\prod_{i=1}^{n}(1 \pm \delta_i)^{\rho_i} = 1 \pm \theta_n$, where, $0 \leq \theta_n < \dfrac{nu}{1 - nu} = \gamma_n$.*

The γ_n notation in Theorem (3) will be used throughout this paper. An advantage of this approach is that quotients of $(1 \pm \epsilon)^n$ terms can also be bounded with the γ function, e.g., $\dfrac{(1 \pm \epsilon)^m}{(1 \pm \epsilon)^n}$, the interval is bounded by $(1 \pm \gamma_{m+n})$, instead of incorrectly $(1 \pm \epsilon)^{m-n}$. This γ_n notation enables us to derive a more concise ray tracing error model.

2 Ray tracing standard formula

2.1 Definitions

In the field of optical design, quadratic rotation surfaces, such as spherical, ellipsoidal, hyperbolic, and parabolic surfaces, are extensively applied and can be uniformly represented in the following form:

$$F(x, y, z) = x^2 + y^2 + pz^2 - 2rz = 0 \tag{2}$$

where p is the quadratic surface coefficient. The surface corresponding to $p < 0$ is a hyperboloid, while the surface corresponding to $p = 0$ is a paraboloid. For $0 < p < 1$, the corresponding surface is an ellipsoid, and for $p = 1$, the corresponding surface is a sphere.

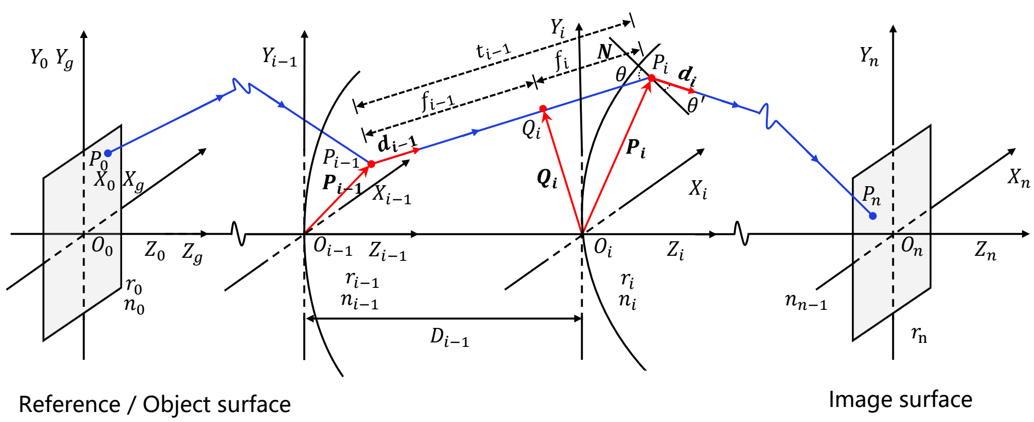

Figure 1: Physical significance of the quantities used in ray tracing.

Fig. 1 schematically illustrates the sequential disposition of the optical surfaces within the system, denoted as the 0th, 1st and nth surfaces from left to right, with O_0, O_1 and O_n designating the vertices of the respective surfaces. In the context of Fig. 1, subscripted data pertains to variables associated with a particular surface within the system or the adjacent space to the right of that surface. For example, for two adjacent $(i-1)$th and ith surfaces, r_{i-1} and r_i represent the curvature radii of those surfaces, respectively, while D_{i-1} denotes the distance between their vertices O_{i-1} and O_i. Unit vectors $\mathbf{d_{i-1}}$ and $\mathbf{d_i}$ point along the ray to the right of the $(i-1)$th and ith surfaces, respectively. Moreover, n_{i-1} and n_i represent the refractive indices to the right of these surfaces, and $\mathbf{P_{i-1}}$ and $\mathbf{P_i}$ are vectors from the vertices of the $(i-1)$th and ith surfaces to the points of incidence of the ray on the respective surfaces. $\mathbf{N}$ is the unit normal vector at point P_i on the ith surface. The angles θ and θ' correspond to the angles formed between the vectors $\mathbf{d_{i-1}}$ and $\mathbf{N}$, and between $\mathbf{d_i}$ and $\mathbf{N}$, respectively, both of these angles are acute in this context. It is important to note that $\mathbf{Q_i}$ is a vector originating from the vertex of the ith surface, perpendicular to the incident ray, and terminating at point Q_i on the ray. Additionally, t_{i-1} represents the length of the ray intercepted between the two refracting surfaces, where $t_{i-1} = f_{i-1} + f_i$.

2.2 Transfer Equations

To begin, light rays originate from the object space point P_0, collide with each surface in order from left to right and refract at the point of collision to proceed to the next surface. Eventually, the rays reach the image space point P_n. In the case of the incident light on the $(i-1)$th surface, the position $P_{i-1}(x_{i-1}, y_{i-1}, z_{i-1})$ and the unit direction vector $\mathbf{d_{i-1}} = L_{i-1}\mathbf{i} + M_{i-1}\mathbf{j}$ $+N_{i-1}\mathbf{k}$ of the incident light are available. We can calculate the subsequent refracted ray projection point $P_i(x_i, y_i, z_i)$, and the unit direction vector $\mathbf{d_i} = L_i\mathbf{i} + M_i\mathbf{j} + N_i\mathbf{k}$ on the ith surface, by applying transfer Eqs. (3)-(10) and refraction Eqs. (11)-(17). Here, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are unit vectors respectively along the X, Y and Z axes. In Eqs. (3)-(17), the variable Q_z denotes the z-coordinate of vector $\mathbf{Q_i}$, while the coefficient p_i represents the quadratic surface coefficient of the ith surface, $c_i = 1/r_i$.

$$f_{i-1} = D_{i-1}N_{i-1} - (x_{i-1}L_{i-1} + y_{i-1}M_{i-1} + z_{i-1}N_{i-1}) \tag{3}$$

$$Q_z = z_{i-1} + f_{i-1}N_{i-1} - D_{i-1} \tag{4}$$

$$Q_i^2 = (x_{i-1} + f_{i-1}L_{i-1})^2 + (y_{i-1} + f_{i-1}M_{i-1})^2 + (z_{i-1} + f_{i-1}N_{i-1} - D_{i-1})^2 \tag{5}$$

$$f_i = \frac{\triangle_1}{\triangle_4} - \frac{\sqrt{\triangle_2 - \triangle_3}}{\triangle_4} \tag{6}$$

$$where\ \triangle_1 = [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}]$$

$$\triangle_2 = [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}]^2$$

$$\triangle_3 = [1-(1-p_i)N_{i-1}^2][Q_i^2 - 2r_iQ_z - (1-p_i)Q_z^2]$$

$$\triangle_4 = 1-(1-p_i)N_{i-1}^2$$

$$t_{i-1} = f_{i-1} + f_i \tag{7}$$

$$x_i = x_{i-1} + t_{i-1}L_{i-1} \tag{8}$$

$$y_i = y_{i-1} + t_{i-1}M_{i-1} \tag{9}$$

$$z_i = z_{i-1} + t_{i-1}N_{i-1} - D_{i-1} \tag{10}$$

2.3 Refraction Equations

$$A = \sqrt{1 + c_i^2(1-p_i)(x_i^2 + y_i^2)} \tag{11}$$

$$\cos\theta = \left|\frac{(1 - z_ip_ic_i)N_{i-1} - x_ic_iL_{i-1} - y_ic_iM_{i-1}}{A}\right| \tag{12}$$

$$\cos\theta^{'} = \sqrt{1 - \frac{n_{i-1}^2}{n_i^2}(1 - \cos^2\theta)} \tag{13}$$

$$g_i = n_i\cos\theta^{'} - n_{i-1}\cos\theta \tag{14}$$

$$L_i = \frac{n_{i-1}}{n_i}L_{i-1} - \frac{g_ic_ix_i}{n_iA} \tag{15}$$

$$M_i = \frac{n_{i-1}}{n_i}M_{i-1} - \frac{g_ic_iy_i}{n_iA} \tag{16}$$

$$N_i = \frac{n_{i-1}}{n_i}N_{i-1} - \frac{g_i(p_ic_iz_i - 1)}{n_iA} \tag{17}$$

Eqs. (3)-(17) are collectively referred to as the standard formula for ray tracing. The correctness of the ray-tracing results is verified using the following validation Eqs. (18)-(19).

$$x_i^2 + y_i^2 + p_iz_i^2 - 2r_iz_i = 0 \tag{18}$$

$$L_i^2 + M_i^2 + N_i^2 = 1 \tag{19}$$

2.4 Derivation of ray tracing standard formula

The simple derivation of Eqs. (3)-(17) is as follows. A more comprehensive illustration detailed in references [1]-[2], although their primary focus is on the formula for spherical tracing.

1. To compute $\mathbf{P_i}$ from $\mathbf{P_{i-1}}$ and $\mathbf{d_{i-1}}$, begin by calculating $\mathbf{Q_i}$ from $\mathbf{P_{i-1}}$ and $\mathbf{d_{i-1}}$, then utilize $\mathbf{Q_i}$ and $\mathbf{d_{i-1}}$ to determine $\mathbf{P_i}$.

(1) Derive the vector formula $\mathbf{P_{i-1}} + f_{i-1}\mathbf{d_{i-1}} = D_{i-1}\mathbf{k} + \mathbf{Q_i}$ from the quadrilateral $O_{i-1}P_{i-1}Q_iO_i$. Taking the dot product of both sides of the equation with $\mathbf{d_{i-1}}$, we obtain $\mathbf{P_{i-1}} \cdot \mathbf{d_{i-1}} + f_{i-1} = D_{i-1}\mathbf{k} \cdot \mathbf{d_{i-1}}$, considering that $\mathbf{d_{i-1}} \cdot \mathbf{d_{i-1}} = 1$ and $\mathbf{Q_i} \cdot \mathbf{d_{i-1}} = 0$ (due to the perpendicularity of $\mathbf{d_{i-1}}$ and $\mathbf{Q_i}$). By substituting the components of $\mathbf{P_{i-1}}$ and $\mathbf{d_{i-1}}$ into the equation, we arrive at Eq. (3): $f_{i-1} = D_{i-1}N_{i-1} - (x_{i-1}L_{i-1} + y_{i-1}M_{i-1} + z_{i-1}N_{i-1})$. Substituting f_{i-1} into $\mathbf{P_{i-1}} + f_{i-1}\mathbf{d_{i-1}} = D_{i-1}\mathbf{k} + \mathbf{Q_i}$ leads to the derivation of Eqs. (4)-(5).

(2) Calculate $\mathbf{P_i}$ from $\mathbf{Q_i}$ and $\mathbf{d_{i-1}}$. From the triangle $\triangle O_iQ_iP_i$, we have the following vector equation: $\mathbf{P_i} = \mathbf{Q_i} + f_i\mathbf{d_{i-1}}$. The point P_i lies on the quadratic surface given by Eq. (2) : $x_i^2 + y_i^2 + p_iz_i^2 - 2r_iz_i = 0$. Expressing it in vector form, we obtain $\mathbf{P_i} \cdot \mathbf{P_i} - 2r_i\mathbf{k} \cdot \mathbf{P_i} - (1 - p_i)(\mathbf{k} \cdot \mathbf{P_i})^2 = 0$. By using $\mathbf{P_i} = \mathbf{Q_i} + f_i\mathbf{d_{i-1}}$ to substitute $\mathbf{P_i}$ in the previous equation with $\mathbf{Q_i}$, $\mathbf{d_{i-1}}$ and f_i, we can solve for f_i, yielding Eq. (6). From Fig. 1, we have $t_{i-1} = f_{i-1} + f_i$, expressed in Eq. (7). Substituting f_i into $\mathbf{P_i} = \mathbf{Q_i} + f_i\mathbf{d_{i-1}}$, we can solve for $\mathbf{P_i}$, resulting in scalar Eqs. (8)-(10).

2. The unit normal vector $\mathbf{N}$ at point P_i on the ith surface is determined. The expression for $\mathbf{N}$ is given as $\{\frac{-x_ic_i}{A}, \frac{-y_ic_i}{A}, \frac{1 - z_ip_ic_i}{A}\}$, by employing calculus principles, where $A = \sqrt{1 + c_i^2(1 - p_i)(x_i^2 + y_i^2)}$. As a result, Eq. (11) is obtained.

3. Calculate $\mathbf{d_i}$ using the law of refraction. The vector form of the law of refraction indicates that $(n_i\mathbf{d_i} - n_{i-1}\mathbf{d_{i-1}}) \times \mathbf{N} = 0$, which implies that $(n_i\mathbf{d_i} - n_{i-1}\mathbf{d_{i-1}})$ is parallel to $\mathbf{N}$, thus $(n_i\mathbf{d_i} - n_{i-1}\mathbf{d_{i-1}}) = g_i\mathbf{N}$, where g_i is a coefficient. By taking the dot product of the above equation with $\mathbf{N}$, we can obtain $g_i = (n_i\mathbf{d_i} \cdot \mathbf{N} - n_{i-1}\mathbf{d_{i-1}} \cdot \mathbf{N})$. According to the property that the dot product of two unit vectors is equal to the cosine of the angle between the vectors, we have the following formulas: $g_i = n_i \cos\theta' - n_{i-1}\cos\theta$, $\cos\theta = \left|\frac{(1 - z_ip_ic_i)N_{i-1} - x_ic_iL_{i-1} - y_ic_iM_{i-1}}{A}\right|$. Applying the law of refrac-

tion yields $\cos\theta' = \sqrt{1 - \frac{n_{i-1}^2}{n_i^2}(1-\cos^2\theta)}$. Hence, Eqs. (12)-(14) are derived as mentioned above. From equation $(n_i\mathbf{d_i} - n_{i-1}\mathbf{d_{i-1}}) = g_i\mathbf{N}$, we can derive $\mathbf{d_i} = \frac{n_{i-1}}{n_i}\mathbf{d_{i-1}} + \frac{g_i}{n_i}\mathbf{N}$, resulting in Eqs. (15)-(17) in scalar form.

3 Ray tracing error model

During the ray tracing process, given the starting point $P_{i-1} = x_{i-1}\mathbf{i} + y_{i-1}\mathbf{j} + z_{i-1}\mathbf{k}$ and directional vector $\mathbf{d_{i-1}} = L_{i-1}\mathbf{i} + M_{i-1}\mathbf{j} + N_{i-1}\mathbf{k}$ of a ray on the $(i-1)$th surface, the spacing D_{i-1} between the $(i-1)$th and ith surfaces, the curvature radius r_i and quadratic surface coefficient p_i of the ith surface, the intersection point $P_i = x_i\mathbf{i} + y_i\mathbf{j} + z_i\mathbf{k}$ of the ith surface with the ray is calculated using the transfer Eqs. (3)-(10), and the directional vector $\mathbf{d_i} = L_i\mathbf{i} + M_i\mathbf{j} + N_i\mathbf{k}$ is calculated using the refraction Eqs. (11)-(17). According to Theorem (1), the input physical quantities have the following representation errors.

$$fl(V_I) = V_I(1 \pm \delta),\ 0 \le \delta \le u, \tag{20}$$

where V_I is one of $x_{i-1}, y_{i-1}, z_{i-1}, L_{i-1}, M_{i-1}, N_{i-1}, D_{i-1}, r_i, p_i$. Meanwhile, when calculating the ray tracing projection point P_i and direction $\mathbf{d_i}$ in the order of Eqs. (3)-(17), it will introduce cumulative errors, such as the calculation of Q_z in the transfer Eq. (4) requires the result of the calculation of f_{i-1} in the transfer Eq. (3), and the calculation of f_{i-1} will introduce cumulative errors $(1 \pm \delta_{f_{i-1}})$, and there exists certain positive integer $n_{f_{i-1}}$ such that $f_{i-1}(1 \pm \delta_{f_{i-1}}) \le f_{i-1}(1 \pm u)^{n_{f_{i-1}}}$. In general, the following formula holds.

$$V_{cum}(1 \pm \delta_{V_{cum}}) \le V_{cum}(1 \pm u)^{n_{V_{cum}}} \tag{21}$$

where V_{cum} denotes any of $f_{i-1}, Q_z, Q_i^2, f_i, t_{i-1}, x_i, y_i, z_i, A, \cos\theta, \cos\theta', g_i$.

3.1 Transfer equation error model

3.1.1 Floating-Point Error for Equation 3

According to Theorem (1), the initial coordinates, directional vector of the light ray, and the distance between the $(i-1)$th surface and ith surface have floating-point representation errors and can be expressed as: $x_{i-1}(1 \pm u)$, $y_{i-1}(1 \pm u)$, $z_{i-1}(1 \pm u)$, $L_{i-1}(1 \pm u)$, $M_{i-1}(1 \pm u)$, $N_{i-1}(1 \pm u)$, and D_{i-1} $(1 \pm u)$, as described in Step 1. Additionally, according to Theorem (2), multiplying D_{i-1} and N_{i-1} introduces rounding errors, resulting in $D_{i-1}(1\pm u)$ $N_{i-1}(1 \pm u)(1 \pm u)$, as shown in Step 2. Similarly, there are comparable conclusions for $x_{i-1}L_{i-1}$, $y_{i-1}M_{i-1}$, and $z_{i-1}N_{i-1}$, as described in Step 2. Adding $x_{i-1}L_{i-1}+$ $y_{i-1}M_{i-1}$ generates a truncation error, and further adding it to $z_{i-1}N_{i-1}$ results in an additional truncation error, as shown in Step 3. The same technique is also used in Step 4. By using Theorem (3), we can obtain the result of Step 5.

$$
\begin{aligned}
&fl(f_{i-1}) = D_{i-1}N_{i-1} - (x_{i-1}L_{i-1} + y_{i-1}M_{i-1} + z_{i-1}N_{i-1})\\
&\subset D_{i-1}(1 \pm u)N_{i-1}(1 \pm u) - [x_{i-1}(1 \pm u)L_{i-1}(1 \pm u)\\
&+ y_{i-1}(1 \pm u)M_{i-1}i(1 \pm u) + z_{i-1}(1 \pm u)N_{i-1}(1 \pm u)] \rhd \textbf{Step 1}\\
&\subset D_{i-1}(1 \pm u)N_{i-1}(1 \pm u)(1 \pm u) - [x_{i-1}(1 \pm u)L_{i-1}(1 \pm u)(1 \pm u)\\
&+ y_{i-1}(1 \pm u)M_{i-1}(1 \pm u)(1 \pm u) + z_{i-1}(1 \pm u)N_{i-1}(1 \pm u)(1 \pm u)] \rhd \textbf{Step 2}\\
&\subset D_{i-1}N_{i-1}(1 \pm u)^3 - (x_{i-1}L_{i-1} + y_{i-1}M_{i-1} + z_{i-1}N_{i-1}i)(1 \pm u)^3\\
&\subset D_{i-1}N_{i-1}(1 \pm u)^3\\
&- [(x_{i-1}L_{i-1} + y_{i-1}M_{i-1})(1 \pm u) + z_{i-1}N_{i-1}](1 \pm u)(1 \pm u)^3 \rhd \textbf{Step 3}\\
&\subset D_{i-1}N_{i-1}(1 \pm u)^3 - [(x_{i-1}L_{i-1} + y_{i-1}M_{i-1})(1 \pm u) + z_{i-1}N_{i-1}](1 \pm u)^4\\
&\subset [D_{i-1}N_{i-1}i(1 \pm u)^3\\
&- [(x_{i-1}L_{i-1} + y_{i-1}M_{i-1})(1 \pm u) + z_{i-1}N_{i-1}](1 \pm u)^4](1 \pm u) \rhd \textbf{Step 4}\\
&\subset D_{i-1}N_{i-1}(1 \pm u)^4 - x_{i-1}L_{i-1}(1 \pm u)^6 - y_{i-1}M_{i-1}(1 \pm u)^6 - z_{i-1}N_{i-1}(1 \pm u)^5\\
&\subset D_{i-1}N_{i-1}(1 \pm \gamma_4) - x_{i-1}L_{i-1}(1 \pm \gamma_6)\\
&- y_{i-1}M_{i-1}(1 \pm \gamma_6) - z_{i-1}N_{i-1}(1 \pm \gamma_5) \rhd \textbf{Step 5}\\
&= D_{i-1}N_{i-1} - (x_{i-1}L_{i-1} + y_{i-1}M_{i-1} + z_{i-1}N_{i-1})\\
&+ [\pm x_{i-1}L_{i-1}\gamma_6 \pm y_{i-1}M_{i-1}\gamma_6 \pm z_{i-1}N_{i-1}\gamma_5 \pm D_{i-1}N_{i-1}\gamma_4] \qquad (22)
\end{aligned}
$$

Therefore, the error term (in square brackets) of f_{i-1} is bounded by $err_{f_{i-1}} = |x_{i-1}L_{i-1}|\gamma_6 + |y_{i-1}M_{i-1}|\gamma_6 + |z_{i-1}N_{i-1}|\gamma_5 + |D_{i-1}N_{i-1}|\gamma_4$.

3.1.2 Floating-Point Error for Equation 4

$$
\begin{aligned}
&fl(Q_z) = z_{i-1} + f_{i-1}N_{i-1} - D_{i-1}\\
&\subset z_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)N_{i-1}(1 \pm u) - D_{i-1}(1 \pm u)
\end{aligned}
$$

$$
\begin{aligned}
&\subset [[z_{i-1}(1 \pm u) + [f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)N_{i-1}(1 \pm u)](1 \pm u)](1 \pm u) \\
&\quad - D_{i-1}(1 \pm u)](1 \pm u) \\
&\subset z_{i-1}(1 \pm u)^3 + f_{i-1}N_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^5 - D_{i-1}(1 \pm u)^2 \\
&\subset z_{i-1}(1 \pm u)^3 + f_{i-1}N_{i-1}(1 \pm u)^{5+n_{f_{i-1}}} - D_{i-1}(1 \pm u)^2 \\
&\subset z_{i-1}(1 \pm \gamma_3) + f_{i-1}N_{i-1}(1 \pm \gamma_{5+n_{f_{i-1}}}) - D_{i-1}(1 \pm \gamma_2) \\
&= z_{i-1} + f_{i-1}N_{i-1} - D_{i-1} + [\pm z_{i-1}\gamma_3 \pm f_{i-1}N_{i-1}\gamma_{5+n_{f_{i-1}}} \pm D_{i-1}\gamma_2] \qquad (23)
\end{aligned}
$$

As a result, the error term (in square brackets) of Q_z is bounded by

$$
err_{Q_z} = |z_{i-1}|\gamma_3 + |f_{i-1}N_{i-1}|\gamma_{5+n_{f_{i-1}}} + |D_{i-1}|\gamma_2 \qquad (24)
$$

3.1.3 Floating-Point Error for Equation 5

$$
\begin{aligned}
&fl(Q_i^2) = (x_{i-1} + f_{i-1}L_{i-1})^2 + (y_{i-1} + f_{i-1}M_{i-1})^2 + (z_{i-1} + f_{i-1}N_{i-1} - D_{i-1})^2 \\
&\subset [x_{i-1} + f_{i-1}(1 \pm \delta_{f_{i-1}})L_{i-1}]^2 + [y_{i-1} + f_{i-1}(1 \pm \delta_{f_{i-1}})M_{i-1}]^2 \\
&+ [z_{i-1} + f_{i-1}(1 \pm \delta_{f_{i-1}})N_{i-1} - D_{i-1}]^2 \\
&\subset [x_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)L_{i-1}(1 \pm u)]^2 \\
&+ [y_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)M_{i-1}(1 \pm u)]^2 \\
&+ [z_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)N_{i-1}(1 \pm u) - D_{i-1}(1 \pm u)]^2 \\
&\subset [[x_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})L_{i-1}(1 \pm u)^3](1 \pm u)]^2(1 \pm u) \\
&+ [[y_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})M_{i-1}(1 \pm u)^3](1 \pm u)]^2(1 \pm u) \\
&+ [[[z_{i-1}(1 \pm u) + f_{i-1}(1 \pm \delta_{f_{i-1}})N_{i-1}(1 \pm u)^3](1 \pm u) - D_{i-1}(1 \pm u)](1 \pm u)]^2(1 \pm u) \\
&\subset [x_{i-1} + f_{i-1}L_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^2]^2(1 \pm u)^5 \\
&+ [y_{i-1} + f_{i-1}M_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^2]^2(1 \pm u)^5 \\
&+ [(z_{i-1} + f_{i-1}N_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^2](1 \pm u) - D_{i-1}]^2(1 \pm u)^5 \\
&\subset x_{i-1}^2(1 \pm u)^5 + 2x_{i-1}f_{i-1}L_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^7 + f_{i-1}^2L_{i-1}^2(1 \pm \delta_{f_{i-1}})^2(1 \pm u)^9 \\
&+ y_{i-1}^2(1 \pm u)^5 + 2y_{i-1}f_{i-1}M_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^7 + f_{i-1}^2M_{i-1}^2(1 \pm \delta_{f_{i-1}})^2(1 \pm u)^9 \\
&+ z_{i-1}^2(1 \pm u)^7 + 2z_{i-1}f_{i-1}N_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^9 + f_{i-1}^2N_{i-1}^2(1 \pm \delta_{f_{i-1}})^2(1 \pm u)^{11} \\
&- 2D_{i-1}z_{i-1}(1 \pm u)^6 - 2D_{i-1}f_{i-1}N_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u)^9 + D_{i-1}^2(1 \pm u)^5 \\
&\subset x_{i-1}^2(1 \pm u)^5 + 2x_{i-1}f_{i-1}L_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u)^7 + f_{i-1}^2L_{i-1}^2(1 \pm u)^{2n_{f_{i-1}}}(1 \pm u)^9 \\
&+ y_{i-1}^2(1 \pm u)^5 + 2y_{i-1}f_{i-1}M_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u)^7 + f_{i-1}^2M_{i-1}^2(1 \pm u)^{2n_{f_{i-1}}}(1 \pm u)^9 \\
&+ z_{i-1}^2(1 \pm u)^7 + 2z_{i-1}f_{i-1}N_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u)^9 + f_{i-1}^2N_{i-1}^2(1 \pm u)^{2n_{f_{i-1}}}(1 \pm u)^{11}
\end{aligned}
$$

$$
\begin{aligned}
&- 2D_{i-1}z_{i-1}(1 \pm u)^6 - 2D_{i-1}f_{i-1}N_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u)^9 + D_{i-1}^2(1 \pm u)^5 \\
&\subset x_{i-1}^2(1 \pm u)^5 + 2x_{i-1}f_{i-1}L_{i-1}(1 \pm u)^{7+n_{f_{i-1}}} + f_{i-1}^2L_{i-1}^2(1 \pm u)^{9+2n_{f_{i-1}}} \\
&+ y_{i-1}^2(1 \pm u)^5 + 2y_{i-1}f_{i-1}M_{i-1}(1 \pm u)^{7+n_{f_{i-1}}} + f_{i-1}^2M_{i-1}^2(1 \pm u)^{9+2n_{f_{i-1}}} \\
&+ z_{i-1}^2(1 \pm u)^7 + 2z_{i-1}f_{i-1}N_{i-1}(1 \pm u)^{9+n_{f_{i-1}}} + f_{i-1}^2N_{i-1}^2(1 \pm u)^{11+2n_{f_{i-1}}} \\
&- 2D_{i-1}z_{i-1}(1 \pm u)^6 - 2D_{i-1}f_{i-1}N_{i-1}(1 \pm u)^{9+n_{f_{i-1}}} + D_{i-1}^2(1 \pm u)^5 \\
&\subset x_{i-1}^2(1 \pm \gamma_5) + 2x_{i-1}f_{i-1}L_{i-1}(1 \pm \gamma_{7+n_{f_{i-1}}}) + f_{i-1}^2L_{i-1}^2(1 \pm \gamma_{9+2n_{f_{i-1}}}) \\
&+ y_{i-1}^2(1 \pm \gamma_5) + 2y_{i-1}f_{i-1}M_{i-1}(1 \pm \gamma_{7+n_{f_{i-1}}}) + f_{i-1}^2M_{i-1}^2(1 \pm \gamma_{9+2n_{f_{i-1}}}) \\
&+ z_{i-1}^2(1 \pm \gamma_7) + 2z_{i-1}f_{i-1}N_{i-1}(1 \pm \gamma_{9+n_{f_{i-1}}}) + f_{i-1}^2N_{i-1}^2(1 \pm \gamma_{11+2n_{f_{i-1}}}) \\
&- 2D_{i-1}z_{i-1}(1 \pm \gamma_6) - 2D_{i-1}f_{i-1}N_{i-1}(1 \pm \gamma_{9+n_{f_{i-1}}}) + D_{i-1}^2(1 \pm \gamma_5) \\
&= (x_{i-1} + f_{i-1}L_{i-1})^2 + (y_{i-1} + f_{i-1}M_{i-1})^2 + (z_{i-1} + f_{i-1}N_{i-1} - D_{i-1})^2 \\
&[\pm(x_{i-1}^2 + y_{i-1}^2 + D_{i-1}^2)\gamma_5 \pm 2x_{i-1}f_{i-1}L_{i-1}\gamma_{7+n_{f_{i-1}}} \pm 2y_{i-1}f_{i-1}M_{i-1}\gamma_{7+n_{f_{i-1}}} \\
&\pm 2z_{i-1}f_{i-1}N_{i-1}\gamma_{9+n_{f_{i-1}}} \pm 2D_{i-1}f_{i-1}N_{i-1}\gamma_{9+n_{f_{i-1}}} \\
&\pm 2D_{i-1}z_{i-1}\gamma_6 \pm f_{i-1}^2L_{i-1}^2\gamma_{9+2n_{f_{i-1}}} \pm f_{i-1}^2M_{i-1}^2\gamma_{9+2n_{f_{i-1}}} \\
&\pm z_{i-1}^2\gamma_7 \pm f_{i-1}^2N_{i-1}^2\gamma_{11+2n_{f_{i-1}}}]
\end{aligned} \tag{25}
$$

Consequently, the error term (in square brackets) of Q_i^2 is bounded by

$$
\begin{aligned}
&err_{Q_i^2} = (x_{i-1}^2 + y_{i-1}^2 + D_{i-1}^2)\gamma_5 + 2|x_{i-1}f_{i-1}L_{i-1}|\gamma_{7+n_{f_{i-1}}} \\
&+ 2|y_{i-1}f_{i-1}M_{i-1}|\gamma_{7+n_{f_{i-1}}} + 2|z_{i-1}f_{i-1}N_{i-1}|\gamma_{9+n_{f_{i-1}}} \\
&+ 2|D_{i-1}f_{i-1}N_{i-1}|\gamma_{9+n_{f_{i-1}}} + 2|D_{i-1}z_{i-1}|\gamma_6 + f_{i-1}^2L_{i-1}^2\gamma_{9+2n_{f_{i-1}}} \\
&+ f_{i-1}^2M_{i-1}^2\gamma_{9+2n_{f_{i-1}}} + z_{i-1}^2\gamma_7 + f_{i-1}^2N_{i-1}^2\gamma_{11+2n_{f_{i-1}}}]
\end{aligned} \tag{26}
$$

3.1.4 Floating-Point Error for Equation 6

To facilitate the derivation of the error model, the following intermediate variables are introduced.

$$\triangle_1 = [N_{i-1}r_i + (1 - p_i)Q_zN_{i-1}] \tag{27}$$

$$\triangle_2 = [N_{i-1}r_i + (1 - p_i)Q_zN_{i-1}]^2 \tag{28}$$

$$\triangle_3 = [1 - (1 - p_i)N_{i-1}^2][Q_i^2 - 2r_iQ_z - (1 - p_i)Q_z^2] \tag{29}$$

$$\triangle_4 = 1 - (1 - p_i)N_{i-1}^2 \tag{30}$$

Thus the transfer Eq. (31) can be expressed as:

$$f_i = \frac{\triangle_1}{\triangle_4} - \frac{\sqrt{\triangle_2 - \triangle_3}}{\triangle_4} \tag{31}$$

First compute the $\triangle_1$ error expression.

$$\begin{aligned}
&fl(\triangle_1) = [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}] \\
&\subset [N_{i-1}r_i(1\pm u)^3 + [1-p_i(1\pm u)](1\pm u)Q_z(1\pm\delta_{Q_z})N_{i-1}(1\pm u)^4] \\
&\subset \{[N_{i-1}r_i(1\pm u)^3 \\
&+ [1-p_i(1\pm u)](1\pm u)Q_z(1\pm u)^{n_{Q_z}}N_{i-1}(1\pm u)^4]\}(1\pm u) \\
&\subset [N_{i-1}r_i(1\pm u)^4 + [1-p_i(1\pm u)]Q_zN_{i-1}(1\pm u)^{6+n_{Q_z}}] \\
&\subset [N_{i-1}r_i(1\pm u)^4 + Q_zN_{i-1}(1\pm u)^{6+n_{Q_z}} - p_iQ_zN_{i-1}(1\pm u)^{7+n_{Q_z}}] \\
&\subset [N_{i-1}r_i(1\pm\gamma_4) + Q_zN_{i-1}(1\pm\gamma_{6+n_{Q_z}}) - p_iQ_zN_{i-1}(1\pm\gamma_{7+n_{Q_z}})] \\
&= [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}] + [\pm N_{i-1}r_i\gamma_4 \pm Q_zN_{i-1}\gamma_{6+n_{Q_z}} \pm p_iQ_zN_{i-1}\gamma_{7+n_{Q_z}}]
\end{aligned} \tag{32}$$

Thus, the $\triangle_1$ error expression is:

$$err_{\triangle_1} = |N_{i-1}r_i|\gamma_4 + |Q_zN_{i-1}|\gamma_{6+n_{Q_z}} + |p_iQ_zN_{i-1}|\gamma_{7+n_{Q_z}} \tag{33}$$

The $\triangle_2$ error expression is derived as follows:

$$\begin{aligned}
&fl(\triangle_2) = [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}]^2 \\
&= N_{i-1}^2r_i^2 + 2N_{i-1}r_iQ_zN_{i-1} - 2p_iN_{i-1}r_iQ_zN_{i-1} + Q_z^2N_{i-1}^2 \\
&- 2p_iQ_z^2N_{i-1}^2 + p_i^2Q_z^2N_{i-1}^2 \\
&\subset N_{i-1}^2r_i^2(1\pm u)^{4+3} + 2N_{i-1}r_iQ_zN_{i-1}(1\pm u)^{4+4}(1\pm\delta_{Q_z}) \\
&- 2p_iN_{i-1}r_iQ_zN_{i-1}(1\pm u)^{5+5}(1\pm\delta_{Q_z}) + Q_z^2N_{i-1}^2(1\pm u)^{4+3}(1\pm\delta_{Q_z})^2 \\
&- 2p_iQ_z^2N_{i-1}^2(1\pm u)^{5+5}(1\pm\delta_{Q_z})^2 + p_i^2Q_z^2N_{i-1}^2(1\pm u)^{6+5}(1\pm\delta_{Q_z})^2 \\
&\subset N_{i-1}^2r_i^2(1\pm u)^{7+5} + 2N_{i-1}r_iQ_zN_{i-1}(1\pm u)^{8+5+n_{Q_z}} \\
&- 2p_iN_{i-1}r_iQ_zN_{i-1}(1\pm u)^{10+4+n_{Q_z}} + Q_z^2N_{i-1}^2(1\pm u)^{7+3+2n_{Q_z}} \\
&- 2p_iQ_z^2N_{i-1}^2(1\pm u)^{10+2+2n_{Q_z}} + p_i^2Q_z^2N_{i-1}^2(1\pm u)^{11+1+2n_{Q_z}} \\
&\subset N_{i-1}^2r_i^2(1\pm\gamma_{12}) + 2N_{i-1}r_iQ_zN_{i-1}(1\pm\gamma_{13+n_{Q_z}}) \\
&- 2p_iN_{i-1}r_iQ_zN_{i-1}(1\pm\gamma_{14+n_{Q_z}}) + Q_z^2N_{i-1}^2(1\pm\gamma_{10+2n_{Q_z}}) \\
&- 2p_iQ_z^2N_{i-1}^2(1\pm\gamma_{12+2n_{Q_z}}) + p_i^2Q_z^2N_{i-1}^2(1\pm\gamma_{12+2n_{Q_z}}) \\
&= [N_{i-1}r_i + (1-p_i)Q_zN_{i-1}]^2
\end{aligned}$$

$$+ [\pm N_{i-1}^2 r_i^2 \gamma_{12} \pm 2N_{i-1} r_i Q_z N_{i-1} \gamma_{13+n_{Q_z}} \pm 2p_i N_{i-1} r_i Q_z N_{i-1} \gamma_{14+n_{Q_z}}$$
$$\pm Q_z^2 N_{i-1}^2 \gamma_{10+2n_{Q_z}} \pm 2p_i Q_z^2 N_{i-1}^2 \gamma_{12+2n_{Q_z}} \pm p_i^2 Q_z^2 N_{i-1}^2 \gamma_{12+2n_{Q_z}}] \quad (34)$$

Therefore, the $\triangle_2$ error expression is:

$$err_{\triangle_2} = N_{i-1}^2 r_i^2 \gamma_{12} + 2|N_{i-1} r_i Q_z N_{i-1}| \gamma_{13+n_{Q_z}} + 2|p_i N_{i-1} r_i Q_z N_{i-1}| \gamma_{14+n_{Q_z}}$$
$$+ Q_z^2 N_{i-1}^2 \gamma_{10+2n_{Q_z}} + 2|p_i| Q_z^2 N_{i-1}^2 \gamma_{12+2n_{Q_z}} + p_i^2 Q_z^2 N_{i-1}^2 \gamma_{12+2n_{Q_z}} \quad (35)$$

The $\triangle_3$ error expression is:

$$fl(\triangle_3) = [1 - (1-p_i)N_{i-1}^2][Q_i^2 - 2r_i Q_z - (1-p_i)Q_z^2]$$
$$\subset Q_i^2(1\pm u)^3(1\pm\delta_{Q_i})^2 - 2r_i Q_z(1\pm u)^4(1\pm\delta_{Q_z}) - Q_z(1\pm u)(1\pm\delta_{Q_z})$$
$$+ p_i Q_z^2(1\pm u)^5(1\pm\delta_{Q_i})^2 - N_{i-1}^2 Q_i^2(1\pm u)^{4+3}(1\pm\delta_{Q_i})^2$$
$$+ 2r_i N_{i-1}^2 Q_z(1\pm u)^{4+4}(1\pm\delta_{Q_z}) + N_{i-1}^2 Q_z^2(1\pm u)^{4+3}(1\pm\delta_{Q_z})^2$$
$$- p_i N_{i-1}^2 Q_z^2(1\pm u)^{5+4}(1\pm\delta_{Q_z})^2 + p_i N_{i-1}^2 Q_i^2(1\pm u)^{5+4}(1\pm\delta_{Q_i})^2$$
$$- 2p_i r_i N_{i-1}^2 Q_z(1\pm u)^{5+5}(1\pm\delta_{Q_z}) - p_i N_{i-1}^2 Q_z(1\pm u)^{5+4}(1\pm\delta_{Q_z})^2$$
$$+ p_i^2 N_{i-1}^2 Q_z(1\pm u)^{6+5}(1\pm\delta_{Q_z})^2$$
$$\subset Q_i^2(1\pm u)^{3+2n_{Q_i}} - 2r_i Q_z(1\pm u)^{4+n_{Q_z}} - Q_z(1\pm u)^{1+n_{Q_z}}$$
$$+ p_i Q_z^2(1\pm u)^{5+2n_{Q_i}} - N_{i-1}^2 Q_i^2(1\pm u)^{7+2n_{Q_i}}$$
$$+ 2r_i N_{i-1}^2 Q_z(1\pm u)^{8+n_{Q_z}} + N_{i-1}^2 Q_z^2(1\pm u)^{7+2n_{Q_i}}$$
$$- p_i N_{i-1}^2 Q_z^2(1\pm u)^{9+2n_{Q_z}} + p_i N_{i-1}^2 Q_i^2(1\pm u)^{9+2n_{Q_i}}$$
$$- 2p_i r_i N_{i-1}^2 Q_z(1\pm u)^{10+n_{Q_z}} - p_i N_{i-1}^2 Q_z(1\pm u)^{9+2n_{q_z}}$$
$$+ p_i^2 N_{i-1}^2 Q_z(1\pm u)^{11+2n_{Q_z}}$$
$$\subset Q_i^2(1\pm\gamma_{3+2n_{Q_i}}) - 2r_i Q_z(1\pm\gamma_{4+n_{Q_z}})$$
$$- Q_z(1\pm\gamma_{1+n_{Q_z}}) + p_i Q_z^2(1\pm\gamma_{5+2n_{Q_i}})$$
$$- N_{i-1}^2 Q_i^2(1\pm\gamma_{7+2n_{Q_i}}) + 2r_i N_{i-1}^2 Q_z(1\pm\gamma_{8+n_{Q_z}})$$
$$+ N_{i-1}^2 Q_z^2(1\pm\gamma_{7+2n_{Q_i}}) - p_i N_{i-1}^2 Q_z^2(1\pm\gamma_{9+2n_{Q_z}})$$
$$+ p_i N_{i-1}^2 Q_i^2(1\pm\gamma_{9+2n_{Q_i}}) - 2p_i r_i N_{i-1}^2 Q_z(1\pm\gamma_{10+n_{Q_z}})$$
$$- p_i N_{i-1}^2 Q_z(1\pm\gamma_{9+2n_{q_z}}) + p_i^2 N_{i-1}^2 Q_z(1\pm\gamma_{11+2n_{Q_z}})$$
$$= [1 - (1-p_i)N_{i-1}^2][Q_i^2 - 2r_i Q_z - (1-p_i)Q_z^2]$$
$$+ [\pm Q_i^2 \gamma_{3+2n_{Q_i}} \pm 2r_i Q_z \gamma_{4+n_{Q_z}} \pm Q_z \gamma_{1+n_{Q_z}} \pm p_i Q_z^2 \gamma_{5+2n_{Q_i}}$$
$$\pm N_{i-1}^2 Q_i^2 \gamma_{7+2n_{Q_i}} \pm 2r_i N_{i-1}^2 Q_z \gamma_{8+n_{Q_z}} \pm N_{i-1}^2 Q_z^2 \gamma_{8+2n_{Q_i}}$$
$$\pm p_i N_{i-1}^2 Q_z^2 \gamma_{9+2n_{Q_z}} \pm p_i N_{i-1}^2 Q_i^2 \gamma_{9+2n_{Q_i}} \pm 2p_i r_i N_{i-1}^2 Q_z \gamma_{10+n_{Q_z}}$$

$$\pm p_i N_{i-1}^2 Q_z \gamma_{9+2n_{Q_z}} \pm p_i^2 N_{i-1}^2 Q_z \gamma_{11+2n_{Q_z}}] \tag{36}$$

Therefore, the $\triangle_3$ error expression is:

$$\begin{aligned} err_{\triangle_3} &= Q_i^2 \gamma_{3+2n_{Q_i}} + 2|r_i Q_z| \gamma_{4+n_{Q_z}} + |Q_z| \gamma_{1+n_{Q_z}} + |p_i| Q_z^2 \gamma_{5+2n_{Q_i}} \\ &+ N_{i-1}^2 Q_i^2 \gamma_{7+2n_{Q_i}} + 2|r_i N_{i-1}^2 Q_z| \gamma_{8+n_{Q_z}} + N_{i-1}^2 Q_z^2 \gamma_{8+2n_{Q_i}} \\ &+ |p_i| N_{i-1}^2 Q_z^2 \gamma_{9+2n_{Q_z}} + |p_i| N_{i-1}^2 Q_i^2 \gamma_{9+2n_{Q_i}} + 2|p_i r_i N_{i-1}^2 Q_z| \gamma_{10+n_{Q_z}} \\ &+ |p_i N_{i-1}^2 Q_z| \gamma_{9+2n_{Q_z}} + p_i^2 N_{i-1}^2 |Q_z| \gamma_{11+2n_{Q_z}} \end{aligned} \tag{37}$$

The $\triangle_4$ error expression is:

$$\begin{aligned} & fl(\triangle_4) = 1 - (1-p_i) N_{i-1}^2 \\ & \subset 1 - N_{i-1}^2 (1 \pm u)^{2+1} - p_i N_{i-1}^2 (1 \pm u)^{3+2} \\ & \subset (1 \pm u)^2 - N_{i-1}^2 (1 \pm u)^5 - p_i N_{i-1}^2 (1 \pm u)^6 \\ & \subset 1 \pm \gamma_2 - N_{i-1}^2 (1 \pm \gamma_5) - p_i N_{i-1}^2 (1 \pm \gamma_6) \\ & = 1 - (1-p_i) N_{i-1}^2 + [\pm \gamma_2 \pm N_{i-1}^2 \gamma_5 \pm p_i N_{i-1}^2 \gamma_6] \end{aligned} \tag{38}$$

Similarly, the $\triangle_4$ error threshold is:

$$err_{\triangle_4} = \gamma_2 + N_{i-1}^2 \gamma_5 + |p_i| N_{i-1}^2 \gamma_6 \tag{39}$$

3.1.5 Floating-Point Error for Equation 7

f_{i-1} denotes the distance between the intersection points of rays between two adjacent surfaces, which is derived as follows:

$$\begin{aligned} & fl(t_{i-1}) = f_{i-1} + f_i \\ & \subset f_{i-1}(1 \pm \delta_{f_{i-1}})(1 \pm u) + f_i (1 \pm \delta_{f_i})(1 \pm u) \\ & \subset f_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u) + f_i (1 \pm u)^{n_{f_i}}(1 \pm u) \\ & \subset [f_{i-1}(1 \pm u)^{n_{f_{i-1}}}(1 \pm u) + f_i (1 \pm u)^{n_{f_i}}(1 \pm u)](1 \pm u) \\ & \subset f_{i-1}(1 \pm u)^{2+n_{f_{i-1}}} + f_i (1 \pm u)^{2+n_{f_i}} \\ & \subset f_{i-1}(1 \pm \gamma_{2+n_{f_{i-1}}}) + f_i (1 \pm \gamma_{2+n_{f_i}}) \\ & \subset f_{i-1} + f_i + [\pm f_{i-1} \gamma_{2+n_{f_{i-1}}} \pm f_i \gamma_{2+n_{f_i}}] \end{aligned} \tag{40}$$

Therefore, the upper and lower bounds of the error are given by:

$$err_{t_{i-1}} = |f_{i-1}| \gamma_{2+n_{f_{i-1}}} + |f_i| \gamma_{2+n_{f_i}} \tag{41}$$

3.1.6 Floating-Point Error for Equation 8-10

$$\begin{aligned}
&fl(x_i) = x_{i-1} + t_{i-1}L_{i-1} \\
&\subset x_{i-1}(1 \pm u) + t_{i-1}(1 \pm \delta_{t_{i-1}})(1 \pm u)L_{i-1}(1 \pm u) \\
&\subset [x_{i-1}(1 \pm u) + t_{i-1}(1 \pm u)^{n_{t_{i-1}}}(1 \pm u)L_{i-1}(1 \pm u)(1 \pm u)](1 \pm u) \\
&\subset x_{i-1}(1 \pm u)^2 + t_{i-1}L_{i-1}(1 \pm u)^{4+n_{t_{i-1}}} \\
&\subset x_{i-1}(1 \pm \gamma_2) + t_{i-1}L_{i-1}(1 \pm \gamma_{4+n_{t_{i-1}}}) \\
&= x_{i-1} + t_{i-1}L_{i-1} + [\pm x_{i-1}\gamma_2 \pm t_{i-1}L_{i-1}\gamma_{4+n_{t_{i-1}}}]
\end{aligned} \tag{42}$$

From Eq. (42), we know:

$$err_{x_i} = |x_{i-1}|\gamma_2 + |t_{i-1}L_{i-1}|\gamma_{4+n_{t_{i-1}}} \tag{43}$$

Similarly, the cumulative error of the intersection of the ray with the $(i+1)$-th surface at y_i and z_i coordinates can be deduced and rewritten as follows:

$$err_{x_i} = |x_{i-1}|\gamma_2 + |t_{i-1}L_{i-1}|\gamma_{4+n_{t_{i-1}}} \tag{44}$$

$$err_{y_i} = |y_{i-1}|\gamma_2 + |t_{i-1}M_{i-1}|\gamma_{4+n_{t_{i-1}}} \tag{45}$$

$$err_{z_i} = |z_{i-1}|\gamma_3 + |t_{i-1}N_{i-1}|\gamma_{5+n_{t_{i-1}}} + |D_{i-1}|\gamma_2 \tag{46}$$

$$\tag{47}$$

3.2 Refractive equation error model

3.2.1 Floating-Point Error for Equation 11

$$\begin{aligned}
&fl(A) = \sqrt{1 + c_i^2(1 - p_i)(y_i^2 + z_i^2)} \\
&\subset \sqrt{1 + c_i^2(1 \pm u)^3[1 - p_i(1 \pm u)][y_i^2(1 \pm \delta_{y_i})^2(1 \pm u)^3 + z_i^2(1 \pm \delta_{z_i})^2(1 \pm u)^3]} \\
&\subset \sqrt{1 + c_i^2(1 \pm u)^3[1 - p_i(1 \pm u)](1 \pm u)[y_i^2(1 \pm u)^{4+2n_{y_i}} + z_i^2(1 \pm u)^{4+2n_{z_i}}]} \\
&\subset \sqrt{1 + c_i^2[1 - p_i(1 \pm u)][y_i^2(1 \pm u)^{2n_{y_i}} + z_i^2(1 \pm u)^{2n_{z_i}}](1 \pm u)^8(1 \pm u)^2} \\
&\subset \sqrt{\{1 + c_i^2[1 - p_i(1 \pm u)][y_i^2(1 \pm u)^{2n_{y_i}} + z_i^2(1 \pm u)^{2n_{z_i}}](1 \pm u)^{10}\}(1 \pm u)} \\
&\subset \sqrt{(1 \pm u) + c_i^2(1 \pm u)^{11}[y_i^2(1 \pm u)^{2n_{y_i}} + z_i^2(1 \pm u)^{2n_{z_i}}][1 - p_i(1 \pm u)]}
\end{aligned}$$

$$\subset \sqrt{1 + c_i^2(1-p_i)(y_i^2+z_i^2) + err},$$
$$err = [\pm\gamma_1 \pm c_i^2 y_i^2 \gamma_{11+2n_{y_i}} + c_i^2 z_i^2 \gamma_{11+2n_{z_i}}$$
$$\pm p_i c_i^2 y_i^2 \gamma_{12+2n_{y_i}} \pm p_i c_i^2 z_i^2 \gamma_{12+2n_{z_i}}] \tag{48}$$

Therefore, the A error threshold can be written as:

$$err_A = \gamma_1 + c_i^2 y_i^2 \gamma_{11+2n_{y_i}} + c_i^2 z_i^2 \gamma_{11+2n_{z_i}} + |p_i| c_i^2 y_i^2 \gamma_{12+2n_{y_i}} + |p_i| c_i^2 z_i^2 \gamma_{12+2n_{z_i}} \tag{49}$$

3.2.2 Floating-Point Error for Equation 12

$$fl(\cos\theta) = \left| \frac{(1 - z_i p_i c_i) N_{i-1} - x_i c_i L_{i-1} - y_i c_i M_{i-1}}{A} \right|$$
$$\subset \left| \frac{[1 - z_i(1\pm\delta_{z_i})(1\pm u) p_i (1\pm u) c_i (1\pm u)] N_{i-1}(1\pm u)}{A(1\pm\delta_A)(1\pm u)} \right.$$
$$- \frac{x_i(1\pm\delta_{x_i})(1\pm u) c_i (1\pm u) L_{i-1}(1\pm u)}{A(1\pm\delta_A)(1\pm u)}$$
$$\left. - \frac{y_i(1\pm\delta_{y_i})(1\pm u) c_i (1\pm u) M_{i-1}(1\pm u)}{A(1\pm\delta_A)(1\pm u)} \right|$$
$$\subset \left| \frac{[1 - z_i(1\pm\delta_{z_i}) p_i c_i (1\pm u)^3] N_{i-1}(1\pm u)}{A(1\pm\delta_A)(1\pm u)} \right.$$
$$\left. - \frac{x_i(1\pm\delta_{x_i}) c_i L_{i-1}(1\pm u)^3}{A(1\pm\delta_A)(1\pm u)} - \frac{y_i(1\pm\delta_{y_i}) c_i M_{i-1}(1\pm u)^3}{A(1\pm\delta_A)(1\pm u)} \right|$$
$$\subset \left| \frac{[1 - z_i(1\pm\delta_{z_i}) p_i c_i (1\pm u)^3 (1\pm u)^2] N_{i-1}(1\pm u)(1\pm u)^2}{A(1\pm\delta_A)(1\pm u)}(1\pm u) \right.$$
$$- \frac{x_i(1\pm\delta_{x_i}) c_i L_{i-1}(1\pm u)^3(1\pm u)^2}{A(1\pm\delta_A)(1\pm u)}(1\pm u)$$
$$\left. - \frac{y_i(1\pm\delta_{y_i}) c_i M_{i-1}(1\pm u)^3(1\pm u)^2}{A(1\pm\delta_A)(1\pm u)}(1\pm u) \right|$$
$$\subset \left| \frac{[1 - z_i(1\pm u)^{n_{z_i}} p_i c_i (1\pm u)^3 (1\pm u)^2] N_{i-1}(1\pm u)^4}{A(1\pm u)^{n_A}(1\pm u)} \right.$$
$$- \frac{x_i(1\pm u)^{n_{x_i}} c_i L_{i-1}(1\pm u)^6}{A(1\pm u)^{n_A}(1\pm u)}$$
$$\left. - \frac{y_i(1\pm u)^{n_{y_i}} c_i M_{i-1}(1\pm u)^6}{A(1\pm u)^{n_A}(1\pm u)} \right|$$

$$
\begin{aligned}
&\subset \Bigg|\Bigg[\Bigg[\frac{[1-z_ip_ic_i(1\pm u)^{5+n_{z_i}}]N_{i-1}(1\pm u)^4}{A(1\pm u)^{1+n_A}}\\
&-\frac{x_ic_iL_{i-1}(1\pm u)^{6+n_{x_i}}}{A(1\pm u)^{1+n_A}}\Bigg](1\pm u)-\frac{y_ic_iM_{i-1}(1\pm u)^{6+n_{y_i}}}{A(1\pm u)^{1+n_A}}\Bigg](1\pm u)\Bigg|\\
&\subset \Bigg|\frac{[1-z_ip_ic_i(1\pm u)^{5+n_{z_i}}]N_{i-1}(1\pm u)^6}{A(1\pm u)^{1+n_A}}\\
&-\frac{x_ic_iL_{i-1}(1\pm u)^{8+n_{x_i}}}{A(1\pm u)^{1+n_A}}-\frac{y_ic_iM_{i-1}(1\pm u)^{7+n_{y_i}}}{A(1\pm u)^{1+n_A}}\Bigg|\\
&\subset \Bigg|\frac{N_{i-1}(1\pm u)^6-z_ip_ic_iN_{i-1}(1\pm u)^{11+n_{z_i}}}{A(1\pm u)^{1+n_A}}\\
&-\frac{x_ic_iL_{i-1}(1\pm u)^{8+n_{x_i}}}{A(1\pm u)^{1+n_A}}-\frac{y_ic_iM_{i-1}(1\pm u)^{7+n_{y_i}}}{A(1\pm u)^{1+n_A}}\Bigg|\\
&\subset \frac{N_{i-1}}{A}(1\pm\gamma_{7+n_A})-\frac{z_ip_ic_iN_{i-1}}{A}(1\pm\gamma_{12+n_A+n_{z_i}})\\
&-\frac{x_ic_iL_{i-1}}{A}(1\pm\gamma_{9+n_A+n_{x_i}})-\frac{y_ic_iM_{i-1}}{A}(1\pm\gamma_{8+n_A+n_{y_i}})\\
&\subset \Bigg|\frac{(1-z_ip_ic_i)N_{i-1}-x_ic_iL_{i-1}-y_ic_iM_{i-1}}{A}\\
&+\Bigg[\pm\frac{N_{i-1}}{A}\gamma_{7+n_A}\pm\frac{z_ip_ic_iN_{i-1}}{A}\gamma_{12+n_A+n_{z_i}})\\
&\pm\frac{x_ic_iL_{i-1}}{A}\gamma_{9+n_A+n_{x_i}}\pm\frac{y_ic_iM_{i-1}}{A}\gamma_{8+n_A+n_{y_i}}\Bigg]\Bigg|
\end{aligned}\tag{50}
$$

From Eq. (50), the $\cos\theta$ upper and lower bound error threshold expressions are:

$$
\begin{aligned}
err_{\cos\theta}=&\left|\frac{N_{i-1}}{A}\right|\gamma_{7+n_A}+\left|\frac{z_ip_ic_iN_{i-1}}{A}\right|\gamma_{12+n_A+n_{z_i}}\\
&+\left|\frac{x_ic_iL_{i-1}}{A}\right|\gamma_{9+n_A+n_{x_i}}+\left|\frac{y_ic_iM_{i-1}}{A}\right|\gamma_{8+n_A+n_{y_i}}
\end{aligned}\tag{51}
$$

3.2.3 Floating-Point Error for Equation 13

$$
fl(\cos\theta')=\sqrt{1-\frac{n_{i-1}^2}{n_i^2}(1-\cos^2\theta)}
$$

$$
\begin{aligned}
&\subset \sqrt{1-\frac{n_{i-1}^2(1\pm u)^3}{n_i^2(1\pm u)^3}[1-\cos^2\theta(1\pm\delta_{\cos\theta})^2(1\pm u)^3]}\\
&\subset \sqrt{\left[1-\left[\frac{n_{i-1}^2(1\pm u)^3}{n_i^2(1\pm u)^3}\right](1\pm u)\left[1-\cos^2\theta(1\pm\delta_{\cos\theta})^2(1\pm u)^3\right](1\pm u)^2\right](1\pm u)}\\
&\subset (1\pm u)\sqrt{(1\pm u)-\frac{n_{i-1}^2(1\pm u)^7}{n_i^2(1\pm u)^3}+\frac{n_{i-1}^2\cos^2\theta(1\pm\delta_{\cos\theta})^2(1\pm u)^{10}}{n_i^2(1\pm u)^3}}\\
&\subset \sqrt{(1\pm u)^3-\frac{n_{i-1}^2(1\pm u)^9}{n_i^2(1\pm u)^3}+\frac{n_{i-1}^2\cos^2\theta[(1\pm u)^{n_{\cos\theta}}]^2(1\pm u)^{12}}{n_i^2(1\pm u)^3}}\\
&\subset \sqrt{1\pm\gamma_3-\frac{n_{i-1}^2}{n_i^2}(1\pm\gamma_{12})+\frac{n_{i-1}^2\cos^2\theta}{n_i^2}(1\pm\gamma_{15+2n_{\cos\theta}})}\\
&\subset \sqrt{1-\frac{n_{i-1}^2}{n_i^2}(1-\cos^2\theta)+\left[\pm\gamma_3\pm\frac{n_{i-1}^2}{n_i^2}\gamma_{12}\pm\frac{n_{i-1}^2}{n_i^2}\cos^2\theta\gamma_{15+2n_{\cos\theta}})\right]} \qquad (52)
\end{aligned}
$$

From the above equation, the threshold value for the upper and lower bounds of $\cos\theta'$ error can be set as:

$$
err_{\cos\theta'} = \gamma_3 + \frac{n_{i-1}^2}{n_i^2}\gamma_{12} + \frac{n_{i-1}^2}{n_i^2}\cos^2\theta\gamma_{15+2n_{\cos\theta}} \qquad (53)
$$

3.2.4 Floating-Point Error for Equation 14

$$
\begin{aligned}
&fl(g_i) = n_i\cos\theta' - n_{i-1}\cos\theta\\
&\subset n_i(1\pm u)\cos\theta'(1\pm\delta_{\cos\theta'})(1\pm u) - n_{i-1}(1\pm u)\cos\theta(1\pm\delta_{\cos\theta})(1\pm u)\\
&\subset n_i(1\pm u)\cos\theta'(1\pm u)^{n_{\cos\theta'}}(1\pm u) - n_{i-1}(1\pm u)\cos\theta(1\pm u)^{n_{\cos\theta}}(1\pm u)\\
&\subset \left[n_i\cos\theta'(1\pm u)^{n_{\cos\theta'}}(1\pm u)^3 - n_{i-1}\cos\theta(1\pm u)^{n_{\cos\theta}}(1\pm u)^3\right](1\pm u)\\
&\subset n_i\cos\theta'(1\pm u)^{4+n_{\cos\theta'}} - n_{i-1}\cos\theta(1\pm u)^{4+n_{\cos\theta}}\\
&\subset n_i\cos\theta'(1\pm\gamma_{4+n_{\cos\theta'}}) - n_{i-1}\cos\theta(1\pm\gamma_{4+n_{\cos\theta}})\\
&\subset n_i\cos\theta' - n_{i-1}\cos\theta + \left[\pm n_i\cos\theta'\gamma_{4+n_{\cos\theta'}} \pm n_{i-1}\cos\theta\gamma_{4+n_{\cos\theta}}\right] \qquad (54)
\end{aligned}
$$

Therefore, the upper and lower error thresholds of g_i can be set in the following form:

$$
err_{g_i} = |n_i\cos\theta'|\gamma_{4+n_{\cos\theta'}} + |n_{i-1}\cos\theta|\gamma_{4+n_{\cos\theta}} \qquad (55)
$$

3.2.5 Floating-Point Error for Equation 15-17

$$
\begin{aligned}
L_i &= \frac{n_{i-1}}{n_i} L_{i-1} - \frac{g_i c_i x_i}{n_i A} \\
&\subset \frac{n_{i-1}(1 \pm u)}{n_i(1 \pm u)} L_{i-1}(1 \pm u) \\
&\quad - \frac{g_i(1 \pm \delta_{g_i})(1 \pm u) c_i (1 \pm u) x_i (1 \pm \delta_{x_i})(1 \pm u)}{n_i(1 \pm u) A (1 \pm \delta_A)(1 \pm u)} \\
&\subset \frac{n_{i-1}(1 \pm u)}{n_i(1 \pm u)} L_{i-1}(1 \pm u)^3 \\
&- \frac{g_i(1 \pm \delta_{g_i})(1 \pm u) c_i (1 \pm u) x_i (1 \pm \delta_{x_i})(1 \pm u)}{n_i A (1 \pm \delta_A)(1 \pm u)^3} (1 \pm u)^3 \\
&\subset \left[\frac{n_{i-1}}{n_i(1 \pm u)} L_{i-1}(1 \pm u)^4 \right. \\
&\quad \left. - \frac{g_i(1 \pm \delta_{g_i}) c_i x_i (1 \pm \delta_{x_i})}{n_i A (1 \pm \delta_A)(1 \pm u)^3} (1 \pm u)^6 \right] (1 \pm u) \\
&\subset \frac{n_{i-1}}{n_i(1 \pm u)} L_{i-1}(1 \pm u)^5 - \frac{g_i(1 \pm \delta_{g_i}) c_i x_i (1 \pm \delta_{x_i})}{n_i A (1 \pm \delta_A)(1 \pm u)^3} (1 \pm u)^7 \\
&\subset \frac{n_{i-1}}{n_i(1 \pm u)} L_{i-1}(1 \pm u)^5 - \frac{g_i(1 \pm u)^{n_{g_i}} c_i x_i (1 \pm u)^{n_{x_i}}}{n_i A (1 \pm u)^{n_A}(1 \pm u)^3} (1 \pm u)^7 \\
&\subset \frac{n_{i-1}}{n_i} L_{i-1}(1 \pm \gamma_6) - \frac{g_i c_i x_i}{n_i A} (1 \pm \gamma_{10+n_{g_i}+n_A+n_{x_i}}) \\
&= \frac{n_{i-1}}{n_i} L_{i-1} - \frac{g_i c_i x_i}{n_i A} + \left[\pm \frac{n_{i-1}}{n_i} L_{i-1} \gamma_6 \pm \frac{g_i c_i x_i}{n_i A} \gamma_{10+n_{g_i}+n_A+n_{x_i}} \right]
\end{aligned} \tag{56}
$$

From the above equation, the upper and lower bounds for the computational error of the direction cosine L_i are:

$$
err_{L_i} = \left| \frac{n_{i-1}}{n_i} L_{i-1} \right| \gamma_6 + \left| \frac{g_i c_i x_i}{n_i A} \right| \gamma_{10+n_{g_i}+n_A+n_{x_i}} \tag{57}
$$

Therefore the error range for the calculation of the direction cosine L_i is:

$$
L_i \subset \left[\frac{n_{i-1}}{n_i} L_{i-1} - \frac{g_i c_i x_i}{n_i A} - err_{L_i}, \frac{n_{i-1}}{n_i} L_{i-1} - \frac{g_i c_i x_i}{n_i A} + err_{L_i} \right] \tag{58}
$$

Referring to L_i, the M_i error expression can be written directly as:

$$
M_i = \frac{n_{i-1}}{n_i} M_{i-1} - \frac{g_i c_i y_i}{n_i A}
$$

$$\subset \frac{n_{i-1}}{n_i}M_{i-1} - \frac{g_i c_i y_i}{n_i A} + \left[\pm\frac{n_{i-1}}{n_i}M_{i-1}\gamma_6 \pm \frac{g_i c_i y_i}{n_i A}\gamma_{10+n_{g_i}+n_A+n_{y_i}}\right] \tag{59}$$

The upper and lower bounds for the computational error of the direction cosine M_i are:

$$err_{M_i} = \left|\frac{n_{i-1}}{n_i}M_{i-1}\right|\gamma_6 + \left|\frac{g_i c_i y_i}{n_i A}\right|\gamma_{10+n_{g_i}+n_A+n_{y_i}} \tag{60}$$

The computational error range of the direction cosine M_i is:

$$M_i \subset \left[\frac{n_{i-1}}{n_i}M_{i-1} - \frac{g_i c_i y_i}{n_i A} - err_{M_i}, \frac{n_{i-1}}{n_i}M_{i-1} - \frac{g_i c_i y_i}{n_i A} + err_{M_i}\right] \tag{61}$$

Derive the error equation for N_i as follows:

$$\begin{aligned}
N_i &= \frac{n_{i-1}}{n_i}N_{i-1} - \frac{g_i(p_i c_i z_i - 1)}{n_i A} \\
&\subset \frac{n_{i-1}(1\pm u)}{n_i(1\pm u)}N_{i-1}(1\pm u) \\
&- \frac{g_i(1\pm\delta_{g_i})(1\pm u)[p_i(1\pm u)c_i(1\pm u)z_i(1\pm\delta_{z_i})(1\pm u) - 1]}{n_i(1\pm u)A(1\pm\delta_A)(1\pm u)} \\
&\subset \frac{n_{i-1}(1\pm u)}{n_i(1\pm u)}N_{i-1}(1\pm u) \\
&- \frac{g_i(1\pm u)^{n_{g_i}}(1\pm u)[p_i(1\pm u)c_i(1\pm u)z_i(1\pm u)^{n_{z_i}}(1\pm u) - 1]}{n_i(1\pm u)A(1\pm u)^{n_A}(1\pm u)} \\
&\subset \frac{n_{i-1}}{n_i(1\pm u)}N_{i-1}(1\pm u)^2 - \frac{g_i(1\pm u)^{1+n_{g_i}}[p_i c_i z_i(1\pm u)^{3+n_{z_i}} - 1]}{n_i A(1\pm u)^{2+n_A}} \\
&\subset \frac{n_{i-1}}{n_i(1\pm u)}N_{i-1}(1\pm u)^4 - \\
&\left[\frac{\left[g_i(1\pm u)^{1+n_{g_i}}\left[p_i c_i z_i(1\pm u)^{5+n_{z_i}} - 1\right](1\pm u)\right](1\pm u)}{n_i A(1\pm u)^{3+n_A}}\right](1\pm u) \\
&\subset \frac{n_{i-1}}{n_i(1\pm u)}N_{i-1}(1\pm u)^4 - \frac{g_i[p_i c_i z_i(1\pm u)^{5+n_{z_i}} - 1]}{n_i A(1\pm u)^{3+n_A}}(1\pm u)^{4+n_{g_i}} \\
&\subset \frac{n_{i-1}}{n_i(1\pm u)}N_{i-1}(1\pm u)^5 - \frac{g_i\left[p_i c_i z_i(1\pm u)^{5+n_{z_i}} - 1\right]}{n_i A(1\pm u)^{3+n_A}}(1\pm u)^{5+n_{g_i}} \\
&\subset \frac{n_{i-1}}{n_i(1\pm u)}N_{i-1}(1\pm u)^5 - \frac{g_i p_i c_i z_i(1\pm u)^{10+n_{z_i}+n_{g_i}}}{n_i A(1\pm u)^{3+n_A}}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{g_i(1 \pm u)^{5+n_{g_i}}}{n_i A(1 \pm u)^{3+n_A}} \\
&\subset \frac{n_{i-1}}{n_i} N_{i-1}(1 \pm \gamma_6) - \frac{g_i p_i c_i z_i}{n_i A}(1 \pm \gamma_{13+n_{z_i}+n_{g_i}+n_A}) \\
&+ \frac{g_i}{n_i A}(1 \pm \gamma_{8+n_{g_i}+n_A}) \\
&= \frac{n_{i-1}}{n_i} N_{i-1} - \frac{g_i(p_i c_i z_i - 1)}{n_i A} \\
&+ \left[\pm \frac{n_{i-1}}{n_i} N_{i-1}\gamma_6 \pm \frac{g_i p_i c_i z_i}{n_i A}\gamma_{13+n_{z_i}+n_{g_i}+n_A}) \pm \frac{g_i}{n_i A}\gamma_{8+n_{g_i}+n_A} \right]
\end{aligned} \tag{62}
$$

The upper and lower bounds for the computational error of the direction cosine N_i are:

$$
\begin{aligned}
err_{N_i} = &\left| \frac{n_{i-1}}{n_i} N_{i-1} \right| \gamma_6 \\
&+ \left| \frac{g_i p_i c_i z_i}{n_i A} \right| \gamma_{13+n_{z_i}+n_{g_i}+n_A} + \left| \frac{g_i}{n_i A} \right| \gamma_{8+n_{g_i}+n_A}
\end{aligned} \tag{63}
$$

References

[1] R. Kingslake, R. B. Johnson, Lens design fundamentals 2nd edition, Academic Press, 2009.

[2] D. P. Feder, Optical calculations with automatic computing machinery, J. Opt. Soc. Am. 41 (1951) 630-635.