LA-DC-6813

Automatic Lens Design by Nonlinear Least Squares, Practice and Malpractice*

Thomas C. Doyle

University of California, Los Alamos Scientific Laboratory,

Los Alamos, New Mexico

yK. f. c. phi

The merit of any trial lens system with continuously variable parameters $\underset{\sim}{u}\equiv(u^1, \cdot \cdot \cdot, u^n)$ may be evaluated by tracing sample rays to form the $N\geqq n$ components E_A of an error vector $\underset{\sim}{E}$ of squared magnitude $\phi\equiv\underset{\sim}{E}\cdot\underset{\sim}{E}$. Regarding the lens $\underset{\sim}{u}$ as a point of arithmetic space A_n, we consider the most general mapping $\mathfrak{m}$ of n-dimensional A_n into N-dimensional Euclidean $\mathcal{E}_N$ and define the direction of steepest descent in A_n of ϕ relative to this general mapping $\mathfrak{m}$. We then restrict $\mathfrak{m}$ to be the orthonormal lens error mapping of A_n into E_N. The resulting least squares equations of steepest descent lead us to a parameter increment vector $\Delta\underset{\sim}{u}\in A_n$ such that $\phi(\underset{\sim}{u}+\Delta\underline{u})<\phi(\underset{\sim}{u})$. Iteration generates a convergent monotone

*Work done under the auspices of the U. S. Atomic Energy Commission.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

decreasing sequence $\{\phi\}\to\phi_L$ and its corresponding trial lens sequence $\{\underset{\sim}{u}\}\to\underset{\sim}{u}_L$, where $\underset{\sim}{u}_L$ is the desired local optimum lens attainable by the process from the first trial lens.

INTRODUCTION

This paper is missionary. Its purpose is to win converts to the cause of correctly applied nonlinear least squares optimization techniques. A sizable portion of the running time of many digital computers is occupied with the following problem: a certain system is completely determined by a choice of n continuously varying parameters, $\underset{\sim}{u}\equiv(u^1, \cdot \cdot \cdot, u^n)$, and hence the system may be idealized as a point $\underset{\sim}{u}$ in arithmetic n-space A_n. The ∞^n family of such systems fills a region of A_n. It is assumed that any sample system of the family may be appraised by forming an error vector $E(\underset{\sim}{u})\equiv(E_1, \cdot \cdot \cdot, E_N)$, $N\geq n$, whose components $E_A(\underline{u})$ measure the departure from perfection of the sample. The positive scalar $\phi\equiv\underset{\sim}{E}\cdot\underset{\sim}{E}\equiv E_Q E_Q$ (a repeated index will imply summation over its range in the absence of any statement to the contrary) is taken as measuring the merit of the sample and ϕ is called the merit function, the smaller ϕ, the better the sample. A sample system $\underset{\sim}{u}$ will be called optimal with respect to its neighbors in A_n when $\phi(\underset{\sim}{u})$ is a local minimum.

The search for an optimal system is begun by the player's blindfolded draw of an initial system $\underset{\sim}{u}_0$ from the urn A_n. He then employs the error vector $\underset{\sim}{E}$ and its partial derivatives $\underset{\sim}{E}_{,i} \equiv (\partial E_1/\partial u^i, \cdot \cdot \cdot, \partial E_N/\partial u^i)$ to determine an improved system $\underset{\sim}{u}_1$. Continuing in this way, he forms a sequence of trial systems $\{\underset{\sim}{u}\} \equiv \underset{\sim}{u}_0, \underset{\sim}{u}_1, \cdot \cdot \cdot$ and the corresponding sequence of merit functions $\{\phi\}$. When the derived matrix $\|E_{A,i}(\underset{\sim}{u})\|$ is independent of $\underset{\sim}{u}$ the problem is called *linear*, and otherwise *nonlinear*. An intuitively obvious theorem of analysis ensures that, when the positive sequence $\{\phi\}$ is *monotone decreasing*, it converges to a limit ϕ_L. Then if $\underset{\sim}{u}_L$ be the limit of the corresponding system sequence $\{\underset{\sim}{u}\}$, $\underset{\sim}{u}_L$ is the desired optimal system.

The chief result of this paper is a constructive proof that a correct application of the method of nonlinear least squares will restrict the computer to the formation of a *monotone decreasing* sequence of merit functions. This is in sharp contrast to a non-monotone sequence wherein the time costly output of a selection cycle must be rejected when the new merit function reveals its system to be poorer than the predecessor.

Definition. Any least squares program which extravagantly commits the computer to the formation of a sequence $\{\phi\}$ which is *not* monotone decreasing will be called a *malpractice* program.

Observation. Most nonlinear least squares optimization programs which have come to our attention have been malpractice programs. In particular, the lens design code employed at Los Alamos and written by J. C. Holladay[1] is such a program.

As a second illustration to support our observation, we quote from D. P. Feder[2] who comments on the nonlinear least squares optical design method: "It is a matter of practical experience that, when the process converges, the speed of convergence is much greater than in the gradient method. On the other hand, if the equations are sufficiently nonlinear, the process will diverge, and if this occurs, the characteristic behavior of the solution vector is to oscillate wildly about the minimum without ever getting close to it. This, in fact, seems to be the normal behavior of the method when used in optical design." Feder's description shows that Holladay does not walk alone along the pathway of malpractice. It is a trail well trodden by the optical fraternity!

1. OPTIMIZATION EXAMPLES

Let it be desired to minimize a given differentiable function $\psi(u^1, \cdot \cdot \cdot, u^n)$ over A_n. One may form the error vector $\underset{\sim}{E} \equiv (\partial\psi/\partial u^1, \cdot \cdot \cdot, \partial\psi/\partial u^n)$ and minimize $\phi \equiv \underset{\sim}{E} \cdot \underset{\sim}{E}$. As a second example, let us seek a solution of the nonlinear algebraic system $E_i(u^1, \cdot \cdot \cdot, u^n)=0$, $i=1, \cdot \cdot \cdot, n$. If $\underset{\sim}{u}$ be a trial solution, we form the error vector $\underset{\sim}{E} \equiv (E_1(\underset{\sim}{u}), \cdot \cdot \cdot, E_n(\underset{\sim}{u}))$ and minimize $\phi \equiv \underset{\sim}{E} \cdot \underset{\sim}{E}$ over A_n. As a third example, let us seek to select from the ∞^n family of plane curves $y=f(x; u^1, \cdot \cdot \cdot, u^n)$ that particular curve of the family which best fits a given set of data points (x_A, y_A), $A=1, \cdot \cdot \cdot, N$, $N \geqq n$, in the least squares sense. We form the error vector components $E_A \equiv f(x_A; \underset{\sim}{u})-y_A$, $A=1, \cdot \cdot \cdot, N$, and minimize $\phi \equiv \underset{\sim}{E} \cdot \underset{\sim}{E}$ over A_n.

A fourth example, which is a model of least squares precision, is the method of R. E. von Holdt[3] for computing the eigenvalues and eigenvectors of a real symmetric matrix $\underset{\sim}{A}$. Here one seeks a vector $\underset{\sim}{u}$ and a related scalar λ such that $\underset{\sim}{A}\underset{\sim}{u}-\lambda\underset{\sim}{u}=0$. The trial system of our introduction is now the trial eigenvector $\underset{\sim}{u}$ and the corresponding trial λ is taken as $\lambda=\underset{\sim}{u}^T\underset{\sim}{A}\underset{\sim}{u}/\underset{\sim}{u}^T\underset{\sim}{u}$. The error vector $\underset{\sim}{E}$ is defined by $\underset{\sim}{E} \equiv (\underset{\sim}{A}\underset{\sim}{u}-\lambda\underset{\sim}{u})/|\underset{\sim}{u}|$ and $\phi \equiv \underset{\sim}{E} \cdot \underset{\sim}{E}$ is minimized over A_n. The writer has coded von Holdt's method for the IBM Stretch and has resolved each eigenvalue and eigenvector pair $(\lambda, \underset{\sim}{u})$ of an exacting 8x8 test matrix[4] with 14 digit accuracy by the iteration sequence {6, 5, 6, 3, 6, 5, 3, 1} per pair $(\lambda, \underset{\sim}{u})$. The rapid convergence of von Holdt's method as applied to a highly nonlinear problem first convinced the writer that lack of monotonicity in the merit function sequence $\{\phi\}$ represents a gross defect in any least squares optimization program which should not be tolerated.

2. LENS DESIGN

A lens designer, faced with a certain imaging problem, draws upon his background of experience with such problems to commit his present design to

1) a certain number n_{surf} of refracting or reflecting surfaces Σ_1, Σ_2, · · · whose sequential order agrees with that in which a light ray from an object point meets the surfaces,

2) achromatization with respect to a specified set of colors β, $\beta=1, \cdot \cdot \cdot, n_{color}$,

3) a definite selection of refracting materials 0, 1, · · · encountered by the ray in 1) in the order 0, Σ_1, 1, Σ_2, 2, · · · whose refractive indexes for the β in 2) are known to be $n_{0\beta}$, $n_{1\beta}$, · · ·,

4) a camera box of specified dimensions,

5) refracting or reflecting surfaces which are to be selected from the 2-parameter family of quadric surfaces (B, C) of revolution defined by Eq. (12.4), of vertex curvature B and of form constant C, where $C<-1$ yields an oblate spheroid, $C=-1$ a sphere of radius $R=1/B$, $-1<C<0$ a prolate spheroid, and $C>0$ an hyperboloid of two sheets.

He then proceeds to optimize his design over the totality of lens systems compatible with his specifications. Thus, if he should choose $n_{surf}=6$ in 1) and define the sequence in 3) to be {a, g, g, a, g, g, a}, a≡air, g≡glass, the resultant family of sample lens systems will be of the Lister type illustrated in Fig. 2.1. The parameter set

<u>insert Fig. 2.1 approximately here</u>

$\underset{\sim}{u}$ is given by

$$\begin{aligned} &u^{-2} \equiv \rho_0 \equiv \text{ radius of entrance pupil,} \\ &u^{-1} \equiv d_{ep} \equiv \text{ distance from } \Sigma_0 \text{ to plane of entrance pupil,} \\ &u^{i} \equiv d_i \equiv \text{ oriented distance } \underline{\text{from}}\ \Sigma_i\ \underline{\text{to}}\ \Sigma_{i+1},\ i=0, \cdot\cdot\cdot, 6, \\ &u^{6+i} \equiv B_i \equiv \text{ vertex curvature of } \Sigma_i,\ i=1, \cdot\cdot\cdot, 7, \\ &u^{13+i} \equiv C_i \equiv \text{ form constant of } \Sigma_i,\ i=1, \cdot\cdot\cdot, 7, \end{aligned} \qquad (2.1)$$

so that Fig. 2.1 depicts a representative member taken from the family of ∞^{23} such lenses obtained by varying the ordered set $\underset{\sim}{u} \equiv (u^{-2}, \cdot\cdot\cdot, u^{20})$ independently. The best performing lens in this set will be found by optimizing the design over A_{23}. Economy in production, however, may possibly be achieved by imposing conditions of symmetry or skew-symmetry and other side conditions on the eligible family. We shall define such conditions for the computer by means of a parameter flag vector $\underset{\sim}{f} \equiv (f^{-2}, \cdot\cdot\cdot, f^{20})$, where

1) $f^i=i$ means that u^i is <u>independent</u>;

2) $f^i=999$ means that u^i is to be held <u>fixed</u> at its input data value;

3) $f^j=+(-)i\neq j$, $i\neq 0$, means that u^j is dependent upon the independent u^i according to the symmetry (skew-symmetry) condition $u^j(\text{dep})=+(-)u^i(\text{indep.})$ (note that if the pair (u^0, u^1) forms a dependent set, then u^1 is flagged as independent and u^0 as dependent);

4) $f^i\equiv -999$, $0\leq i\leq n_{surf}$, means that the spacing $u^i\equiv d_i$ will be assigned a sequence of input data values as in the case of the sliding lens in a zoom assembly as discussed in section 11;

5) $f^\alpha=9999$, $0\leq\alpha\leq n_{surf}$ means that the spacing $u^\alpha\equiv d_\alpha$ is dependent upon the independent d's according to the dependency

$$d_\alpha\equiv D_\alpha+D'_{\alpha r'}d_{r'}, \tag{2.2}$$

where $d_{r'}$ is summed over the independent d's and D_α and $D'_{\alpha r'}$ are data input constants.

As an example of the use of the parameter flag vector $\underset{\sim}{f}$ in constraining the design, let us introduce the following constraints on the Lister lens of Fig. 2.1:

a) Let the distance from Σ_0 to Σ_7 be a fixed L.

b) Let the lens (Σ_1, Σ_2) be identical with (Σ_4, Σ_5).

c) Let the surfaces Σ_3 and Σ_6 be skew-symmetrical.

d) Let $d_0=d_6$ and $d_2=d_5$.

e) Let the photographic plate be plane ($B_7=0$, $C_7=-1$).

f) Let all refracting surfaces be spherical (all C's are -1). The following flag vector $\underset{\sim}{f}$ imposes these constraints:

$f^{-2}=-2$, $f^{-1}=-1$; $f^0=6$, $f^1=1$, $f^2=2$; $f^3=9999$; $f^4=1$, $f^5=2$, $f^6=6$;
$f^7=7$, $f^8=8$, $f^9=9$; $f^{10}=7$, $f^{11}=8$, $f^{12}=-9$; $f^{13}=999$;
$f^i=999$, $i=14, \cdot \cdot \cdot 20$, (all C's held fixed at common input data value -1).

Here the dependency (2.2) imposed on d_3 by a), b), and d) is $d_3=L-2d_1-2d_2-2d_6$. Thus $\alpha=3$ defines the α range in (2.2), and $d_{r'}$ ranges over the independent d's, namely d_1, d_2, d_6. The independent parameters of our constrained system have been chosen to be (ρ_0, d_{ep}; d_1, d_2, d_6; B_7, B_8, B_9) so that our constraints have reduced the dimension of our eligible family from 23 to 8. In general, the unconstrained lens optimized over A_{23} will be superior in performance to the constrained lens optimized over A_8. Only under the highly improbable circumstance that the optimum over A_{23} lies itself in A_8 will the two optima be the same.

In the sequel it should be understood that the parameter set $\underset{\sim}{u}\equiv(u^1, \cdot \cdot \cdot, u^n)$ refers to the n independent parameters of the trial lens as defined by the parameter flag vector $\underset{\sim}{f}$ and a possible renaming. Thus n=23 for the unconstrained Lister lens of Fig. 2.1 while n=8 for the constrained lens of our example.

3. LIGHT CONE SAMPLING

In deference to a well established convention, Fig. 2.1 shows an entrance pupil of radius ρ_0 with its plane normal to the optical axis and at a distance d_{ep} from the object plane. If an axial object point Q be regarded as the vertex of a right circular cone having the entrance pupil as base, then any ray from Q lying within this cone will traverse the lens system and reach the plate. Now let the entrance pupil be subdivided into small squares by a square mesh and consider the totality of rays emanating from the axial point Q and passing through the centers of the mesh squares. This family of rays constitutes a planar approximation to a finite uniform sampling of the continuous family of ∞^2 rays from the axial object point Q to all the ∞^2 points of the entrance pupil. A truly uniform sampling of the rays from the axial point Q would result from properly choosing the mesh points to lie equally spaced on a sphere about Q as center rather than equally spaced on a tangent plane to such a sphere as we have done. A desire for simplicity in computation, however, prompts us to settle for the planar approximation to uniform ray sampling just described.

Now, with each sample ray still joining the light cone vertex Q with a mesh square center, continuously move Q away from its former axial position along the meridional line normal to the axis and observe the gradually increasing nonuniformity of the finite ray sampling. As the principal ray from Q becomes steeply inclined to the optical axis the finite population sample becomes grossly distorted from a uniform sample.

The Holladay lens design code, which employs a conventional entrance pupil normal to the axis as in Fig. 2.1, has been found by B. Brixner to be inadequate for the design of a wide angle lens. As a remedy Brixner proposes that, for a given pair (d_{ep}, ρ_0) of Fig. 3.1, a 1-parameter family of inclined entrance pupils as in Fig. 3.1 replace the conventional unique normal entrance pupil of Fig. 2.1. Thus Brixner proposes that a lens be designed by insisting that, for Q any object point, all light in the right circular cone of Fig. 3.1 should traverse the lens and reach the plate.

insert Fig. 3.1 approximately here

4. AIMING OF SAMPLE RAYS

Let the orthonormal triad ($\underset{\sim}{e}_x$, $\underset{\sim}{e}_y$, $\underset{\sim}{e}_z$) of Fig. 3.1 take the directions of a rectangular Cartesian set of axes with origin at the center O' of the inclined entrance pupil of Fig. 3.1. Let the plane of the inclined entrance pupil be spanned by the orthonormal pair ($\underset{\sim}{e}_x$, $\underset{\sim}{e}_y'$), where the pair ($\underset{\sim}{e}_y'$, $\underset{\sim}{e}_z'$) of Fig. 3.1 results from the pair ($\underset{\sim}{e}_y$, $\underset{\sim}{e}_z$) by a clockwise rotation in the y-z plane through the angle θ. Sighting from Q we see the inclined entrance pupil as in Fig. 4.1 The

insert Fig. 4.1 approximately here

dots show those mesh square centers which are interior to the inclined entrance pupil of radius ρ_0, the mesh squares having a side $a_0 \equiv \rho_0 / q_0$.

For a given mesh refinement q_0 the sample rays from the object point Q will be traced in the order shown in Fig. 4.1. The circular symmetry of the lens ensures that the fate of rays through the left-hand semicircle is known when that through the right-hand semicircle has been computed. The j-th component q_j of the vector $\underset{\sim}{q} \equiv (q_1, q_2, \cdot \cdot \cdot)$ gives the number of mesh points in the j-th column of a quadrant. We regard a single mesh point ray as approximating the refractive fate of the ∞^2 rays from Q through the subarea represented by the mesh point. All mesh points which are the centers of squares wholly interior to the quadrant are tentatively assigned unit weight. A planimeter, a compass and some common sense serve to establish the weights of boundary mesh points and occasionally the tentative unit weight of a near boundary mesh point is altered. The first three weight vectors were found to be

$$\begin{aligned} &q_0=1,\ \underset{\sim}{w}\equiv(0.785,\ 0.785),\\ &q_0=2,\ \underset{\sim}{w}\equiv(1.071,\ 1,\ 1,\ 1.071;\ 1.071,\ 1.071),\\ &q_0=3,\ \underset{\sim}{w}=(0.948,\ 1,\ 1,\ 1,\ 1,\ 0.948;\ 0.586,\ 1,\ 1,\ 1,\ 1,\ 0.586;\\ &\qquad\qquad (0.586,\ 0.948,\ 0.948,\ 0.586). \end{aligned} \tag{4.1}$$

Flawless planimeter measurements would result in a $\underset{\sim}{w}$ whose components w_i satisfy $\Sigma w_i=(\pi/2)q_0{}^2$. The weight vectors $\underset{\sim}{w}$ through $q_0=10$ were carefully determined by Mrs. R. E. Luders of Los Alamos.

Let us now determine the unit vector $\underset{\sim}{u}$ directed from the object point Q of Fig. 3.1 toward a typical inclined entrance pupil mesh point $(x_0,\ y_0')$ whose coordinates are referred to the inclined triad $(\underset{\sim}{e}_x,\ \underset{\sim}{e}_y')$ of Fig. 3.1. With j designating the column of mesh points in Fig. 4.1 for a given q_0, then

$$\begin{aligned} &x_0=(j-\tfrac{1}{2})a_0,\ a_0\equiv\rho_0/q_0,\ j=1,\ \cdot\ \cdot\ \cdot,\ q_0,\\ &y_0'=(i-\tfrac{1}{2})a_0,\ i=-(q_j-1),\ -(q_j-2),\ \cdot\ \cdot\ \cdot,\ q_j. \end{aligned} \tag{4.2}$$

Inspection of Fig. 3.1 shows that the mesh point with coordinates $(x_0,\ y_0',\ 0)$ relative to the inclined triad $(\underset{\sim}{e}_x,\ \underset{\sim}{e}_y',\ \underset{\sim}{e}_z')$ with origin at O' has the coordinates

$(x_0, y_0'\cos\theta, d_{ep}+y_0'\sin\theta)$ relative to the triad $(\underset{\sim}{e}_x, \underset{\sim}{e}_y, \underset{\sim}{e}_z)$ with origin at O. Thus the desired unit aiming vector $\underset{\sim}{u}$ referred to the triad $(\underset{\sim}{e}_x, \underset{\sim}{e}_y, \underset{\sim}{e}_z)$, has the components

$$\begin{aligned}&u_x \equiv x_0/M_{ag},\ u_y \equiv (y_0'\cos\theta-h)/M_{ag},\ u_z \equiv (d_{ep}+y_0'\sin\theta/M_{ag},\\ &M_{ag}^2 \equiv x_0^2+(y_0'\cos\theta-h)^2+(d_{ep}+y_0'\sin\theta)^2,\\ &(\cos\theta,\ \sin\theta)=(d_{ep}/(d_{ep}^2+h^2)^{\frac{1}{2}},\ h/(d_{ep}^2+h^2)^{\frac{1}{2}}).\end{aligned} \tag{4.3}$$

The same program may compare the design resulting from the conventional normal entrance pupil with that resulting from Brixner's inclined pupil by granting that $(\cos\theta, \sin\theta)$ be defined by either (1, 0) or Eq. (4.3) according to option.

5. THE IDEAL LENS

Returning to our introduction, we begin any least squares machine pass with a certain trial lens $\underset{\sim}{u}$ in storage, where $\underset{\sim}{u}$ is better in a least squares sense than any of its predecessors, and proceed to appraise the current lens by forming its error vector $\underset{\sim}{E}$ whose components measure the departure of the trial $\underset{\sim}{u}$ from an ideal lens. The current machine storage values of the parameter pair (ρ_0, d_{ep}) of Fig. 3.1, when taken together with the constant h defining a particular object point Q, determine a current machine storage light cone $C(h, \rho_0, d_{ep})$ emanating from Q as in Fig. 3.1. A choice of q_o in Fig. 4.1 defines a set of mesh points in the base of the light cone and the rays from Q through these mesh points define our finite planar approximation to a uniform sample population of the cone's light.

Definition. We shall say that a sample ray in the light cone $C(h, \rho_0, d_{ep})$ vignettes when it fails to reach the photographic plate for any reason whatever.

A sample ray which traverses the lens and reaches the plate scores, in a vignetting sense, a successful event, while one which fails to reach the plate scores a failure. An ideal lens, which we strive to approximate at convergence, is characterized by the following properties at *convergence*:

1) There is *no vignetting* of any sample ray in the light cone $C(h, \rho_O, d_{ep})$ associated with any sample object point Q of height h.

2) There is *perfect achromatized focusing*, that is, all sample rays of all design colors β, $\beta=1, \cdot \cdot \cdot, n_{color}$, within the light cone $C(h, \rho_O, d_{ep})$ associated with each tested object point Q will focus on a unique image point Q'.

3) All design specifications will be met precisely. Such requirements may include assigned magnification, Gaussian focal length, f-number, oriented distance from the last vertex to the exit pupil, minimal and maximal allowable spacing between successive vertices and compatibility of the physical dimensions of the lens with those of the intended camera box.

6. CENTROIDAL ERROR COMPONENTS

We now begin defining the error vector $\underset{\sim}{E}$ for optical design. F. Wachendorf[5] formulated the focusing deficiency of a trial lens $\underset{\sim}{u}$ by a centroidal analysis of the spot diagram resulting from the multiple image of any test object point Q projected on the photographic plate by the sample ray population of Fig. 4.1 of specified color β. An important extension of the Wachendorf method was the outcome of a series of consultations between B. Brixner and J. C. Holladay to which the latter gave a mathematical formulation. The Holladay program , however, does not weight each sample ray according to the illumination which it represents as described for Fig. 4.1, nor does it provide an automatic correction of vignetting rays or steer the computation away from shaping a lens which cannot be made in a shop (the entrance pupil radius goes negative or some oriented spacing of successive vertices changes sign). We present in this and succeeding sections our own modification of the Brixner-Holladay extension of the Wachendorf method which recognizes the aspects of lens design just described.

Let Q_Γ, at height h_Γ in Fig. 3.1, $\Gamma=1, \cdot \cdot \cdot, n_{object}$, be a set of object test points chosen arbitrarily as a discrete representation of a continuous object line segment. From Q_Γ we trace a ray of color β, $\beta=1, \cdots, n_{color}$, aimed at an inclined trial entrance pupil mesh point resulting from some arbitrarily chosen q_o subdivision of Fig. 4.1. The ordered components of the illumination weight vector $\underset{\sim}{w}$ of (4.1) serve to identify the mesh points. In general, not all mesh points of $\underset{\sim}{w}$ determine a plate-reaching ray. We select that *subset* $\underset{\sim}{w}_{\beta\Gamma}$ of w which is plate-reaching, as determined by ray-tracing, where $\underset{\sim}{w}_{\beta\Gamma}$ has the components $w_{\tilde{E}}$, $\tilde{E}=1, \cdots, N_{\beta\Gamma}$, taken by the tracing order of Fig. 4.1 from the plate-reaching components of $\underset{\sim}{w}$.

The plate-reaching ray $L_{\tilde{E}\beta\Gamma}$ of color β aimed from the object point Q_Γ through the mesh point $\tilde{E}$ will image Q_Γ into the plate point with plate rectangular coordinates $(x_{\tilde{E}\beta\Gamma}, y_{\tilde{E}\beta\Gamma})$, where $(0, h_\Gamma)$ would be a plate point on a line through Q_Γ parallel to the optical axis. For every right-hand semicircular mesh point of Fig. 4.1 there is a companion left-hand semicircular mesh point which would yield the symmetrical companion image$(-x_{\tilde{E}\beta\Gamma}, y_{\tilde{E}\beta\Gamma})$. The $N_{\beta\Gamma}$ images of Q_Γ, together with their companions, form a plate *spot cloud* $D_{\beta\Gamma}$ of color β with centroid $(0, p_{\beta\Gamma})$, where

$$p_{\beta\Gamma} \equiv \Sigma_{\tilde{E}} w_{\tilde{E}} y_{\tilde{E}\beta\Gamma} / W_{\beta\Gamma}, \quad W_{\beta\Gamma} = \Sigma_{\tilde{E}} w_{\tilde{E}}, \tag{6.1}$$

(for convenience in presentation it is assumed here that not all sample rays vignette, so $W_{\beta\Gamma} \neq 0$). The first set of centroidal error components of $\underset{\sim}{E}$ are defined to be

$$(x_{\tilde{E}\beta\Gamma}-0) \text{ and } (y_{\tilde{E}\beta\Gamma}-p_{\beta\Gamma}) \text{ of weight } w_{\tilde{E}}. \qquad (6.2)$$

Definition. The statement that a component E_A of $\underset{\sim}{E}$ is of weight w_A means that $E_A \equiv w_A \xi_A$, where ξ_A measures a certain deficiency of the lens and w_A is a weight which we assign to ξ_A in the least squares process of minimizing $\phi \equiv |\underset{\sim}{E}|^2$.

The greater the weight $w(\xi)$ of an error measurement ξ, the more will ξ be driven toward 0 at the possible expense of other less weighted measurements. The vanishing of (6.2) for all $\tilde{E}$, with β and Γ fixed, would imply that the color spot cloud $D_{\beta\Gamma}$ has coalesced to coincide with its centroid $(0, p_{\beta\Gamma})$.

Next, consider the composite spot cloud $D_\Gamma \equiv \Sigma_\beta D_{\beta\Gamma}$ with centroid $(0, p_\Gamma)$, where

$$p_\Gamma \equiv \Sigma_\beta W_{\beta\Gamma} p_{\beta\Gamma} / W_\Gamma, \quad W_\Gamma \equiv \Sigma_\beta W_{\beta\Gamma}. \qquad (6.3)$$

The second set of centroidal error components of $\underset{\sim}{E}$ are defined to be

$$(p_{\beta\Gamma}-p_\Gamma) \text{ of assigned partial weight } c_{\beta\Gamma}. \qquad (6.4)$$

To define what is meant by the "partial weight $c(\xi)$ of an error measurement ξ," we first state that only the first set of centroidal components (6.2) of $\underset{\sim}{E}$ are associated immediately with known weights. We seek to automate the selection of the weight $w(\xi)$ of any other error measurement ξ such as (6.4) by giving it a data input partial weight $c(\xi)$. At the end of a complete ray-trace we shall arrive at a root mean square error number R_{rmsq}, defined later by Eq. (7.3), which measures the composite focusing error. This number R_{rmsq} having been determined, we then define

$$w(\xi)\equiv(c(\xi)/c_{rmsq})/(|\xi|/R_{rmsq}), \qquad (6.5)$$

where $c(\xi)$ and c_{rmsq} are data input constants, both chosen as 1 prior to testing for better values. Thus $w(\xi)$ is automatically set larger when $|\xi|>>R_{rmsq}$ and automatically approaches $c(\xi)/c_{rmsq}$ as $|\xi|\rightarrow R_{rmsq}$. It is hoped that testing will reveal our trial data entries $c(\xi)\equiv c_{rmsq}\equiv 1$ to be satisfactory and so justify their ultimate removal from input data. The coding device for coping with the delayed setting of weights implied by (6.5) will be described in section 30.

The vanishing of <u>both</u> (6.2) and (6.4) for given Γ and all $\tilde{E}$ and β would imply that a) each color spot cloud $D_{\beta\Gamma}$ has coalesced to coincide with its centroid $(0, p_{\beta\Gamma})$ and b) all the color centroids have coalesced into a <u>single</u> point $(0, p_{\Gamma})$, a perfect multicolored image of Q_{Γ}.

7. AREAL ERROR COMPONENTS

Aware that the design parameters comprising the components of the trial lens $\underset{\sim}{u}$ are generally too few in number to attain the simultaneous vanishing of the centroidal components (6.2) and (6.4), Brixner and Holladay continued to form additional error components in order to bring the image spot clouds into agreement with practical requirements of good focusing. Generalizing their treatment to pay heed to our introduction of weighted illumination and plate-reaching discrimination, a ray trace from Q_Γ of color β through the plate-reaching mesh points defined by $\underset{\sim}{w}_{\beta\Gamma}$ of the previous section enables us to form the weighted mean squared radius

$$R^2_{\beta\Gamma} \equiv \{\Sigma_{\underset{\sim}{E}} w_{\underset{\sim}{E}} [x^2_{\underset{\sim}{E}\beta\Gamma} + (y_{\underset{\sim}{E}\beta\Gamma} - p_{\beta\Gamma})^2]\}/W_{\beta\Gamma}, \quad W_{\beta\Gamma} \equiv \Sigma_{\underset{\sim}{E}} w_{\underset{\sim}{E}}, \tag{7.1}$$

of the spot cloud $D_{\beta\Gamma}$ of color β relative to its centroid $(0, p_{\beta\Gamma})$. We shall say that $R^2_{\beta\Gamma}$ measures the <u>area</u> of $D_{\beta\Gamma}$, a spot cloud of weight $W_{\beta\Gamma} \equiv \Sigma_{\underset{\sim}{E}} w_{\underset{\sim}{E}}$. The composite spot cloud $D_\Gamma \equiv \Sigma_\beta D_{\beta\Gamma}$ of weight W_Γ will then have the area

$$R^2_\Gamma \equiv (\Sigma_\beta W_{\beta\Gamma} R^2_{\beta\Gamma})/W_\Gamma, \qquad W_\Gamma \equiv \Sigma_\beta W_{\beta\Gamma}, \tag{7.2}$$

while the composite spot cloud $D \equiv \Sigma_\Gamma D_\Gamma$ of weight W will have the area

$$R^2_{rmsq} \equiv (\Sigma_\Gamma W_\Gamma R^2_\Gamma)/W, \qquad W \equiv \Sigma_\Gamma W_\Gamma. \tag{7.3}$$

The areal components of $\underset{\sim}{E}$ are defined to be

$$(R_\Gamma - R_{rmsq}) \text{ of assigned partial weight } c_\Gamma. \tag{7.4}$$

The vanishing of (7.4) would imply that each multicolored spot cloud D_Γ which images Q_Γ, $\Gamma=1, \cdot \cdot \cdot, n_{object}$, is of equal area.

8. CIRCULAR SHAPE ERROR COMPONENTS

The centroidal components (6.2) and (6.4) and the areal components (7.4) exhaust our generalizations of the spot cloud components of $\underset{\sim}{E}$ defined by Brixner and Holláday. It is possible, however, that the composite spot clouds D_Γ, having nearly the same area at convergence, may remain distorted from circular shape. To ensure both nearly equal and circular areas we form

$$\begin{aligned} X^2_{\beta\Gamma} &\equiv (\Sigma_{\underset{\sim}{E}} w_{\underset{\sim}{E}} x^2_{\underset{\sim}{E}\beta\Gamma})/W_{\beta\Gamma}, \quad & Y^2_{\beta\Gamma} &\equiv [\Sigma_{\underset{\sim}{E}} w_{\underset{\sim}{E}} (y_{\underset{\sim}{E}\beta\Gamma} - p_{\beta\Gamma})^2]/W_{\beta\Gamma}, \\ X^2_\Gamma &\equiv (\Sigma_\beta W_{\beta\Gamma} X^2_{\beta\Gamma})/W_\Gamma, \quad & Y^2_\Gamma &\equiv (\Sigma_\beta W_{\beta\Gamma} Y^2_{\beta\Gamma})/W_\Gamma, \end{aligned} \tag{8.1}$$

and define the <u>circular</u> <u>shape</u> components of $\underset{\sim}{E}$ to be

$$(X_\Gamma - Y_\Gamma) \text{ of assigned partial weight } c_\Gamma^{\mathrm{circ}}. \tag{8.2}$$

Referring to the three characterizations of an ideal lens in section 5, the error components (6.2), (6.4), (7.4) and (8.2) measure the departure of our trial lens $\underset{\sim}{u}$ from the ideal property 2), namely the perfect focusing of Q_Γ into its unique image Q'_Γ.

9. VIGNETTING ERROR COMPONENTS

We consider in this section those events which may be encountered during a ray-trace as supplying the first evidence that the sample ray will not traverse the entire lens system and reach the plate to form its image of the object point Q. The occurrence of such an event will be the signal to halt the further tracing of the errant ray, to form a vignetting component of the error vector $\underset{\sim}{E}$, and to start tracing the next ray in the mesh of Fig. 4.1.

1) *Discriminant component of* $\underset{\sim}{E}$. Looking ahead to Fig. 12.1 and anticipating that the refracting (or reflecting) surface Σ_i will be assumed to be a quadric of revolution, then the axial coordinate z of the refraction point $Q_i^*(x,y,z)$ will be determined as the solution of a quadratic equation with a discriminant $\bar{q}^2$ as defined by Eq. (13.1). We compare $\bar{q}^2$ with a bias input constant $\epsilon_1>0$ and form the *biased discriminant* error component

$\xi_1' \equiv \bar{q}^2 - \epsilon_1$ when $\bar{q}^2 < \epsilon_1$ of assigned partial weight c_1' (9.1)

2) *Feathering component of* $\underset{\sim}{E}$. If a pair of successive refracting surfaces (Σ_{i-1}, Σ_i) should intersect on the axis side of the ray segment cut off by them, the ray will pierce these surfaces in the *reverse* order (Σ_i, Σ_{i-1}) which will be said to constitute a *feathering* violation. We must consider two cases:

a) $d_{i-1}>0$, i.e. the ray has suffered an even number of reflections at the completion of its encounter with Σ_{i-1}. Define $F_i'\equiv d_{i-1}+z_i-z_{i-1}$ and compare F_i' with a bias input constant $\epsilon_2'>0$ and form the *biased feathering* error component

$\xi_2'\equiv F_i'-\epsilon_2'$ when $F_i'<\epsilon_2'$ of assigned partial weight c_2'; (9.2)

b) $d_{i-1}<0$, i.e. the ray has suffered an odd number of reflections at the completion of its encounter with Σ_{i-1}. Using the *same* bias input constant $\epsilon_2'>0$ of a), compare F_i' of a) with $-\epsilon_2'$ and form the *biased feathering* error component

$\xi_2'\equiv F_i'+\epsilon_2'$ when $F_i'>-\epsilon_2'$ of assigned partial weight c_2'. (9.3)

3) *Total reflection component of* $\underset{\sim}{E}$. In Eq. (13.6) we shall meet a second discriminant, $\bar{q}_1'^2$, with the optical significance that a ray is transmitted by a refracting surface Σ_i when $\bar{q}_i'^2>0$ and is totally reflected when $\bar{q}'^2\leq 0$. At a refracting surface we compare $\bar{q}'^2$ with an input bias constant $\epsilon_3'>0$ and form the *biased total reflection* error component

$\xi_3'\equiv\bar{q}_i'^2-\epsilon_3'$ when $\bar{q}'^2<\epsilon_3'$ of assigned partial weight c_3'. (9.4)

4) *Intolerable refraction point radius component of* $\underset{\sim}{E}$. As input data we enter $\underset{\sim}{R}^{max}\equiv(R_1^{max}, R_2^{max}, \cdot\cdot\cdot)$ whose i-th component measures the maximum refraction point radius relative to the axis that will be tolerated during a ray-

trace. At the refraction point (x_i, y_i, z_i) we compare $R_i \equiv (x_i^2+y_i^2)^{\frac{1}{2}}$ with R_i^{max} and form the intolerable refraction point radius error component

$$\xi_4' \equiv R_i - R_i^{max} \text{ when } R_i > R_i^{max} \text{ of assigned partial weight } c_4'. \quad (9.5)$$

This completes the description of the four vignetting components of $\underset{\sim}{E}$.

10. DESIGN DEMAND ERROR COMPONENTS

1) Magnification demand. From section 6 and Fig. 3.1 the object points $Q_\Gamma(0, h_\Gamma, 0)$, $\Gamma=1, \cdot \cdot \cdot, n_{object}$, are imaged into spot clouds D_Γ with centroids at the plate points $(0, p_\Gamma)$ as defined by (6.3). In terms of a data input magnification vector $\underset{\sim}{M}^{mag}$ with components M_Γ^{mag}, +(-) for inverted (upright) image, we form the Brixner-Holladay magnification demand components of $\underset{\sim}{E}$,

$$\xi_\Gamma^{mag} \equiv p_\Gamma + M_\Gamma h_\Gamma \text{ of assigned partial weight } c_\Gamma^{mag}. \qquad (10.1)$$

2) Second focal length demand. If a designer schooled in Gaussian optics should wish to prescribe the second focal length, f'_{assign}, of the system, he need merely trace a paraxial ray parallel to the axis at an object height $h=\Omega_{parax}\ \rho_0$, with $\Omega_{parax} \equiv 10^{-5}$ a likely choice, measure the observed second focal length f'_{obs} and form the Brixner-Holladay second focal length demand component of $\underset{\sim}{E}$,

$$\xi^f \equiv 1/f'_{obs} - 1/f'_{assign} \text{ of assigned partial weight } c^f. \qquad (10.2)$$

3) f-number demand. As a working definition of the f-number of a lens outside of the Gaussian domain, Brixner aims a meridional ray from Q_0, the intersection of the axis with the object plane Σ_0, to a normal entrance pupil point at the radial distance $\Omega\rho_0$, with $\Omega \equiv 1/\sqrt{2}$ a

likely choice, and observes the acute angle θ which the refracted ray makes with the axis on emerging from the system. He defines the working observed f-number to be

$$(f/\#)_{obs} \equiv \tfrac{1}{2}\Omega\cot\theta. \qquad (10.3)$$

We accordingly form the f-number demand component of $\underset{\sim}{E}$,

$$\xi^{f/\#} \equiv 1/(f/\#)_{obs} - 1/(f/\#)_{assign} \text{ of partial weight } c^{f/\#}. \qquad (10.4)$$

4) Exit pupil oriented distance demand. Here Brixner aims a principal ray from an object point at a specified height h above the axis. If the refracted output ray be observed to intersect the axis at an oriented distance $d_{expuobs}$ from the vertex of the last Σ_i, then we form the Brixner-Holladay exit pupil demand component of $\underset{\sim}{E}$,

$$\xi^{expu} \equiv 1/d_{expuobs} - 1/d_{expuassign} \text{ of assigned partial weight } c^{expu}. \qquad (10.5)$$

5) Oriented upper bound demand. Referring to $u^{-1} \equiv \rho_0$, $u^{-1} = d_{ep}$, $u^i \equiv d_1$ as defined by Eqs. (2.1), we shall bound the domain of design freedom of each of these parameters by entering as input data a lower oriented bound vector $\underset{\sim}{m} \equiv (m^{-2}, m^{-1}, \cdot\cdot\cdot)$ and an upper oriented bound vector $\underset{\sim}{M} \equiv (M^{-2}, M^{-1}, \cdot\cdot\cdot)$, where sign u^i=sign m_1^i=sign M_1. u^{-2}, u^{-1}, u^0 are always positive, but the spacings $u^i = d_1$ may be of either sign due to possible

reflections. We define the *oriented upper bound demand* components of $\underset{\sim}{E}$ by

$\xi^{up} \equiv u^{i} - M^{i}$ when $|u^{i}| > |M^{i}|$ of assigned partial weight c^{up}. (10.6)

It is not possible to constrain a design parameter to stay on the admissible side of its lower oriented bound by means of forming a lower bound error component, for change of sign might then occur during successive least squares adjustments and during such a sign reversal the machine would be appraising a lens of impossible construction. This absurdity is permitted by the Holladay code in the hope that the sign barrier may again be crossed and so permit the machine to resume design in the domain of a constructible lens. We shall discuss the treatment of the lower oriented bound $\underset{\sim}{m}$ imposed on $\underset{\sim}{u}$ through the d's in section 28.

A zero input data entry for a design demand is a signal to by-pass that demand as being of no design interest. Thus a designer who has severed his pre-computer ties with Gaussian optics will normally consider 2) to be of scarcely more than textbook interest.

11. ZOOM LENS DESIGN

We consider here a zoom lens previously designed by Brixner[6] using the Holladay code. It illustrates type I of three common zoom types I, II, III, where types II and III will be presented at the end of this section. Brixner found that the design shown schematically in Fig. 11.1

insert Fig. 11.1 approximately here

provides enough degrees of freedom to yield good results. A set of 2m+1 arbitrarily spaced object points $O_1, \cdot \cdot \cdot,$ O_{2m+1} are to be imaged into a common point O' by 2m+1 equally spaced positions of the sliding lens pair (B, D) joined by an inextensible bar of length L/2. For simplicity in drawing, Fig. 11.1 shows all 15 refracting surfaces to be plane and shows the sliding lenses B and D to be midway between the fixed lens pairs (A, C) and (C, E) respectively.

We assume now that we have selected a common maximum allowable glass thickness M compatible with the distance L/2 within which we must allot the 2m+1 equally spaced stations for the movable lens B. Letting Δ_B represent the common air space between successive stations of B, the condition on Δ_B for given L and common maximum glass thickness M is

$$3M(\text{end glass})+(2m+1)2M(\text{station glass B})+2(m+1)\Delta_B(\text{air gap})=\tfrac{1}{2}L, \qquad (11.1)$$

$\therefore \Delta_B = [\frac{1}{2}L-(4m+5)M]/2(m+1)$.

If we demand that $\Delta_B \geqq 2M$, then (11.1) places an upper bound on our choice of M, namely

$$M \leqq M_{max} \equiv (\tfrac{1}{2}L)/(8m+9), \quad \Delta_B = 2M \text{ for } M = M_{max}. \qquad (11.2)$$

The 2m+1 stations for B are now determined by the zoom station vector d_{3i}, i=1, · · ·, 2m+1, with the components

$$\Delta_B,\ \Delta_B+(\Delta_B+2M),\ \Delta_B+2(\Delta_B+2M),\ \cdot\cdot\cdot,\ \Delta_B+2m(\Delta_B+2M). \qquad (11.3)$$

To define a general zoom lens of type I we now relax Brixner's demand that all the object points O_i of height h_i image into a common point O'. We accomplish this by introducing data input station magnification demand components M_i, i=1, · · ·, 2m+1, and asking that the magnification error components $p_i+M_ih_i$ (i not summed) approach 0 at convergence, where (0, p_i) is the centroid of the multicolored spot cloud imaging O_i. If we choose M_i to satisfy M_ih_i=constant for all i, our generalized type I zoom lens will satisfy Brixner's imaging demand.

To implement a type I design we fix our attention upon an arbitrary one of the continuous object segments (O_0, O_i) and approximate this by a discrete data input object point spectrum $Q_{\Gamma i}$, Γ=1, · · ·, n_{object}, whose last point coincides with O_i. Doing this for each i, the data input object height vector of section 6 with components h_Γ now becomes generalized to a matrix of heights $h_{\Gamma i}$, $\Gamma=1, \cdot \cdot \cdot, n_{object}$, $i=1, \cdot \cdot \cdot, 2m+1$. The sample ray $L_{E\beta\Gamma}$ of section 6 aimed there from Q_{Γ} now becomes $L_{E\beta\Gamma i}$ aimed from the object point $Q_{\Gamma i}$. A review of the centroid analysis of section 6 shows that the extra index i must be added to the centroid quantities and to their corresponding error components in order to form the composite zoom stationed error vector $\underset{\sim}{E}$. Thus the components of $\underset{\sim}{E}$ corresponding to the zoom station i are measuring the departure from perfection of the lens when set at station i. The least squares process will minimize the zoom composite $\phi\equiv\underset{\sim}{E}\cdot\underset{\sim}{E}$ by stealing a bit from the performance at one station to improve that at another. The i-th zoom station should cause the type I magnification error components of the form

$${}_{I}\xi^{mag}_{\Gamma i}\equiv p_{\Gamma i}+M_i h_{\Gamma i} \text{ of partial weight } {}_{I}c^{mag}_{\Gamma i}, \qquad (11.4)$$

to approach 0, where $(0,p_{\Gamma i})$ is the centroid of the multicolored spot cloud imaging $Q_{\Gamma i}$ of height $h_{\Gamma i}$.

We begin the zoom design by locking the movable lens pair (B, D) of Fig. 11.1 in its <u>central</u> position corresponding to $i=m+1$ and proceed to optimize the non-zoom system of 5 lenses A, B, C, D, E relative to the discrete central object point spectrum $Q_{\Gamma m+1}$, $\Gamma=1, \cdot \cdot \cdot n_{object}$, with all 10 glass thicknesses free to vary independently within the constraints m$\leq$glass thickness$\leq$M and with the air gaps subject to the constraints

$$d_3=\tfrac{1}{4}L-d_1-d_2-d_4, \quad d_{12}=\tfrac{1}{4}L-d_{11}-d_{13}-d_{14},$$
$$d_6=\tfrac{1}{4}L-d_5-d_7, \quad d_9=\tfrac{1}{4}L-d_8-d_{10}, \tag{11.5}$$

of the type given by Eq. (2.2). Let $\underset{\sim}{u}_{\text{central}}$ be the resulting optimal parameter vector with components analogous to those of Eq. (2.1). The components of $\underset{\sim}{u}_{\text{central}}$ other than $u^3=d_3$ then define the initial trial lens in the minimization of the composite zoom error vector resulting from holding d_3 fixed at the sequence of values given by Eq. (11.3), as signaled to the program by setting the parameter flag component $f^3=-999$ as in 4) of section 2. For <u>each</u> <u>zoom</u> <u>station</u> <u>value</u> of the air gap d_3 the remaining air gaps are subject to the zoom constraints

$$d_6=(L/2-d_3)-d_1-d_2-d_4-d_5-d_7,$$
$$d_9=(d_3)+d_1+d_2+d_4-d_8-d_{10}, \tag{11.6}$$
$$d_{12}=(L/2-d_3)-d_1-d_2-d_4-d_{11}-d_{13}-d_{14},$$

of the type given by Eq. (2.2).

The zoom lens of type II results from placing the 2m+1 object points O_i on a <u>horizontal</u> line at height h above the optical axis rather than on a vertical line as in Fig. 11.1. Let O_{0i} be the foot of the perpendicular from O_i to the optical axis at the distance d_{0i} from Σ_1. For given i we approximate the continuous object segment

$(O_{0i}O_i)$ by a discrete set of object points $Q_{\Gamma i}$ of heights h_Γ, $\Gamma=1, \cdot\cdot\cdot, n_{object}$, and request that the i-th zoom station reduce the magnification error components $_{II}\xi^{mag}_{\Gamma i} \equiv p_{\Gamma i}+M_i h_\Gamma$, of assigned partial weight $_{II}c^{mag}_{\Gamma i}$, to small absolute values. (11.7)

The zoom lens of type III has but one object segment $(O_0 O_1)$ of height h which is approximated by a discrete spectrum of object points Q_Γ of height h_Γ, $\Gamma=1, \cdot\cdot\cdot, n_{object}$. Data input magnification demand components M_i, $i=1, \cdot\cdot\cdot, 2m+1$, serve to define the image requirements that the i-th zoom station should cause the near vanishing of the magnification error components $_{III}\xi^{mag}_{\Gamma i} \equiv p_{\Gamma i}+M_i h_\Gamma$ of assigned partial weight $_{III}c^{mag}_{\Gamma i}$.

12. HERZBERGER'S METHOD OF RAY-TRACING

We begin a summary of M. Herzberger's[7] method of ray-tracing. Figure 12.1 shows the input ray emerging

<u>insert Fig. 12.1 approximately here</u>

at Q_i from the plane E_i tangent to the optical surface Σ_i at its vertex O_i. This ray suffers refraction (or reflection) at Q^* on Σ_i and pierces E_{i+1} at Q_{i+1} as the input ray for the next refraction at Σ_{i+1}. We define the input vector $\underset{\sim}{s}_i$ by

$$\underset{\sim}{s}_i \equiv n_{i-1}\underset{\sim}{u}_{i-1} \equiv \underset{\sim}{S}_i + \zeta_{i-1}\underset{\sim}{e}_z \equiv (\Phi^{(i)}_{21}\underset{\sim}{e}_x + \Phi^{(i)}_{22}\underset{\sim}{e}_y) + \zeta_{i-1}\underset{\sim}{e}_z,$$
$$|\underset{\sim}{s}_i|^2 = n^2_{i-1} = \underset{\sim}{S}_i \cdot \underset{\sim}{S}_i + \zeta^2_{i-1}, \quad i=1, \cdot \cdot \cdot, n_{surf}. \qquad (12.1)$$

In terms of the transverse position vector $\underset{\sim}{A}_i$ giving the displacement of the tangent plane piercing point Q_i with respect to the vertex O_i of Σ_i, we define the proportional transverse vector

$$\underset{\sim}{P}_i \equiv \zeta_{i-1}\underset{\sim}{A}_i \equiv (\Phi^{(i)}_{11}\underset{\sim}{e}_x + \Phi^{(i)}_{12}\underset{\sim}{e}_y), \quad i=1, \cdot \cdot \cdot, n_{surf}. \qquad (12.2)$$

The matrix $\underset{\sim}{\Phi}^{(i)}$ completely defines the ray $\underset{\sim}{s}_i$ of Fig. 12.1 and will be called the <u>ray identification matrix</u>. The next input pair $(\underset{\sim}{P}_{i+1}, \underset{\sim}{S}_{i+1})$, identified by $\underset{\sim}{\Phi}^{(i+1)}$, is related to the previous input pair $(\underset{\sim}{P}_i, \underset{\sim}{S}_i)$ identified by $\Phi^{(i)}$, according to

$$\underset{\sim}{\Phi}^{(i+1)} = \underset{\sim}{\Theta}^{(i+1,i)}\underset{\sim}{\Phi}^{(i)}, \quad i=1, \cdot \cdot \cdot, n_{surf} \qquad (12.3)$$

The problem of ray-tracing as posed by Herzberger is the problem of computing the four elements of the <u>refraction</u>

matrix $\underset{\sim}{\Theta}^{(i+1,i)}$.

Referred to axes along the orthonormal triad $(\underset{\sim}{e}_x, \underset{\sim}{e}_y, \underset{\sim}{e}_z)$ of Fig. 12.1 with origin at the vertex O_i of Σ_i, the equation of Σ_i, restricted now to be a quadric of revolution about the optical axis, will be expressed as

$$\Sigma_i \equiv \Sigma(B_i, C_i): \quad z+\tfrac{1}{2}B_i C_i z^2-\tfrac{1}{2}B_i(x^2+y^2)=0. \tag{12.4}$$

We shall call (x, y, z) of Eq. (12.4) the <u>tangential</u> coordinates of a point of Σ_i since they are referred to the tangential frame $(O_i; \underset{\sim}{e}_x, \underset{\sim}{e}_y, \underset{\sim}{e}_z)$ of Fig. 12.1. We also introduce a parallel <u>absolute</u> frame with origin at O_0, the intersection of the object plane Σ_0 with the optical axis. The object point Q of Fig. 3.1 has the absolute coordinates (0, h, 0).

In order to determine $\underset{\sim}{\Phi}^{(1)}$ we set i=0 in Fig. 12.1 and recall that now there is no refraction at the object plane Σ_0 so $\underset{\sim}{s}_0 \equiv \underset{\sim}{s}_1 \equiv n_0 \underset{\sim}{u}$, where $\underset{\sim}{u}$ is the unit aiming vector with components given by Eq. (4.3). Inspection of Fig. 12.1 shows that

$$\underset{\sim}{A}_1=-d_0\underset{\sim}{e}_z+\underset{\sim}{A}_0+(d_0/\zeta_0)(\underset{\sim}{S}_1+\zeta_0\underset{\sim}{e}_z), \quad \zeta_0 \equiv n_0 u_z,$$

$$\therefore \underset{\sim}{P}_1=\zeta_0\underset{\sim}{A}_0+d_0\underset{\sim}{S}_1 \text{ from Eq. (12.2)} \tag{12.5}$$

From the definitions (12.1) and Eq. (12.5) it follows that

$$\begin{Vmatrix} \Phi_{11}^{(1)} & \Phi_{12}^{(1)} \\ \Phi_{21}^{(1)} & \Phi_{12}^{(1)} \end{Vmatrix} \equiv \begin{Vmatrix} n_0 d_0 u_x & n_0(d_0 u_y + h u_z) \\ n_0 u_x & n_0 u_y \end{Vmatrix}, \tag{12.6}$$

where $\underset{\sim}{u}$ is given by (4.3).

13. HERZBERGER'S COMPUTATIONAL PROCEDURE

We summarize the order of computation in making a Herzberger ray-trace.

1) Select a specific object point $Q(0, h_\Gamma, 0)$ from the input data set of object points Q_Γ, $\Gamma=1, \cdot \cdot \cdot, n_{object}$, introduced in section 6.

2) Select a specific inclined entrance pupil mesh refinement number q_0 as in Fig. 4.1.

3) Select a specific mesh point (i, j) with coordinates (x_0, y_0') given by Eq. (4.2) relative to the inclined entrance pupil frame $(0; \underset{\sim}{e}_x, \underset{\sim}{e}_y', \underset{\sim}{e}_z')$ of Fig. 3.1.

4) Use the definitions (4.3) to form the unit aiming vector $\underset{\sim}{u}$ referred to the absolute frame $(0; \underset{\sim}{e}_x, \underset{\sim}{e}_y, \underset{\sim}{e}_z)$ of Fig. 3.1.

5) Use the definitions (12.6) to form $\underset{\sim}{\Phi}^{(1)}$.

We now employ induction by assuming that we have formed the ray identification matrix $\underset{\sim}{\Phi}^{(i)}$ and that we presently seek to form Herzberger's refraction matrix $\underset{\sim}{\theta}^{(i+1,i)}$ in order to form $\underset{\sim}{\Phi}^{(i+1)}$ by means of Eq. 12.3). With the understanding that our attention is now fixed on the ray $\underset{\sim}{s}_i$ as defined by Eq. (12.1) which is about to be refracted by Σ_i, we simplify notation by omitting all indices in our computations.

6) Form the discriminant

$$\bar{q}^2 \equiv (B\underset{\sim}{P}\cdot\underset{\sim}{S}-\zeta^2)^2-B^2\underset{\sim}{P}\cdot\underset{\sim}{P}(\underset{\sim}{S}\cdot\underset{\sim}{S}-C\zeta^2), \quad \zeta\equiv\zeta_{i-1} \text{ in } (12.1), \qquad (13.1)$$

of the quadratic equation

$$[Bz(\underset{\sim}{S}\cdot\underset{\sim}{S}-C\zeta^2)+(B\underset{\sim}{P}\cdot\underset{\sim}{S}-\zeta^2)]^2=(B\underset{\sim}{P}\cdot\underset{\sim}{S}-\zeta^2)^2-B^2\underset{\sim}{P}\cdot\underset{\sim}{P}(\underset{\sim}{S}\cdot\underset{\sim}{S}-C\zeta^2) \qquad (13.2)$$

yielding the z coordinate of the refraction point Q^* relative to the tangential frame $(O_i;\ \underset{\sim}{e}_x,\ \underset{\sim}{e}_y,\ \underset{\sim}{e}_z)$ of Fig. 12.1.

7) Is $\bar{q}^2<\epsilon_1$ of (9.1)? If so, form the biased discriminant error component of (9.1) and start tracing the next ray in the mesh (i, j) of Fig. 4.1.

8) Form $\bar{q}\equiv(\bar{q}^2)^{\frac{1}{2}}$ and solve (13.2) by factoring to obtain

$$z=B\underset{\sim}{P}\cdot\underset{\sim}{P}/D, \quad D\equiv\bar{q}-B\underset{\sim}{P}\cdot\underset{\sim}{S}+\zeta^2\neq 0, \qquad (13.3)$$

which gives z=0 for B=0.

9) Compute the transverse coordinates of the refraction point $Q^*(x,\ y,\ z)$,

$$x=(\Phi_{11}+z\Phi_{21})/\zeta, \quad y=(\Phi_{12}+z\Phi_{22})/\zeta. \qquad (13.4)$$

10) Form $R\equiv(x^2+y^2)^{\frac{1}{2}}$ and test for magnitude. If too large, form the intolerable radius error component (9.5) and start tracing the next ray in the entrance pupil mesh of Fig. 4.1.

11) Form the feathering expression $F_i^!\equiv d_{i-1}+z_i-z_{i-1}$, test for feathering as in 2) of section 9 and, if feathering is revealed, form the feathering error component of section 9 and start tracing the next ray.

12) Form for future use the quantities

$$\bar{d}\equiv Bd,\ \bar{z}\equiv Bz,\ \bar{t}\equiv 1+(C+1)\bar{z},\ \bar{R}^2\equiv[\bar{t}^2-\bar{z}(\bar{t}-1)]\zeta^2>0. \tag{13.5}$$

13) Test the product $d_i d_{i-1}$ for sign. If negative, then Σ_i is a *mirror* and we define $\bar{q}'\equiv-\bar{q}$, where $\bar{q}$ is defined by 8), and go to 17).

14) A positive product $d_i d_{i-1}$ indicates refraction at Σ_i and so we form

$$\bar{q}'^2\equiv\bar{q}^2-(n^2-n'^2)\bar{R}^2,\ n\equiv n_{i-1},\ n'\equiv n_i, \tag{13.6}$$

and recognize, as Herzberger seemingly did not, that the condition for total reflection at Σ_i is $\bar{q}'^2\leqq 0$. To verify this, observe that from the refraction law $n\sin\theta=n'\sin\theta'$ for the case $n>n'$ the criticality condition $\sin^2\theta=n'^2/n^2$ may be expressed as $n^2-n'^2=(o\cdot s)^2$, where $\underset{\sim}{s}\equiv n\underset{\sim}{u}$, $|\underset{\sim}{u}|=1$, and $\underset{\sim}{o}$ is the unit normal to Σ as given by (12.4), directed from the medium of index n to that of index n'. The vanishing of $\bar{q}'^2$ in (13.6) would imply $n^2-n'^2=\bar{q}^2/\bar{R}^2$ and the equality $(\underset{\sim}{o}\cdot\underset{\sim}{s})^2=\bar{q}^2/\bar{R}^2$ may be established by observing that z satisfies Eq. (13.2).

15) Is $\bar{q}'^2<\epsilon_3'$ of (9.4)? If so, form the biased total reflection error component (9.4) and start tracing the next ray in the mesh (i, j).

16) Form $\bar{q}'\equiv(\bar{q}'^2)^{\frac{1}{2}}$.

17) Form $\bar{\Psi}\equiv(\bar{q}'-\bar{q})/\bar{R}^2$,

$$\left\|\begin{matrix}\theta_{11} & \theta_{12}\\ \theta_{21} & \theta_{22}\end{matrix}\right\| \equiv \left\|\begin{matrix}1+\bar{\Psi}(\bar{t}-\bar{d}) & d+z\bar{\Psi}(\bar{t}-\bar{d})\\ -B\bar{\Psi} & 1-Bz\bar{\Psi}\end{matrix}\right\|. \tag{13.7}$$

This definition of the refraction matrix $\underset{\sim}{\Theta}$ completes the computational description of Herzberger's method of ray-tracing.

14. APERTURE STOP DETERMINATION

In order to locate the aperture stop of a lens at the termination of convergence, Brixner and Holladay transfer machine control to a <u>twin ray</u> diagnostic subroutine which traces the two rays ${}_1\underset{\sim}{R}$ and ${}_2\underset{\sim}{R}$ of Fig. 14.1 aimed from a selected object point O(0, h, 0) and of prescribed color β.

<u>insert Fig. 14.1 approximately here</u>

${}_1\underset{\sim}{R}$ and ${}_2\underset{\sim}{R}$ are meridional rays aimed at the extreme ends of a diameter of the entrance pupil (d_{ep}, ρ_0) of Fig. 3.1 as defined at convergence. Figure 14.1 shows the twin rays as they transit the space between successive vertex tangent planes E_i and E_{i+1}, where by (12.1) and (12.2),

$${}_\alpha\underset{\sim}{S}_i \equiv {}_\alpha\Phi_{22}^{(i)}\underset{\sim}{e}_y + {}_\alpha\zeta_{i-1}\underset{\sim}{e}_z, \quad {}_\alpha\underset{\sim}{A}_i \equiv ({}_\alpha\Phi_{12}^{(i)}/{}_\alpha\zeta_{i-1})\underset{\sim}{e}_y, \quad \alpha=1,\ 2. \qquad (14.1)$$

Figure 14.1 shows a point Q on that segment of the optical axis cut out by E_i and E_{i+1} such that a circle with Q as center and plane normal to the axis cuts the direction lines of the twin rays in ${}_1Q$ and ${}_2Q$. For a given pair (E_i, E_{i+1}) there will usually be no such point Q between them, but when it does exist, and when in addition the medium between Σ_i and Σ_{i+1} is <u>air</u>, then Q determines the center of the required aperture stop of radius R_i.

Our point of view will be to allow Q to move along the entire axis interval $(-\infty, \infty)$ and to halt it when the oriented segments bear the desired relation $({}_{Q}\ {}_{1}Q)=-({}_{Q}\ {}_{2}Q)$. If z_i be the oriented spacing out to Q as in Fig. 14.1, then we shall have located the aperture stop when

$$z_i d_i > 0 \text{ and } |z_i| < |d_i| \text{ in an engineering sense.} \tag{14.2}$$

Given ${}_{\alpha}\underset{\sim}{s}_i$ and ${}_{\alpha}\underset{\sim}{A}_i$ as defined by (14.1), we seek properly chosen scalars λ and μ and a positive radius R_i such that

$$\begin{aligned} &R_i \underset{\sim}{e}_y = -z_i \underset{\sim}{e}_z + {}_1\underset{\sim}{A}_i + \lambda\, {}_1\underset{\sim}{s}_i = -(-z_i \underset{\sim}{e}_z + {}_2\underset{\sim}{A}_i + \mu\, {}_2\underset{\sim}{s}_i), \\ &-z_i + \lambda\, {}_1\zeta_{i-1} = -z_i + \mu\, {}_2\zeta_{i-1} = 0. \end{aligned} \tag{14.3}$$

Eliminating λ and μ we obtain

$$\begin{aligned} &z_i = -[{}_2\zeta_{i-1} \cdot {}_1\Phi_{12}^{(i)} + {}_1\zeta_{i-1} \cdot {}_2\Phi_{12}^{(i)}]/[{}_2\zeta_{i-1} \cdot {}_1\Phi_{22}^{(i)} + {}_1\zeta_{i-1} \cdot {}_2\Phi_{22}^{(i)}], \\ &R_i = [{}_1\Phi_{12}^{(i)} + z_i \cdot {}_1\Phi_{22}^{(i)}]/{}_1\zeta_{i-1}, \end{aligned} \tag{14.4}$$

when the denominator of z_i is not zero. Inspection of Fig. 14.1 shows that z_i becomes indeterminate when the twin rays are mirror conjugate with respect to the optical axis, meaning that

$${}_1\zeta_{i-1} = {}_2\zeta_{i-1}, \quad {}_1\Phi_{22}^{(i)} = -{}_2\Phi_{22}^{(i)}, \quad {}_1\Phi_{12}^{(i)} = -{}_2\Phi_{12}^{(i)}. \tag{14.5}$$

In such an event z_i may be chosen arbitrarily. When the first two equalities of (14.5) hold, but not the third, then there is no solution for z_i.

We have presented the determination of the aperture stop by the tracing of twin rays of chosen color β from a chosen object point of height h. It is anticipated that different stops will result from varying β and h, but it seems plausible that a well designed lens may yield a family of stops within engineering agreement.

15. ANALYTIC DIFFERENTIATION

The knowledge of all the first partial derivatives of the ray identification pair $(\underset{\sim}{P}_s, \underset{\sim}{S}_s)$ as defined by (12.1) and (12.2) at each vertex tangent plane E_s will enable us to differentiate all components of the error vector $\underset{\sim}{E}$. From Eq. (12.3),

$$\partial_i \underset{\sim}{\Phi}^{(s+1)} = \underset{\sim}{\Theta}^{(s+1,s)} \partial_i \underset{\sim}{\Phi}^{(s)} + \partial_i \underset{\sim}{\Theta}^{(s+1,s)} \underset{\sim}{\Phi}^{(s)}, \quad s = 1, \cdots, n_{surf}, \tag{15.1}$$

so that we initiate the induction on $\partial_i \underset{\sim}{\Phi}^{(s)}$ by differentiating $\underset{\sim}{\Phi}^{(1)}$ as defined by (12.6), where $\underset{\sim}{u}$ is defined by Eqs. (4.2) and (4.3). Preparing to form $\partial_i \underset{\sim}{\Phi}^{(1)}$, we recall the definitions (2.1) of the parameter vector $\underset{\sim}{u}$ of the Lister type lens of Fig. 2.1. From (12.6),

$$\partial_i \underset{\sim}{\Phi}^{(1)} \equiv \begin{Vmatrix} n_0(d_0\partial_i u_x + u_x \partial_i d_0) & n_0(d_0 \partial_i u_y + u_y \partial_i d_0 + h\partial_i u_z) \\ \\ n_0 \partial_i u_x & n_0 \partial_i u_y \end{Vmatrix}, \quad i = -2, -1, 0, \tag{15.2}$$

$$\partial_i \underset{\sim}{\Phi}^{(1)} \equiv 0, \quad i \geq 1.$$

From Eqs. (4.2) and (4.3) we obtain

$$\partial_{-2} u_x \equiv (j-\tfrac{1}{2}) q_0^{-1} M_{ag}^{-1} + x_0 \partial_{-2} M_{ag}^{-1},$$

$$\partial_{-2} u_y \equiv (i-\tfrac{1}{2}) q_0^{-1} \cos\theta M_{ag}^{-1} + (y_0' \cos\theta - h) \partial_{-2} M_{ag}^{-1},$$

$$\partial_{-2} u_z \equiv (i-\tfrac{1}{2}) q_0^{-1} \sin\theta M_{ag}^{-1} + (d_{ep} + y_0' \sin\theta) \partial_{-2} M_{ag}^{-1},$$

$$\partial_{-2} M_{ag}^{-1} \equiv -q_0^{-1} M_{ag}^{-3} [(j-\tfrac{1}{2}) x_0 + (i-\tfrac{1}{2})(y_0' + d_{ep} \sin\theta - h\cos\theta)],$$

$$\partial_{-1} u_x \equiv x_0 \partial_{-1} M_{ag}^{-1},$$

$$\partial_{-1}u_y \equiv y_0' M_{ag}^{-1}\partial_{-1}\cos\theta + M_{ag}u_y\partial_{-1}M_{ag}^{-1}, \tag{15.3}$$

$$\partial_{-1}u_z \equiv (1+y_0'\partial_{-1}\sin\theta)M_{ag}^{-1} + M_{ag}u_z\partial_{-1}M_{ag}^{-1},$$

$$\partial_{-1}\sin\theta \equiv - h^{-1}\sin^2\theta\cos\theta, \quad \partial_{-1}\cos\theta \equiv h^{-1}\sin^3\theta,$$

$$\partial_{-1}M_{ag}^{-1} \equiv - M_{ag}^{-2}[u_y y_0'\partial_{-1}\cos\theta + u_z(1+y_0'\partial_{-1}\sin\theta)],$$

$$\partial_i d_0 = \delta_{i0}, \quad \partial_0 u_x = \partial_0 u_y = \partial_0 u_z = 0.$$

Using (15.2) and (15.3) we may evaluate the non-vanishing $\partial_i\underset{\sim}{\Phi}^{(1)}$, $i = -2, -1, 0$. We may then use (15.1) to evaluate successively $\partial_i\underset{\sim}{\Phi}^{(s+1)}$ for $s = 1,2,\cdots$, providing that we can form $\partial_i\underset{\sim}{\Theta}^{(s+1,s)}$ for $s = 1,2,\cdots$. Now (13.7) gives the elements of the refraction matrix $\underset{\sim}{\Theta}^{(s+1,s)}$ for $s = 1,2,\cdots$ when we associate the surface index s of Σ_s with $\bar{\psi},\bar{t},\bar{d},z,B$ in (13.7). Assuming that $\underset{\sim}{\Phi}^{(s)}$ and $\partial_i\underset{\sim}{\Phi}^{(s)}$ are known, and recalling from (12.1) and (12.2) that

$$\underset{\sim}{P}_s \equiv \Phi_{11}^{(s)}\underset{\sim}{e}_x+\Phi_{12}^{(s)}\underset{\sim}{e}_y, \quad n_{s-1}^2 = \underset{\sim}{S}_s\cdot\underset{\sim}{S}_s + \zeta_{s-1}^2,$$

$$\underset{\sim}{S}_s \equiv \Phi_{21}^{(s)}\underset{\sim}{e}_x + \Phi_{22}^{(s)}\underset{\sim}{e}_y,$$

we seek to form $\partial_i\underset{\sim}{\Theta}^{(s+1,s)}$. Dropping the refracting surface index s for convenience, we seek the derivative of $\underset{\sim}{\Theta}$ as defined by (13.7). From our assumed knowledge of $\underset{\sim}{\Phi}$ and $\partial_i\underset{\sim}{\Phi}$ we may form

$$\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{P}) \equiv 2(\Phi_{11}\partial_i\Phi_{11}+\Phi_{12}\partial_i\Phi_{12}),$$

$$\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{S}) \equiv \Phi_{11}\partial_i\Phi_{21}+\Phi_{21}\partial_i\Phi_{11} + \Phi_{12}\partial_i\Phi_{22} + \Phi_{22}\partial_i\Phi_{12},$$

$$\partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S}) \equiv 2(\Phi_{21}\partial_i\Phi_{21} + \Phi_{22}\partial_i\Phi_{22}), \quad \partial_i\zeta^2 \equiv - \partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S}),$$

$$\partial_i\bar{q} \equiv \tfrac{1}{2}\bar{q}^{-1}\{2(B\underset{\sim}{P}\cdot\underset{\sim}{S}-\zeta^2)[(\underset{\sim}{P}\cdot\underset{\sim}{S})\partial_i B + B\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{S}) + \partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S})]$$

$$- [2B(\underset{\sim}{P}\cdot\underset{\sim}{P})\partial_i B + B^2\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{P})](\vec{S}\cdot\vec{S}-C\zeta^2)$$

$$- B^2(\underset{\sim}{P}\cdot\underset{\sim}{P})[(1+C)\breve{\partial}_i(\underset{\sim}{S}\cdot\underset{\sim}{S})-\zeta^2\partial_i C]\}, \text{ from (13.1)},$$

$$\partial_i D \equiv \partial_i\bar{q} - (\underset{\sim}{P}\cdot\underset{\sim}{S})\partial_i B - B\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{S}) - \partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S}), \text{ from (13.3)},$$

$$\partial_i z \equiv D^{-1}[(\underset{\sim}{P}\cdot\underset{\sim}{P})\partial_i B + B\partial_i(\underset{\sim}{P}\cdot\underset{\sim}{P}) - B(\underset{\sim}{P}\cdot\underset{\sim}{P})D^{-1}\partial_i D],$$

$$\partial_i(\bar{t}-\bar{d}) \equiv Bz\partial_i C+(C+1)(B\partial_i z+z\partial_i B)-(B\partial_i d+d\partial_i B), \text{ from (13.5)},$$

$$\partial_i\bar{R}^2 \equiv [(2\bar{t}-\bar{z})\partial_i\bar{t}-(\bar{t}-1)(B\partial_i z+z\partial_i B)]\zeta^2-\bar{R}^2\zeta^{-2}\partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S}) \text{ from (13.5)},$$

$$\partial_i\bar{q}' = - \partial_i\bar{q} \text{ for reflection at } \Sigma \text{ from 13) section 13},$$

$$\partial_i\bar{q}' \equiv \tfrac{1}{2}\bar{q}'^{-1}[2\bar{q}\partial_i\bar{q}-(n^2-n'^2)\partial_i\bar{R}^2] \text{ for refraction at } \Sigma \text{ from (13.6)},$$

$$\partial_i\bar{\psi} \equiv \bar{R}^{-2}[\partial_i\bar{q}'-\partial_i\bar{q} - \bar{\psi}\partial_i\bar{R}^2] \qquad \text{from (13.7)}.$$

From these equations and (13.7) we obtain the desired derivatives

$$\begin{aligned}
\partial_i\Theta_{11} &\equiv \bar{\psi}\partial_i(\bar{t}-\bar{d}) + (\bar{t}-\bar{d})\partial_i\bar{\psi},\\
\partial_i\Theta_{12} &\equiv \partial_i d + \bar{\psi}(\bar{t}-\bar{d})\partial_i z + z(\bar{t}-\bar{d})\partial_i\bar{\psi} + z\bar{\psi}\partial_i(\bar{t}-\bar{d}),\\
\partial_i\Theta_{21} &\equiv - (\bar{\psi}\partial_i B+B\partial_i\bar{\psi}),\\
\partial_i\Theta_{22} &\equiv - (z\bar{\psi}\partial_i B + B\bar{\psi}\partial_i z + Bz\partial_i\bar{\psi}).
\end{aligned} \tag{15.4}$$

The method of combined ray-tracing and analytic differentiation is now clear. A ray is traced from an object point Q to the plane E_1 of Fig. 12.1 and is defined at Q_1, its piercing point of E_1, by the ray identification matrix of Eq. (12.6), $\underset{\sim}{\Phi}^{(1)}$, sensitive only to the 3 parameters (ρ_0, d_{ep}, d_0). We interrupt the ray-trace at E_1 to form $\partial_i\underset{\sim}{\Phi}^{(1)}$ by the definition (15.2). We then resume the ray-trace by forming the refraction matrix $\underset{\sim}{\Theta}^{(2,1)}$ at Σ_1 as defined by (13.7). We observe that $\underset{\sim}{\Theta}^{(2,1)}$ is sensitive only to the parameters $(\rho_0, d_{ep}, d_0; d_1; B_1; C_1)$. We then use Eq. (12.3) to form $\underset{\sim}{\Phi}^{(2)} \equiv \underset{\sim}{\Theta}^{(2,1)}\underset{\sim}{\Phi}^{(1)}$ and differentiation by means of Eqs. (15.2) through (15.4) provides $\partial_i\underset{\sim}{\Phi}^{(2)}$. Successively determining the matrix pairs $(\underset{\sim}{\Phi}^{(s)}, \partial_i\underset{\sim}{\Phi}^{(s)})$, both sensitive only to the parameters

$$(d_0, d_{ep}, d_0; d_1, \cdots, d_{s-1}; B_1, \cdots, B_{s-1}; C_1, \cdots, C_{s-1}),$$

at each vertex tangent plane E_s, we eventually obtain $\underset{\sim}{\Phi}$ and $\partial_i\underset{\sim}{\Phi}$ at the vertex tangent plane E of the curved photographic plate $\Sigma(B,C)$. For a plane plate $E = \Sigma(0,-1)$. We now use (13.4) to compute the plate spot (x,y), and differentiation gives

$$\begin{aligned} \partial_i x &\equiv \zeta^{-1}[\partial_i\Phi_{11} + \Phi_{21}\partial_i z + z\partial_i\Phi_{21} + \tfrac{1}{2}\zeta^{-1}x\partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S})], \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15.5) \\ \partial_i y &\equiv \zeta^{-1}[\partial_i\Phi_{12} + \Phi_{22}\partial_i z + z\partial_i\Phi_{22} + \tfrac{1}{2}\zeta^{-1}y\partial_i(\underset{\sim}{S}\cdot\underset{\sim}{S})]. \end{aligned}$$

The plate spot derivatives (15.5) and the knowledge of the ray identification matrix $\underset{\sim}{\Phi}^{(s)}$ at each vertex tangent plane E_s and of its derivatives $\partial_i\underset{\sim}{\Phi}^{(s)}$ enable us to differentiate all the components of $\underset{\sim}{E}$ as defined in sections 6, 7, 8, 9, and 10.

16. EUCLIDEAN M-SPACE $\mathcal{E}_M$

Let $\underset{\sim}{e}_A$, $A = 1, \cdots, M$, be an orthonormal set of base vectors spanning Euclidean M-space $\mathcal{E}_M$. This means that the scalar product $\underset{\sim}{e}_A \cdot \underset{\sim}{e}_B$ is given by $\underset{\sim}{e}_A \cdot \underset{\sim}{e}_B = \delta_{AB}$, where $\delta_{AB} = 1(0)$ for $A = (\neq)B$. Any matrix $\|F_{AB}\|$ of constant elements for which $|F_{AB}| \neq 0$ defines a new set of base vectors $\underset{\sim}{f}_A$ spanning $\mathcal{E}_N$ defined by $\underset{\sim}{f}_A \equiv \underset{\sim}{e}_Q F_{QA}$ (sum repeated index over its range 1, $\cdots$, M) whose scalar products define a positive-definite matrix $\|G_{AB}\|$, where

$$G_{AB} \equiv \underset{\sim}{f}_A \cdot \underset{\sim}{f}_B \equiv F_{QA} F_{QB}. \tag{16.1}$$

These new base vectors $\underset{\sim}{f}_A$ are in general skew and of non-unit length, the conditions that they themselves be likewise orthonormal and obtainable from the orthonormal base $\underset{\sim}{e}_A$ by a rigid rotation in $\mathcal{E}_N$ being $G_{AB} \equiv F_{QA} F_{QB} = \delta_{AB}$. We shall assume that $\|F_{AB}\|$ is not a rotation matrix and shall agree that $\underset{\sim}{e}_A$ will designate an orthonormal base of $\mathcal{E}_M$ and $\underset{\sim}{f}_A$ a skew base. If $\underset{\sim}{V}$ be any vector of $\mathcal{E}_M$, its components may be referred either to $\underset{\sim}{e}_A$ or $\underset{\sim}{f}_A$,

$$\begin{gathered} \underset{\sim}{V} = \underset{\sim}{e}_Q E_Q = \underset{\sim}{f}_R V^R = \underset{\sim}{e}_Q F_{QR} V^R, \quad E_A = F_{AR} V^R, \\ |\underset{\sim}{V}|^2 = \delta_{QR} E_Q E_R = G_{QR} V^Q V^R. \end{gathered} \tag{16.2}$$

The Kronecker delta, δ_{AB}, gives the components of the Euclidean metric tensor $\underset{\sim}{G}$ in the base $\underset{\sim}{e}_A$ while G_{AB} gives the components of this same tensor in the base $\underset{\sim}{f}_A$.

If we confine our attention to $\mathcal{E}_3$ and a rectangular Cartesian system $(0; \underset{\sim}{e}_1, \underset{\sim}{e}_2, \underset{\sim}{e}_3)$ with origin at 0 and axes along the orthonormal base directions $\underset{\sim}{e}_A$, we are concerned with the Cartesian analytical description of the space of Euclidean geometry. It is rich in geometric concepts associated with points, distances between points, lines, angles between lines, curves, planes, surfaces, gradient direction at a point on a surface, etc.

17. ARITHMETIC SPACE A_n

Arithmetic n-space A_n is as poor in geometry as Euclidean M-space is rich. An arithmetic point is merely an ordered set of n numbers, $\underset{\sim}{u} \equiv (u^1, \dots, u^n)$, and the totality of such ordered sets constitutes A_n by sterile definition. It is a grinding task to squeeze any fruitful ideas out of A_n because it has no structure, no metric which serves to measure the "distance" between neighboring arithmetic points $\underset{\sim}{u}$ and $\underset{\sim}{u} + d\underset{\sim}{u}$. Our trial lens has been idealized as a point of A_n. As a prerequisite for measuring the distance between a neighboring pair of lenses $(\underset{\sim}{u}, \underset{\sim}{u} + d\underset{\sim}{u})$ taken from A_n we must relieve the poverty of A_n by endowing it from outside itself with a metric, a commodity which it does not have in its own right. This may be accomplished by mapping the geometrically barren A_n upon the bountiful $\mathcal{E}_M$.

18. GENERAL MAP OF A_n ONTO $\mathcal{E}_M$

To map the arithmetic points $\underset{\sim}{u}$ of A_n into the geometric points $\underset{\sim}{V}$ of Euclidean space $\mathcal{E}$, we make an arbitrary choice of the mapping arguments $[M, F_{AB}, V^A(\underset{\sim}{u})]$, where M defines the dimension of $\mathcal{E}$, F_{AB} defines a skew base $\underset{\sim}{f}_A \equiv \underset{\sim}{e}_Q F_{QA}$ in $\mathcal{E}$, and $V^A(\underset{\sim}{u})$ are differentiable functions defining the mapping according to the equations

$$\underset{\sim}{u}\epsilon A_n \to \underset{\sim}{V} \equiv \underset{\sim}{e}_R F_{RQ} V^Q(\underset{\sim}{u}) \equiv \underset{\sim}{f}_Q V^Q(\underset{\sim}{u})\epsilon\mathcal{E}_M, \quad M \geq n. \qquad (18.1)$$

As $\underset{\sim}{u}$ ranges over n-dimensional A_n its geometric image point $\underset{\sim}{V}(\underset{\sim}{u})$ ranges over a hypersurface $S_n\epsilon\mathcal{E}_M$ swept out by the free terminus of the displacement vector $\underset{\sim}{V}$. The tangent n-plane $\Sigma_n(\underset{\sim}{u})$ to S_n at $V(\underset{\sim}{u})\epsilon S_n$ is spanned by the base vectors

$$\underset{\sim}{\sigma}_i(\underset{\sim}{u}) \equiv \underset{\sim}{f}_Q V^Q_{,i}(\underset{\sim}{u}), \quad V^A_{,i} \equiv \partial_i V^A, \qquad (18.2)$$

and its maximum possible dimension will be n. At a singular point $\underset{\sim}{u}\epsilon A_n$, for which rank $\|V^A_{,i}(\underset{\sim}{u})\| = r < n$, the linear tangent space of S_n is r-dimensional.

19. METRIC INDUCED IN A_n

The neighboring arithmetic pair $(\underset{\sim}{u}, \underset{\sim}{u} + d\underset{\sim}{u})$ defines an infinitesimal n-tuple $d\underset{\sim}{u} \epsilon A_n$ which maps under (18.1) by

$$d\underset{\sim}{u} \epsilon A_n \rightarrow d\underset{\sim}{V}(\underset{\sim}{u}, d\underset{\sim}{u}) \equiv \sigma_r(\underset{\sim}{u}) du^r \epsilon \Sigma_n(\underset{\sim}{u}) \tag{19.1}$$

In the sense of the mapping we may now associate with $d\underset{\sim}{u} \epsilon A_n$ a scalar magnitude $|d\underset{\sim}{u}|$ defined by

$$\begin{aligned} |d\underset{\sim}{u}|^2 &\equiv |d\underset{\sim}{V}(\underset{\sim}{u}, d\underset{\sim}{u})|^2 \equiv g_{rs}(\underset{\sim}{u}) du^r du^s, \\ g_{ij}(\underset{\sim}{u}) &\equiv \underset{\sim}{\sigma}_i(\underset{\sim}{u}) \cdot \underset{\sim}{\sigma}_j(\underset{\sim}{u}) \equiv G_{QR} V^Q{}_{,i}(\underset{\sim}{u}) V^R{}_{,j}(\underset{\sim}{u}). \end{aligned} \tag{19.2}$$

One says that the mapping (18.1) has <u>induced</u> the metric g_{ij} of (19.2) on A_n, itself devoid of any metric.

20. DIRECTION OF STEEPEST DESCENT

There has been much discussion in the literature of "the direction of steepest descent" at a specified trial point $\underset{\sim}{u}\epsilon A_n$ of a scalar function $\varphi(\underset{\sim}{v})$, $\underset{\sim}{v}$ variable over A_n. All such discussions that have come to our attention have shared a common oversight, a failure to define the word "steepest." We present here a correct formulation of steepest descent.

Since steepness is a metric concept, there can be no measurement of the instantaneous rate of change of a function $\varphi(\underset{\sim}{v})$ as the arithmetic point $\underset{\sim}{v}$ moves along an arithmetic curve $C\epsilon A_n$ until *after* a metric has been induced in A_n by a mapping of A_n into Euclidean space. We accordingly assume that the mapping (18.1) has been applied so that we have thereby induced in A_n the metric $g_{ij}(\underset{\sim}{u})$ of (19.2). The situation is now that we find ourselves at a trial arithmetic point $\underset{\sim}{u}\epsilon A_n$ at which our φ takes the value $\varphi(\underset{\sim}{u})$ and we consider an arbitrary arithmetic curve $C\epsilon A_n$ on the trial point $\underset{\sim}{u}$. Equations of the form $v^i = v^i(s)$, $v^i(0) = u^i$, parameterize C with respect to arc length s along C when the arithmetic tangent $\underset{\sim}{t}$ with components $t^i \equiv dv^i/ds|_0$ is a *unit vector*, the condition from (19.2) being $g_{rs}(\underset{\sim}{u})t^r t^s = 1$. With arc length now defined, the directional derivative $d\varphi/ds|_0 = \varphi_{,r}(\underset{\sim}{u})t^r$ measures the initial rate of change of φ with respect to the arc length parameter s as we evaluate $\varphi(\underset{\sim}{v})$ along C issuing from $\underset{\sim}{u}$. With the observer at the trial $\underset{\sim}{u}\epsilon A_n$, there exists a family of ∞^{n-1} such unit directions $\underset{\sim}{t}$ issuing from $\underset{\sim}{u}$ and we seek that $\underset{\sim}{t}$ from the family along which the directional derivative $\varphi_{,r}(\underset{\sim}{u})t^r$ is *stationary*. Now the problem of seeking a stationary value of $\varphi_{,r}(\underset{\sim}{u})t^r$ as $\underset{\sim}{t}$ ranges over the unit sphere $g_{rs}(\underset{\sim}{u})t^r t^s - 1 = 0$ may be solved by introducing a Lagrange multiplier[8] λ to form $\psi(t,\lambda) \equiv \lambda(g_{rs}t^r t^s - 1) + \varphi_{,r}t^r$ and by imposing the conditions $\frac{1}{2}\partial\psi/\partial t^i = 0$, namely

$$g_{ir}(\underset{\sim}{u})z^r + \tfrac{1}{2}\,\varphi_{,i}(\underset{\sim}{u}) = 0, \quad z^i = \lambda t^i. \tag{20.1}$$

These are the equations of steepest descent for a function $\varphi(\underset{\sim}{v})$ at a trial point $\underset{\sim}{u}\epsilon A_n$ relative to the metric $g_{ij}(\underset{\sim}{u})$ of (19.2) induced on A_n by the mapping (18.1) of A_n into $\mathcal{E}_M$.

When $\underset{\sim}{u}$ is a non-singular point of the mapping (18.1), for which rank $\|V^A_{;i}(\underset{\sim}{u})\| = n$, then $|g_{ij}(\underset{\sim}{u})| \neq 0$ from the definition of g_{ij} in (19.2). In this case the steepest descent system (20.1) has a unique solution $\underset{\sim}{z}$. We eliminate the Lagrange multiplier λ from $z^i = \lambda t^i$ in (20.1) to obtain as the unit direction of steepest descent $t^i = z^i/|\underset{\sim}{z}|$, where from (16.1) and (19.2)

$$|\underset{\sim}{z}|^2 \equiv g_{rs}z^r z^s \equiv F_Q F_Q > 0, \quad F_A \equiv F_{AR}V^R_{,r}(\underset{\sim}{u})z^r, \tag{20.2}$$

which yields the minimizing directional derivative

$$\varphi_{,r}t^r = \varphi_{,r}z^r/|\underset{\sim}{z}| = -\,2g_{rs}z^r z^s/|\underset{\sim}{z}| = -\,2|\underset{\sim}{z}|,$$

when $\varphi_{,i}$ is eliminated by (20.1).

21. PRIMITIVE MAP OF A_n INTO $\mathcal{E}_n$

We shall say that the choice of mapping arguments $[M \equiv n,\ F_{ij} \equiv \delta_{ij},\ V^i(\underset{\sim}{u}) \equiv u^i]$ of section 18 defines the <u>primitive</u> map of A_n into $\mathcal{E}_n$ for which (18.1), (19.2), and (20.1) reduce to

$$\begin{aligned} &\underset{\sim}{u}\epsilon A_n \rightarrow \underset{\sim}{V} \equiv \underset{\sim}{e}_r u^r \epsilon \mathcal{E}_n, \\ &g_{ij}(\underset{\sim}{u}) \equiv \delta_{ij}, \quad z^i = -\tfrac{1}{2}\varphi_{,i}(\underset{\sim}{u}). \end{aligned} \tag{21.1}$$

Otherwise expressed, the primitive map of A_n into $\mathcal{E}_n$ results from setting up a rectangular Cartesian coordinate system in Euclidean $\mathcal{E}_n$ and defining the image of $\underset{\sim}{u}\epsilon A_n$ to be the geometric point with rectangular Cartesian coordinates $(u^1, \cdots, u^n)$. This primitive mapping yields Cauchy's[9] formulation of the direction of steepest descent of φ as that of the negative gradient of φ. Since the derived mapping matrix $\|V^i_{,j}(\underset{\sim}{u}) \equiv \delta^i_j\|$ is now the identity matrix, <u>all</u> points $\underset{\sim}{u}\epsilon A_n$ are nonsingular points of the mapping. The simplicity of the metric $g_{ij} = \delta_{ij}$ causes the steepest descent system (20.1) to appear in the solved form of (21.1).

22. ORTHONORMAL ERROR MAP OF A_n INTO $\mathcal{E}_N$

The primitive mapping of the previous section appeared conspicuously as the mapping of least imagination. As a mapping for optimizing a lens system, it ignores the intrinsic error vector $\underset{\sim}{E}$ of components E_A, $A = 1, \cdots, N$. In order to enrich the mapping by the inclusion of $\underset{\sim}{E}$, we now choose the mapping arguments $[M \equiv N,\ F_{AB} \equiv \delta_{AB},\ V^A(\underset{\sim}{u}) \equiv E_A(\underset{\sim}{u})]$ as defining the *orthonormal error* map of A_n into $\mathcal{E}_N$ for which (18.1), (19.2), and (20.1) become

$$\begin{gathered} \underset{\sim}{u} \epsilon A_n \to \underset{\sim}{E} \equiv \underset{\sim}{e}_Q E_Q(\underset{\sim}{u}) \epsilon \mathcal{E}_N, \\ g_{ir} z^r + \tfrac{1}{2}\varphi_{,i} = 0, \quad g_{ij}(\underset{\sim}{u}) \equiv E_{Q,i}(\underset{\sim}{u}) E_{Q,j}(\underset{\sim}{u}). \end{gathered} \tag{22.1}$$

Here also we are open to the charge of lacking imagination, for why have we chosen to treat our error components $E_A(\underset{\sim}{u})$ as components of $\underset{\sim}{E}$ relative to an *orthonormal* base $\underset{\sim}{e}_A$ rather than relative to some intrinsically chosen skew base $\underset{\sim}{f}_A(\underset{\sim}{u}) \equiv \underset{\sim}{e}_Q F_{QA}(\underset{\sim}{u})$, with $F_{AB}(\underset{\sim}{u})$ presumably dependent on the current trial lens $\underset{\sim}{u} \epsilon A_n$ which we seek to improve? There does indeed exist such an intrinsic skew base $\underset{\sim}{f}_A(\underset{\sim}{u})$, but its computation for the optics case $N \gg n$ is too formidable to be practical for today's computers and so we shall present this in another paper.

23. GEOMETRY OF ERROR MAPPING

A significant achievement of the error mapping leading to (22.1) is that the function φ which we seek to minimize is now intrinsically related to the mapping $[N,\delta_{AB},E_A(\underset{\sim}{u})]$ by $\varphi \equiv |\underset{\sim}{E}|^2$. This circumstance enables us to derive by geometric reasoning the steepest descent system

$$E_{Q,i}E_{Q,r}z^r + E_QE_{Q,i} = 0 \tag{23.1}$$

for $\varphi \equiv |\underset{\sim}{E}|^2$ relative to this mapping. Namely, as $\underset{\sim}{v}$ ranges over A_n its geometric image ranges over the hypersurface $S_n \epsilon \mathcal{E}_N$ swept out by the free terminus of the mapping vector $\underset{\sim}{E} \equiv \underset{\sim}{e}_Q E_Q(\underset{\sim}{u})$ with its fixed terminus at the origin O of a rectangular Cartesian coordinate system in $\mathcal{E}_N$ with axes in the directions of $\underset{\sim}{e}_A$, $A = 1, \cdots, N$. The current trial lens $\underset{\sim}{u}$ is imaged into the geometric point $\underset{\sim}{E}(\underset{\sim}{u}) \epsilon S_n$. We shall minimize $\varphi(\underset{\sim}{v}) \equiv |E(\underset{\sim}{v})|^2$ in the neighborhood of our current trial $\underset{\sim}{u}$ by finding that $\underset{\sim}{u}_m \epsilon A_n$ for which $\underset{\sim}{E}(\underset{\sim}{u}_m)$ is that point of S_n in the neighborhood of $E(\underset{\sim}{u}) \epsilon S_n$ which is the closest to O. Seeking an arithmetic n-tuple $\underset{\sim}{z} \epsilon A_n$ such that the arithmetic line $\underset{\sim}{u} + \eta \underset{\sim}{z}$, $\eta > 0$, will give the direction of steepest descent of $\varphi \equiv |\underset{\sim}{E}|^2$ at $\eta = 0$, we replace the curved $S_n \epsilon \mathcal{E}_N$ by its tangent n-plane $\Sigma_n(\underset{\sim}{u}) \epsilon \mathcal{E}_N$ spanned by the n vectors $\underset{\sim}{\sigma}_i(\underset{\sim}{u}) \equiv \underset{\sim}{e}_Q E_{Q,i}(\underset{\sim}{u})$ when rank $\|E_{A,i}(\underset{\sim}{u})\| = n$. Since the flat Σ_n approximates the curved S_n in the neighborhood of the point of tangency $\underset{\sim}{E}(\underset{\sim}{u})$, the direction $\underset{\sim}{z} \epsilon A_n$ of steepest descent of $|\underset{\sim}{E}(\underset{\sim}{v})|^2$ as $\underset{\sim}{E}(\underset{\sim}{v})$ sweeps out the curved S_n will be identical with that of steepest descent of $|\underset{\sim}{\Sigma}|^2$ as $\underset{\sim}{\Sigma}$ sweeps out the tangent n-plane $\Sigma_n(\underset{\sim}{u})$. The displacement vector $\underset{\sim}{\Sigma}$ from O to any point of the tangent n-plane $\Sigma_n(\underset{\sim}{u})$ is of the form $\underset{\sim}{\Sigma} \equiv \underset{\sim}{\sigma}_r(\underset{\sim}{u})z^r + \underset{\sim}{E}(\underset{\sim}{u})$.

Now the desired direction in $\Sigma_n(\utilde{u})$ of steepest descent of $|\utilde{\Sigma}|^2$ as $|\utilde{\Sigma}|$ ranges over $\Sigma_n(\utilde{u})$ is given by the <u>direction</u> of the tangential vector $\utilde{\sigma}_r(\utilde{u})z^r$ when $\underset{\wedge\wedge}{z} \equiv (z^1, \cdots, z^n)\epsilon A_n$ is chosen such that $\utilde{\sigma}_r(\utilde{u})z^r$ gives the displacement in $\Sigma_n(\utilde{u})$ from the contact point $\utilde{E}(\utilde{u})$ to the foot of the perpendicular from the origin 0 onto $\Sigma_n(\utilde{u})$. The conditions on z^i giving the equations of steepest descent are thus identical with (23.1),

$$\utilde{\sigma}_i \cdot (\utilde{\sigma}_r z^r + \utilde{E}) \equiv E_{Q,i}E_{Q,r}z^r + E_Q E_{Q,i} = 0. \tag{23.2}$$

2

24. SINGULAR POINTS OF ERROR MAP

The debugging of our IBM Stretch lens design code was undertaken from an initial trial Lister type lens $\underset{\sim}{u}$ as shown in Fig. 2.1. This had been previously obtained by Brixner as a completed design using the Holladay program as revised and written by C. A. Lehman[10] for the IBM 7090. It may therefore be assumed that our trial $\underset{\sim}{u}$ was near an optimum. We constrained the system by a certain choice of the parameter flag vector $\underset{\sim}{f}$ of section 2 which left us with 12 design parameters out of a possible 23. A nonsingular linear system solver showed that the elements of the coefficient matrix $\|g_{ij}\|$ of the 12x12 linear system (23.1) ranged in absolute value from order 10^{-9} to 10^{3}. The exponent sequence of the Gaussian pivots was $\{-2, -4, -2, -2, -5, -5, -4, -7, -9, -12, -17, -18\}$ and the determinant was of order 10^{-77}! If we define the approximate machine zero for this matrix $\|g_{ij}\|$ to be of order 10^{-12}, obtained as the product of the maximum element of order 10^{3} by the machine accuracy 10^{-15}, we see that the system is of rank 10 at most and *should not be processed* as a nonsingular system.

A trial lens $\underset{\sim}{u} \in A_n$ is a *singular* point of the error map $\underset{\sim}{E}(\underset{\sim}{u})$ when the columns $i = 1, \cdots, n$ of the derived matrix $\|E_{A,i}(\underset{\sim}{u})\|$ are linearly dependent, meaning that there exists a set of constants $\underset{\sim}{w} \equiv (w_1, \cdots, w_n)$ not all 0, such that $E_{A,r}(\underset{\sim}{u})w^r = 0$ for $A = 1, \cdots, N$. Since $g_{ij} \equiv E_{Q,i}E_{Q,j}$ from Eq. (22.1), we have $g_{ir}w_r = 0$ and hence $|g_{ij}| = 0$. Conversely, let $|g_{ij}| = 0$ and let $\underset{\sim}{w}$ now be *any* solution of $g_{ir}w_r = 0$. Then $g_{rs}w_rw_s = V_QV_Q = 0$, where $V_A \equiv E_{A,r}w^r$, and so $V_A = 0$. But this means that any $\underset{\sim}{w}$ satisfying $g_{ir}w_r = 0$ also satisfies $E_{A,r}w_r = 0$. Thus rank $\|g_{ij}\|$ = rank $\|E_{A,j}\|$.

To test the singular system (23.1) for consistency we choose $\underset{\sim}{w}$ to be a solution of $w_s g_{sj} = 0$ and form $w_s g_{sr} z^r = - E_Q E_{Q,s} w_s$. The coefficients of the z's are now zero and the necessary and sufficient condition for consistency of the singular system is that $E_Q E_{Q,s} w_s$ should likewise vanish. But this condition is satisfied since we know that $E_{A,s} w_s = 0$ for $A = 1, \cdots, N$.

Consider now a singular point $\underset{\sim}{u} \epsilon A_n$ for which rank $\|E_{A,j}(\underset{\sim}{u})\| = r < n$. Then rank $\|g_{ij}(\underset{\sim}{u})\| = r$ and the homogeneous system $g_{ir} w^r = 0$ admits n-r linearly independent solutions $\underset{\sim}{w}_\alpha$. Let $\underset{\sim}{\zeta}$ be some arbitrarily chosen particular solution of the consistent singular system $g_{ir} z^r + E_Q E_{Q,i} = 0$, then the most general particular solution is of the form $\underset{\sim}{z} \equiv \underset{\sim}{\zeta} + c_\rho \underset{\sim}{w}_\rho$ and

$$|\underset{\sim}{z}|^2 \equiv g_{rs} z^r z^s \equiv g_{rs}(\zeta^r + c_\rho w_\rho{}^r)(\zeta^s + c_\sigma w_\sigma{}^s) = g_{rs}\zeta^r \zeta^s = |\underset{\sim}{\zeta}|^2. \qquad (24.1)$$

From this we conclude that the steepest descent directional derivative $- 2|\underset{\sim}{z}|$ of Eq. (20.3) is independent of the choice of the particular solution $\underset{\sim}{z}$ of the singular consistent system (23.1) resulting from a singular trial point $\underset{\sim}{u}$.

We follow R. E. von Holdt[3] by selecting from the ∞^{n-r} particular solutions $\underset{\sim}{z} \equiv \underset{\sim}{\zeta} + c_\rho \underset{\sim}{w}_\rho$ that uniquely determined solution for which $\|\underset{\sim}{z}\|^2 \equiv z^r z^r$ is a minimum. To arrive at this selection we first orthonormalize the n-r $\underset{\sim}{w}$'s such that $\underset{\sim}{w}_\alpha \cdot \underset{\sim}{w}_\beta \equiv w_\alpha{}^r w_\beta{}^r = \delta_{\alpha\beta}$ and then choose the c's so that the particular solution $\underset{\sim}{z}$ is orthogonal to these orthonormalized $\underset{\sim}{w}$'s, the conditions being

$$\underset{\sim}{w}_\alpha \cdot \underset{\sim}{z} \equiv \underset{\sim}{w}_\alpha \cdot \underset{\sim}{w}_\rho c_\rho + \underset{\sim}{w}_\alpha \cdot \underset{\sim}{\zeta} = c_\alpha + \underset{\sim}{w}_\alpha \cdot \underset{\sim}{\zeta} = 0. \qquad (24.2)$$

We designate by $\underset{\sim}{z}_m$ the particular solution resulting from $\underset{\sim}{z} \equiv \underset{\sim}{\zeta} + c_\rho \underset{\sim}{w}_\rho$ when the c's have been chosen by Eq. (24.2). Then any other particular solution is of the form $\mathbf{z} \equiv \mathbf{z}_m + \tau_\rho \mathbf{w}_\rho$ and

$$||\mathbf{z}||^2 \equiv (\mathbf{z}_m + \tau_\rho \mathbf{w}_\rho)\cdot(\mathbf{z}_m + \tau_\sigma \mathbf{w}_\sigma \equiv ||\mathbf{z}_m||^2 + \tau_\rho \tau_\rho > ||\mathbf{z}_m||^2 \qquad (24.3)$$

since now $(\mathbf{z}_m \cdot \mathbf{w}_\alpha) = 0$.

Conclusion: An efficient subroutine for solving the steepest descent system (23.1) should process the system as nonsingular when no Gaussian pivot drops in absolute value below the approximate machine zero for the matrix $||g_{ij}||$. When the trial lens $\mathbf{u}$ is a singular point, the rank r of $g_{ij}(\mathbf{u})$ should be determined and an orthonormal set of n-r solutions $\mathbf{w}_\alpha$ of the homogeneous system $g_{ir}w^r = 0$ should be computed. From these, and any arbitrarily chosen particular solution $\boldsymbol{\zeta}$, the unique particular solution $\mathbf{z}_m$ which minimizes the norm $||\mathbf{z}||$, $\mathbf{z}$ ranging over the ∞^{n-r} particular solutions, is determined.

He who has coded R. E. von Holdt's excellent eigenvalue-eigenvector subroutine mentioned in section 1 has but to reach in his card file for such a linear system solver; others face a job of work. The futility of attempting to use a nonsingular linear system solver at a singular trial $\mathbf{u}$ has driven malpractitioners to the artful dodge of varying only a few parameters at a time.

25. GRADIENT MAP OF A_n INTO $\mathcal{E}_n$

Let $\varphi(\underset{\sim}{u})$, $\underset{\sim}{u}\epsilon A_n$, be any scalar function φ which we seek to minimize over A_n. We shall formulate this minimization problem by the following chain of procedures:

1) Map A_n into $\mathcal{E}_n$ by the primitive mapping $[n,\delta_{ij},u^i]$ of (21.1) so that the arithmetic function φ defined over A_n now becomes the geometric point function $\varphi[\underset{\sim}{V}]$, $\underset{\sim}{V} \equiv \underset{\sim}{e}_r u^r \epsilon \mathcal{E}_n$, defined over $\mathcal{E}_n$.

2) Form the gradient error vector $\underset{\sim}{E}_g \equiv \underset{\sim}{e}_r \varphi_{,r}$, $\varphi_{,i} \equiv \partial\varphi/\partial u^i$, and map Euclidean $\mathcal{E}_n$ onto itself by the choice of mapping arguments $[n,\delta_{ij},\varphi_{,i}]$. The steepest descent equations (20.1) for minimizing any scalar point function $\psi[\underset{\sim}{V}]$, $\underset{\sim}{V} \equiv \underset{\sim}{e}_r u^r \epsilon \mathcal{E}_n$, are written in terms of the gradient mapping metric (19.2),

$$g_{ir}z^r + \tfrac{1}{2}\psi_{,i} = 0, \quad g_{ij} \equiv \varphi_{,si}\varphi_{,sj}, \quad \varphi_{,ij} \equiv \partial^2\varphi/\partial u^i \partial u^j. \tag{25.1}$$

3) Now define the scalar ψ to be intrinsically related to the gradient mapping vector $\underset{\sim}{E}_g$ by the definition $\psi \equiv |\underset{\sim}{E}_g|^2 \equiv \varphi_{,r}\varphi_{,r}$. Then (25.1) reduces to

$$\begin{gathered}\varphi_{,si}(\varphi_{,sr}z^r + \varphi_{,s}) = 0 \text{ when } |\varphi_{,ij}(\underset{\sim}{u})| = 0 \\ \therefore \varphi_{,ir}z^r + \varphi_{,i} = 0 \text{ when } |\varphi_{,ij}(\underset{\sim}{u})| \neq 0.\end{gathered} \tag{25.2}$$

If we now restrict the general φ of (25.2) by making the choice $\varphi \equiv |\underset{\sim}{E}|^2$, where $\underset{\sim}{E}$ is the optical error vector, then the nonsingular point ($|\varphi_{,ij}(\underset{\sim}{u})| \neq 0$) form of (25.2) gives the nonsingular point second order equations of steepest descent for the scalar $\psi \equiv \varphi_{,r}\varphi_{,r}$ **relative** to the *gradient* *mapping* $[n,\delta_{ij},\varphi_{,i}]$,

$$(E_{Q,i}E_{Q,r} + E_Q E_{Q,ir})z^r + E_Q E_{Q,i} = 0. \tag{25.3}$$

At a singular point $\mathbf{u}$ the system (25.3) must be replaced by the first of (25.2).

The preparatory primitive mapping in 1) is motivated by the awareness that $\varphi_{,ij}(\mathbf{u})$ are components of a covariant tensor only when the u's are *rectangular* *Cartesian* *coordinates*.

26. STATIONARY POINT CLASSIFICATION

We assume now that our trial point of A_n is a stationary point $\underset{\sim}{u}_s$ of the scalar $\varphi \equiv |\underset{\sim}{E}|^2$, where $\underset{\sim}{E}$ is the optical error vector. We seek to determine whether $\underset{\sim}{u}_s$ be a maximum, a minimum, or a saddle point of $|\underset{\sim}{E}|^2$ by the following procedure.

1) Consider any neighboring point pair $(\underset{\sim}{u},\underset{\sim}{v})\epsilon A_n$ and define $\underset{\sim}{z} \equiv \underset{\sim}{v}-\underset{\sim}{u}$.

2) Make the primitive mapping $[n,\delta_{ij},u^i]$ of A_n into $\mathcal{E}_n$ so that $\varphi \equiv |\underset{\sim}{E}|^2$, an arithmetic function defined over A_n, now becomes a geometric point function φ defined in rectangular Cartesian coordinates u^i over $\mathcal{E}_n$.

3) Expand $\varphi \equiv |\underset{\sim}{E}|^2$ in a Taylor series about $\underset{\sim}{u}$,

$$\varphi(\underset{\sim}{u}+\underset{\sim}{z}) = \varphi(\underset{\sim}{u}) + \varphi_{,r}(\underset{\sim}{u})z^r + \tfrac{1}{2}\varphi_{,rt}(\underset{\sim}{u})z^r z^t + \cdots, \tag{26.1}$$

where the dots indicate terms of higher order in the z's.

4) Now restrict $\underset{\sim}{u}$ to be a <u>stationary</u> point $\underset{\sim}{u}_s$ of φ so that $\varphi_{,i}(\underset{\sim}{u}_s) = 0$. By choosing all $|z^i|$ sufficiently small the truncated Taylor series φ_T,

$$\varphi_T(\underset{\sim}{u}_s+\underset{\sim}{z}) \equiv \varphi(\underset{\sim}{u}_s) + \tfrac{1}{2}\varphi_{,rt}(\underset{\sim}{u}_s)z^r z^t, \tag{26.2}$$

approximates $\varphi(\underset{\sim}{u}_s+\underset{\sim}{z})$ to within any desired accuracy. Inspection of (26.2) gives the following classifications of a stationary point $\underset{\sim}{u}_s \epsilon A_n$:

$\underset{\sim}{u}_s$ is a <u>minimum</u>, <u>maximum</u>, or <u>saddle point</u> of $\varphi(\underset{\sim}{u})$ when $\varphi_2(\underset{\sim}{z}) \equiv \varphi_{,rs}(\underset{\sim}{u}_s)z^r z^s$ is <u>positive definite</u>, <u>negative definite</u>, or <u>indefinite</u> respectively.

27. THE CAUCHY LINE L_C

We return to the system (23.1) for the direction of steepest descent of $\varphi \equiv |\underset{\sim}{E}|^2$ relative to the orthonormal error mapping $[N,\delta_{ij},E_A(\underset{\sim}{u})]$ of section 22. Cauchy[9] showed how one could depart from a trial $\underset{\sim}{u}\epsilon A_n$ to arrive at a $\underset{\sim}{v}\epsilon A_n$ such that $\varphi(\underset{\sim}{v}) < \varphi(\underset{\sim}{u})$ by considering $\varphi(\underset{\sim}{u}+\eta\underset{\sim}{z})$, η increasing continuously from 0. We shall call this *arithmetic half line* $\underset{\sim}{v} \equiv \underset{\sim}{u}+\eta\underset{\sim}{z}\epsilon A_n$, departing from our trial $\underset{\sim}{u}$ in the direction of steepest descent, the *Cauchy line* L_C.

We begin by considering the *geometric half line* $L_\Sigma\epsilon\Sigma_n(\underset{\sim}{u})$,

$$L_\Sigma\colon\ \underset{\sim}{\Sigma}(\eta) \equiv \underset{\sim}{E}(\underset{\sim}{u}) + \eta\underset{\sim}{\sigma}_r(\underset{\sim}{u})z^r,\ \underset{\sim}{z}\text{ satisfying (23.1)}, \tag{27.1}$$

which departs from the contact point $\underset{\sim}{E}(\underset{\sim}{u})$ (given by $\eta = 0$) of the tangent n-plane $\Sigma_n(\underset{\sim}{u})$ to S_n and is directed toward the foot $\underset{\sim}{\Sigma} \equiv \underset{\sim}{E}(\underset{\sim}{u}) + \underset{\sim}{\sigma}_r(\underset{\sim}{u})z^r$ (given by $\eta = 1$) of the perpendicular from 0 onto $\Sigma_n(\underset{\sim}{u})$. As η increases from 0 the displacement vector $\underset{\sim}{\Sigma}(\eta)$ of squared magnitude

$$|\underset{\sim}{\Sigma}(\eta)|^2 \equiv |\underset{\sim}{E} + \eta\underset{\sim}{\sigma}_{,r}z^r|^2 \equiv |\underset{\sim}{E}|^2 + 2\eta\underset{\sim}{E}\cdot\underset{\sim}{\sigma}_r z^r + \eta^2 g_{rs}z^r z^s \tag{27.2}$$

sweeps out L_Σ. The minimum of $|\underset{\sim}{\Sigma}(\eta)|^2$ occurs for η satisfying

$$\eta g_{rs}z^r z^s = -E_Q E_{Q,r}z^r = g_{rs}z^r z^s \text{ from (23.1)}$$

so that $\eta = 1$ minimizes $|\underset{\sim}{\Sigma}(\eta)|^2$ along $L_\Sigma\epsilon\Sigma_n(\underset{\sim}{u})$. This exercise in formalism merely rediscovers what we already know, namely that $\underset{\sim}{\Sigma}(1) \equiv \underset{\sim}{E} + \underset{\sim}{\sigma}_r z^r$, $\underset{\sim}{z}$ satisfying (23.1), gives the least displacement from 0 to points on $\Sigma_n(\underset{\sim}{u})$.

Of much greater concern to us is the behavior of

$$\varphi[\eta] \equiv |\mathbf{E}[\eta]|^2 \equiv |\mathbf{E}(\mathbf{u}+\eta\mathbf{z})|^2, \quad \mathbf{z} \text{ a solution of }(23.1), \tag{27.3}$$

as the Euclidean displacement vector $\mathbf{E}(\mathbf{u}+\eta\mathbf{z})$ traces a curve $C_\eta \epsilon S_n$ while $\mathbf{v} \equiv \mathbf{u}+\eta\mathbf{z}$ moves along $L_C \epsilon A_n$. Expanding (27.3) in a Taylor series and substituting (27.2) gives

$$\varphi[\eta] = |\mathbf{\Sigma}(\eta)|^2 + \mathbf{E}(\mathbf{u})\cdot\mathbf{E}_{,rs}(\mathbf{u})z^r z^s \eta^2 + \cdots . \tag{27.4}$$

If the trial $\mathbf{u}\epsilon A_n$ is *near* a stationary point $\mathbf{u}_s$ of φ so that $|\varphi_{,i}(\mathbf{u})| \approx 0$, then the solutions z^i of Eqs. (23.1) satisfy $|z^i| \approx 0$ regardless of whether $|g_{ij}(\mathbf{u})|$ be singular or nonsingular, in the latter case because we then choose the particular solution minimizing $\|\mathbf{z}\|^2 \equiv z^r z^r$. We conclude that for $\mathbf{u}$ near $\mathbf{u}_s$ the truncation of the expansion (27.4) resulting from neglecting terms of order higher than 2 in the z's yields a good approximation to $\varphi[\eta]$ for $\eta \le 1$.

For the trial $\mathbf{u}$ well removed from a stationary point $\mathbf{u}_s$ of φ, and this is the situation in the early stages of a lens design, truncation of the terms of order 3 and higher in the expansion (27.4) cannot be justified. Because of this and because in any case evaluating the terms $\mathbf{E}(\mathbf{u})\cdot\mathbf{E}_{,ij}(\mathbf{u})$ would require formidable analysis and computation for a lens problem, we *renounce any attempt at approximating $\varphi[\eta]$ at the points $\mathbf{u}+\eta\mathbf{z}$ of the Cauchy line L_C in favor of evaluating $\varphi[\eta]$ to within machine accuracy by ray-tracing*.

28. MINIMIZING Φ ALONG L_C

With a trial lens $\underset{\sim}{u}$ in the machine, we follow Cauchy[9] by narrowing our search for the next trial lens to an inspection of the ∞^1 lenses $\underset{\sim}{v}(\eta)$ on the Cauchy line L_C: $\underset{\sim}{v}(\eta) \equiv \underset{\sim}{u} + \eta\underset{\sim}{z}\epsilon A_n$ passing through $\underset{\sim}{u}$ in the direction of steepest descent for φ. Thus, along C,

$$\varphi[\eta] \equiv \varphi(\underset{\sim}{u}+\eta\underset{\sim}{z}),$$

$$d\varphi/d\eta|_0 \equiv \varphi_{,r}(\underset{\sim}{u})z^r = -2\,g_{rs}z^r z^s < 0 \text{ from } (20.2). \tag{28.1}$$

Figure 28.1 shows a graph of $\varphi[\eta]$. Its negative slope at $\eta = 0$, as evidenced by (28.1), ensures the possibility of finding a point $\underset{\sim}{v}(\eta)\epsilon L_C$ for which $\varphi(\underset{\sim}{v}) \equiv \varphi[\eta] < \varphi[0] \equiv \varphi(\underset{\sim}{u})$. Any reasonable approximation to a local minimum of $\varphi[\eta]$ on L_C will be satisfactory. We must be aware that machine time spent in refining our approximation to a local minimum on L_C might possibly be spent more effectively by getting started on the next least squares pass proceeding from the acceptance of a less refined local minimum on L_C. We present what seems to us an adequate exploration of L_C.

insert Fig. 28.1 approximately here

Prior to beginning our march along L_C as in Fig. 28.1 we must refer back to our introduction near Eq. (10.6) of a data input oriented lower bound vector $\underset{\sim}{m}$ which imposes the constraints $u^i >(<)m^i$ when $u^i>(<)0, i=-2,\cdots, n_{surf}$. These constraints are saying that ρ_o and the various axial spacings must be bounded away from a possible reversal of sign. Consideration of these yields an upper bound η_{max} with the property that for any η satisfying $\eta \leq \eta_{\max}$ the Cauchy replacement new $\underset{\sim}{u} \equiv$ old $\underset{\sim}{u} + \eta\underset{\sim}{z}$ will avoid any such violation of $\underset{\sim}{m}$. Any attempt to minimize $\varphi[\eta]$ for $\eta > \eta_{\max}$ signals a) the end of our march along L_C, b) the acceptance as optimal η along L_C of the last ray-trace value of η preceding the violation $\eta > \eta_{\max}$, and c) the freezing of the distance parameter, or parameters, which caused the setting of $\eta_{\max}$ from further design variations over the next n_{freeze} least squares passes, where n_{freeze} is a data input control integer.

Now choose a first Cauchy line probing point $\eta = \Delta\eta$ which on the *first* least squares pass we select as $\Delta\eta \equiv \min(\frac{1}{2},\eta_{\max})$, this preference for $\frac{1}{2}$ resulting purely from ignorance. Evaluate $\varphi[\Delta\eta]$ by a ray-trace and compare $\varphi[0]$ with $\varphi[\Delta\eta]$. Either we have case II, $\varphi[\Delta\eta] \geq \varphi[0]$, or

Case I, $\varphi[\Delta\eta] < \varphi[0]$: Here we proceed as in Fig. 28.1. We seek to extend the monotone decreasing sequence $\{\varphi[0],\varphi[\Delta\eta]\}$ by successive ray-trace evaluations of φ at the abscissa sequence $\{\eta_i\}$, where $\eta_0 \equiv 0$ and $\eta_i \equiv \eta_{i-1} + 2^{i-1}\Delta\eta$, $i = 1,2,\cdots$, to obtain the corresponding ordinate sequence $\{\varphi[\eta_i]\}$. An interruption in the monotone decrease of $\{\varphi[\eta_i]\}$, such as occurs at $i = 5$ in Fig. 28.1, signals a ray-trace evaluation of φ at the *midpoint of the last interval* (η_{i-1},η_i). We are now left with 4 *equally spaced* terminating points (P_1,P_2,P_3,P_4) with abscissae $(\tilde{\eta}_1,\tilde{\eta}_2,\tilde{\eta}_3,\tilde{\eta}_4)$ and φ values $(\tilde{\varphi}_1,\tilde{\varphi}_2,\tilde{\varphi}_3,\tilde{\varphi}_4)$. We compare the end values $\tilde{\varphi}_1$ and $\tilde{\varphi}_4$ and reject the larger. This reduces our terminating quartet to a triplet which, after a possible renaming, we call $P_i(\tilde{\eta}_i,\tilde{\varphi}_i)$, $i = 1,2,3$. We prepare now to pass a parabola with a vertical axis through this triplet by testing the point triplet discriminant $D \equiv \tilde{\varphi}_1 - 2\tilde{\varphi}_2 + \tilde{\varphi}_3$. When $D = (<)0$ we by-pass the interpolating parabola since now the points (P_1,P_2,P_3) are collinear (lie on a parabola with vertex concave downward). For D > 0 we form $\Delta\tilde{\eta}$, the vertex abscissa relative to the middle abscissa $\tilde{\eta}_2$,

$$\Delta\tilde{\eta} \equiv \tfrac{1}{2}(N/D)(\tilde{\eta}_3-\tilde{\eta}_2), \quad N \equiv \tilde{\varphi}_1-\tilde{\varphi}_3, \tag{28.2}$$

and reject the interpolation when $|\Delta\tilde{\eta}| \geq \tilde{\eta}_3 - \tilde{\eta}_2$. Otherwise we evaluate $\varphi[\tilde{\eta}_2+\Delta\tilde{\eta}]$ by ray-tracing and choose the minimum of $\tilde{\varphi}_2, \tilde{\varphi}_3$, and $\varphi[\tilde{\eta}_2+\Delta\tilde{\eta}]$ whose abscissa will serve as the optimal η along L_C.

Case II, $\varphi[\Delta\eta] \geq \varphi[0]$: Our first trial probe at $\eta_0 \equiv \Delta\eta$ is now too far to the right and so we evaluate φ at the abscissa sequence $\{\eta_i\}, \eta_i = \eta_{i-1} - 2^{-i}\Delta\eta$, $i = 1,2,\cdots$, to obtain the ordinate sequence $\{\varphi[\eta_i]\}$. We ignore $\varphi[\eta_i]$ until we reach an i for which $\varphi[\eta_i] < \varphi[0]$, which in Fig. 28.2 occurs at i = 2. This signals the beginning of a monotone decreasing

Insert Fig. 28.2 approximately here

sequence $\{\varphi\}$ which is first interrupted in Fig. 28.2 at i = 4. This interruption at η_i is the signal to go back to the right and make a ray-trace at the midpoint of the interval (η_{i-1}, η_{i-2}) to establish the intermediate P_3 in Fig. 28.2. We are now left with a terminating quartet of equally spaced station points (P_1, P_2, P_3, P_4) and we proceed to approximate the optimal η as in case I.

The blind choice $\eta = 1$ as optimal along L_C is the hallmark of the least squares malpractitioner. We present a case history to show how disastrous this choice may be. C. A. Lehman coded at our request a curve fitting program which determines the circle $(x-u^1)^2 + (y-u^2)^2 - (u^3)^2 = 0$ best fitting a set of data points (x_A, y_A), $A = 1,\cdots,N$, which appear to lie on a circle. A test run was made on a set of 5 points chosen to lie on the circle $\underset{\sim}{u} \equiv (40,20,100)$. We chose $\underset{\sim}{u} \equiv (1,1,1)$ as the first trial $\underset{\sim}{u}$ which gave $\varphi = 3.8(10^8) \equiv 3.8+08$ and $\underset{\sim}{z} \equiv (39,19,4058.5)$. Thus the malpractitioner's choice $\eta = 1$ would yield as the next trial circle $\underset{\sim}{u} \equiv (40,20,4059.5)$ with $\varphi = 3.4+14$! Here we have case II of Fig. 28.2 and the best station value of η was $\eta = 0.03125$, yielding the Cauchy optimal trial circle (2.22,1.59,128) for which $\varphi = 1.7+07$. The monotone decreasing sequence $\{\varphi\}$ promised in our introduction was $\{1.7+07, 6.8+05, 1.3+00, 4.7-08\}$ corresponding to the Cauchy optimal η's $\{.03125, .939, 1.004, .99996\}$ and the u's at convergence showed 8 digit accuracy.

We used this same problem to test a curve fitting program written by the statisticians at Los Alamos who adhere to the popular choice $\eta = 1$. This yielded the nonmonotone sequence $\{\varphi\} \equiv \{1.5+09, 1.4+15, 8.5+13, 5.3+12, 3.2+11, 1.9+10, 8.6+08, 1.7+07, 2.7+04, 8.8-02, 2.9-07\}$ and the solution was found to 8 digit accuracy. It is perhaps the occasional "success" of his $\eta = 1$ program which makes the malpractitioner's conversion difficult.

Some malpractitioners concede that $\eta = 1$ may not be optimal at the early stages of convergence, but they are convinced that optimal $\eta \to 1$ as the trial $u \to u_s$, a stationary point of φ. This conviction may be challenged by the example

$$E \equiv (u,v,h^2+u^2+v^2), \quad z \equiv (-\psi u,-\psi v),$$

$$\psi \equiv [1+2(h^2+u^2+v^2)]/[1+4(u^2+v^2)].$$

Clearly $\eta = 1/\psi$ is optimal along L_C and yields the minimizing point

$u_m \equiv (0,0)$ at the end of the first least squares pass. As the trial $u \rightarrow (0,0)$ the optimal $\eta \rightarrow 1/(1+2h^2) \neq 1$ for $h \neq 0$. One may be tempted to extrapolate from this example and conjecture that when minimum $\varphi \rightarrow 0$ optimal $\eta \rightarrow 1$.

If it should be found that in the majority of cases optimal η tends to 1 as $u \rightarrow u_s$, then the choice $\Delta\eta \equiv \frac{1}{2}$(previous optimal η) would be a favorable selection for the first probe of L_C after the first least squares pass. Near convergence this would cause the intermediate point P_3 of Fig. 28.1 to be established by a ray-trace near the expected optimal η.

29. HOLLADAY'S OPTICS CODE

It may interest current users of the Holladay lens design code to learn what is in it. The code employs what some might call numerical differentiation. This is initiated by the input data entry of a parameter increment vector $\Delta\mathbf{u}$ and the program forms by ray-tracing

$$\mathbf{I}_i \equiv \mathbf{E}(\mathbf{u}+\Delta u^i) - \mathbf{E}(\mathbf{u}), \quad \mathbf{u}+\Delta u^i \equiv (u^1,\cdots,u^i+\Delta u^i,\cdots,u^n). \tag{29.1}$$

Definition. We shall say that $\mathbf{I}_i/\Delta u^i$ *is approximating the partial derivative* $\mathbf{E}_{,i}$ *in the Holladay sense*.

If we write the steepest descent system (23.1) in the vector form $\mathbf{E}_{,i}\cdot\mathbf{E}_{,r}z^r+\mathbf{E}\cdot\mathbf{E}_{,i} = 0$, approximate $\mathbf{E}_{,i}$ in the Holladay sense, and multiply by Δu^i, with no summation on i, we obtain Holladay's approximation to (23.1)

$$\mathbf{I}_i\cdot\mathbf{I}_r\lambda^r + \mathbf{E}\cdot\mathbf{I}_i = 0, \; z^i \equiv \lambda^i\Delta u^i, \qquad \text{(i not summed)}. \tag{29.2}$$

The accuracy of the approximate system (29.2) is measured by the accuracy of the approximations $\mathbf{I}_i/\Delta u^i$ to $\mathbf{E}_{,i}$.

We define $\|\Delta\mathbf{u}\| \equiv (\Delta u^r\Delta u^r)^{\frac{1}{2}}$ and observe that, with decreasing $\|\Delta\mathbf{u}\|$, $\mathbf{I}_i\cdot\mathbf{I}_j \to 0$ as $\|\Delta\mathbf{u}\|^2$ and $|\mathbf{I}_i\cdot\mathbf{I}_j| \to 0$ as $\|\Delta\mathbf{u}\|^{2n}$. When not near a stationary point $\mathbf{u}_s$, we have $\|\mathbf{z}\| \not\to 0$, and hence $1/\lambda^i \to 0$ as Δu^i. The ill-chosen scaling resulting from Holladay's differential approximation (29.2) to the correct system (23.1) imposes severe demands upon a computer with only a finite number of bits. The Holladay program attempts to solve (29.2) with a nonsingular linear system solver, a highly unjustified commitment which readily leads to trouble.

The self-imposed near singularity of his system (29.2) compelled Holladay to abandon any attempt to vary all design parameters at once in favor of the expedient of varying only a few parameters at a time. Since test runs confirmed our analysis following Eq. (29.2) of the extreme sensitivity of the coefficient matrix $\|\underset{\sim}{I}_i \cdot \underset{\sim}{I}_j\|$ of (29.2) to the smallness of $\|\Delta \underset{\sim}{u}\|$, Holladay sought to mollify the robot's protest by rescaling his $\Delta \underset{\sim}{u}$ at the end of each least squares pass in preparation for the next according to the storage substitution

$$\text{new } \Delta u^i \equiv [e_i{}^2/(\underset{\sim}{I}_i \cdot \underset{\sim}{I}_i)^{\frac{1}{2}}] \ (\text{old } \Delta u^i), \ (\text{i not summed}),$$

where $e_i{}^2, i=1,\cdots,n$, are data input control constants. What this perturbation in $\Delta \underset{\sim}{u}$ does simultaneously to the coefficient matrix $\|\underset{\sim}{I}_i \cdot \underset{\sim}{I}_j\|$ and to the Holladay approximation to $\underset{\sim}{E}_{,i}$ we leave to the reader's speculation.

Holladay makes the usual malpractice choice $\eta \equiv 1$, namely new $u^i \equiv$ old $u^i + \lambda^i \Delta u^i$, i not summed. His only concession to Cauchy's 1847 recommendation in this matter is to compare $\lambda_{max} \equiv \max|\lambda^i|, i=1,\cdots, n$, with a positive input data control constant Λ, and when $\lambda_{max} > \Lambda$ he defines

$$\text{new } u^i \equiv \text{old } u^i + (\Lambda/\lambda_{max})\lambda^i \Delta u^i, \quad (\text{i not summed}).$$

The Holladay code does not generate a monotone decreasing sequence $\{\varphi\}$ and it contains no convergence exit. To obtain compliance with the designer's data input f-number demand it does not form the error component (10.4) as we have done but rather perturbs the new lens obtained by the least squares process to force compliance with the required f-number. Obviously any such perturbation may be sufficient to upset the monotonicity of the sequence $\{\varphi\}$. The sample rays of Fig. 4.1 are not weighted according to the illumination they represent nor is there any attempt to automate the choice of weights corresponding to the error components. There is no formation of vignetting error components as presented in section 9 to steer the next least squares pass so as to remove the current vignetting.

It is easy to see how Holladay's expedient of varying only a few parameters at a time can lead to a successful result. The directional derivative (20.3) of φ is *negative* regardless of the number n of independent parameters that are allowed to vary at any least squares pass. Thus varying only a few parameters at a time will still generate a monotone decreasing sequence $\{\varphi\}$ *if* one proceeds correctly by using the method of this paper.

Since we have been drawn into presenting a machine approximation of $\underset{\sim}{E}_{,i}$ in the Holladay sense, we conclude this section by reporting our own experience with the machine approximation of $\underset{\sim}{E}_{,i}$ in the mathematical sense. To obtain this we consider the α sequence $\{Q_{Ai\alpha}\}$, A and i being held *fixed*, where

$$Q_{Ai\alpha} \equiv [E_A(\underset{\sim}{u}+\Delta u^i_\alpha) - E_A(\underset{\sim}{u})]/\Delta u^i_\alpha, \quad \Delta u^i_\alpha \equiv 10^{-\alpha}\Delta u^i_{start} \quad \alpha = 1,2,\cdots, \tag{29.3}$$

with $\Delta u^i_{start} \equiv 1$ a likely data input choice to initiate the *first* least squares pass. Next we make the following assumptions:

1) $E_{A,i}(\underset{\sim}{u})$ is best approximated by that α which signals the interruption in the monotone trend of the difference sequence $\{|Q_{Ai\alpha} - Q_{Ai\alpha+1}|\}$.

2) With optimal α determined by 1), this *same* α yields the best practically obtainable machine approximation to $E_{A,i}(\underset{\sim}{u})$ for *any* A.

3) To simplify the determination of the optimal α in 1) we may replace E_A in (29.3) by the y coordinate of the plate spot resulting from tracing a ray of arbitrarily chosen color β from the highest object point Q_T aimed through the lowest entrance pupil mesh point 1 of Fig. 4.1.

Under these assumptions we define $y'_\alpha \equiv [y(\underset{\sim}{u}+\Delta u^i_\alpha)-y(\underset{\sim}{u})]/\Delta u^i_\alpha$ and tune Δu^i_α for numerical differentiation by choosing $\Delta u^i_\alpha \equiv 10^{-\alpha}\Delta u^i_{start}$, where α signals the interruption in the monotone trend of the difference sequence $\{|y'_{\alpha+1} - y'_\alpha|\}$. For the Lister type lens of Fig. 2.1, with the trial $\underset{\sim}{u}$ chosen as the near optimum from a previous Holladay code run, and for i = 7 giving $u^7 \equiv B_1 \equiv 0.0459$, the vertex curvature of Σ_1, the tuning sequence $\{y'_\alpha\}$ gave $\{23.7, 21.74, 21.641, 21.6332, 21.6324, 21.632313, 21.6323040, 21.6323041, 21.63233, \cdots\}$ where the dots represent difference quotients diverging from the best obtainable approximation 21.6323041. Tests confirm that a 15 digit machine can approximate an optical derivative for the Lister lens of Fig. 2.1 accurately to about 8 digits. The derivatives may be retuned at the beginning of the next least squares pass by starting with a coarsened increment vector Δu^i_{start} where $\Delta u^i_{start} \equiv 10^2\Delta u^i_{old}$ would be a likely choice. This permits the tuning sequence to arrive at a coarser increment Δu^i at the next parameter point if it should seek to do so. Tuning may be by-passed as the trial $\underset{\sim}{u} \rightarrow \underset{\sim}{u}_s$, a stationary point of φ. Our Stretch code now contains numerical differentiation and later we shall include the option of analytic differentiation. In this way we may compare numerical and analytic derivatives for agreement and may obtain comparative running times and resolutions.

30. STORAGE REQUIREMENT

We examine the storage necessary for forming the steepest descent system (23.1). If we should choose to form $g_{ij} \equiv E_{Q,i}E_{Q,j}$ by first forming and storing the rectangular $N \times n$ matrix $\|E_{A,i}\|$, inspection of (23.1) shows that such a procedure would demand $(n+1)(n+N)$ cells. Referring to sections 6, 7, 8 we see that N is dominated by the need to store spot coordinates and the number of spots is dominated by the number $n(q_0)$, the number of semicircular entrance pupil mesh points corresponding to the mesh refinement choice of Fig. 4.1. In approximating N we consider the case wherein all rays from a test object point Q_Γ, $\Gamma=1, \cdot\cdot\cdot, n_{object}$ reach the plate to form a spot. We list side by side the spot components of $\underset{\sim}{E}$ and their storage demands:

$x_{E\beta\Gamma}$, $y_{E\beta\Gamma}-p_{\beta\Gamma}$ of (6.2),	$2n(q_0)\cdot n_{color}\cdot n_{object}$,
$p_{\beta\Gamma}-p_\Gamma$ of (6.4),	$n_{color}\cdot n_{object}$,
$R_\Gamma-R_{rmsq}$ of (7.4), $X_\Gamma-Y_\Gamma$ of (8.2),	$2n_{object}$,

and conclude that

$$N \approx n_{object}[n_{color}(2n(q_0)+1)+2]. \qquad (30.1)$$

Only testing can reveal how many object points Q_Γ should be used to approximate a continuous object segment and how fine the entrance pupil mesh of Fig. 4.1 should be chosen. The code provides for starting a new design by converging on $(n_{object}, q_0)\equiv(1, 1)$, then converging on $(2, 1)$, $(3, 1)$, $\cdot\cdot\cdot$ until all desired object points have been brought to focus for $q_0=1$. Then the lens is brought to a sequence of convergences resulting from $(n_{object}, 2)$, $(n_{object}, 3)$, $\cdot\cdot\cdot$ until the last refinement in the entrance pupil mesh shows no discernible improvement over the previous refinement. Testing will reveal whether this gradual procedure will minimize preconvergence vignetting and running time as compared to starting at once with all object points and a fully refined mesh.

We anticipate that $(n_{object}, n_{color}, q_0)\equiv(5, 3, 3)$ will produce a fine Lister type lens, but let us be pessimistic and count the storage requirement for a (10, 6, 10) design of the Lister lens of Fig. 2.1. We have $n(10)=158$. Let us design on all the $n\equiv23$ parameters (2.1) of this lens. Then, from (30.1),

$$N\approx10[6(317)+2]=19,040 \text{ and } (n+1)(n+N)\approx457,512.$$

The Holladay code would demand this number of words, 457,512, for the formation of its system (29.2) since it forms this system by first forming and storing $\|E_A\|$ and the $N\times n$ matrix $\|I_{Ai}\|$, where $\underset{\sim}{I}_i\cdot\underset{\sim}{I}_j\equiv I_{Qi}I_{Qj}$.

The Holladay storage demand is of course preposterous.

Returning to the steepest descent Eqs. (23.1), which we now write compactly as

$$g_{ir}z^{r}+b_{i}=0, \quad g_{ij}\equiv E_{Q,i}E_{Q,j}, \quad b_{i}\equiv E_{Q}E_{Q,i}, \tag{30.2}$$

we observe that the full matrices $\|E_A\|$ and $\|E_{A,i}\|$ are of no interest, for we are concerned only with the contributions of their elements to the formation of the coefficients g_{ij} and b_i of (30.2). To record this contribution we have merely to initial the cells g_{ij} and b_i with 0 and then add in the contribution from any error component E_A and its partial derivatives $E_{A,i}$ just as soon as they have been formed by ray-tracing. Thus the spot component $x_{E\beta\Gamma}$ of (6.2) may be differentiated immediately after the completion of the ray-trace which determines it and the contributions of $x_{E\beta\Gamma}$ and $x_{E\beta\Gamma,i}$ may be promptly added to g_{ij} and b_i after which both $x_{\beta\Gamma}$ and $x_{E\beta\Gamma,i}$ may be forgotten. This avoidance of storing the full matrices $\|E_A\|$ and $\|E_{A,i}\|$ reduces the storage demand in forming (23.1) from Holladay's 457,512 words to our 5,523 words.

We return now to Eq. (6.5) which shows the "delayed" weight assignment $w(\xi)\equiv[c(\xi)/(c_{rmsq}R_{rmsq})]|\xi|$ attached to any error measurement ξ, other than the centroidal error measurements (6.2), in forming the error component $E(\xi)\equiv w(\xi)\xi$. We say "delayed" because R_{rmsq} defined by (7.3) is not known at the time of the ray-trace measurement of ξ and hence we must postpone the division by R_{rmsq} until the entire entrance pupil mesh has been traced. The contribution of all such delayed weight components $E(\xi)$ of E are added in auxiliary storage blocks g'_{ij} and b'_i, initialed by 0, and at the final time of determining R_{rmsq} we make the storage replacement new $g_{ij} \equiv \text{old } g_{ij} + g'_{ij}/R_{rmsq}$ and similarly for the b_i.

31. COLASL CODING LANGUAGE

Our lens design code has been written in the COLASL language published in 1963 at Los Alamos by G. L. Carter, K. G. Balke, and B. A. Bacon. As an illustration of COLASL we form $[\frac{1}{2}(g_{rs}z_rz_s)^{\frac{1}{2}}-\sqrt{5}]/13$:

```
     "Form $|z|^2=g_{rs}z_rz_s$, initial sum"  $z_{mag}=0$.
     Thru R1(s=1, 2, · · ·, n).
R1   $z_{mag}=z_{mag}+g_{r,s}z_rz_s$, (r=1, 2, · · ·, n).
     $z_{mag}=(0.5\sqrt{z_{mag}} - \sqrt{5})/13$.
```

This coding language is outstandingly superior to any other such language that we have seen in its close resemblance to standard mathematical formalism and as a universal programming Latin for communicating the details of a code to any computing laboratory.

We intend to publish the complete COLASL lens design program, order by order, as a LASL report, to be ready hopefully in 1966. With the help of this publication of our COLASL code and with a COLASL manual, the latter now available on request from Glenn L. Carter, Los Alamos Scientific Laboratory, Los Alamos, N. M., any experienced programmer can write our lens design program in a language suitable to his machine. We emphasize that, as an automatic coding device, COLASL is restricted to our own IBM 7030 (Stretch) machine. It is possible, however, that by

1966 we may be able to supply a binary lens design deck capable of being run on any IBM 7030.

32. TWO PROPHECIES

It seems fitting that a paper begun with missionary intent and developed with evangelistic zeal should end with a prophecy. Such an utterance from the sibyl's cave is not without precedent in the literature. Herzberger concludes his book, MODERN GEOMETRICAL OPTICS, with this prophecy.

"The author believes that the most valuable development in theoretical optics in the near future will consist of analyzing and simplifying fifth-order approximation formulae and studying fifth-order models of various types of optical systems."

The truncated Taylor series in optical theory was the natural approach to approximating tedious, but exact, computation. This began by retaining the first order terms, and was successively refined to retain next the third and now the fifth order terms. Do these fifth order truncations provide the ultimate in desired precision? We think not. A new era of computational feasibility has dawned. The lens designer has but to rub the magic lamp and the jinni of the lamp appears to do his bidding.

Perhaps it is the specification "theoretical optics" which may save Herzberger's prophecy from becoming discredited. If we replace this by "practical lens design," then our own assault on prophecy prompts us to declare that the truncated Taylor series method in practical lens design will be encased in glass, properly labeled, and given an honored place in the Smithsonian Museum between the Wright brothers' plane that was launched at Kitty Hawk and Lindbergh's _Spirit of St. Louis_!

ACKNOWLEDGMENTS

We owe our participation in this problem to Berlyn Brixner, who has long contended that, since ray-tracing tests a lens, so logically should it be capable of designing it, and to John C. Holladay who supported Brixner by bidding a mathematician's adieu to the truncated Taylor series in lens design. We thank Charles A. Lehman for explaining the interior of his IBM 7090 version of the Holladay code. To Glenn C. Carter and his assistant Barbara A. Bacon we are grateful for their excellent scientific coding language COLASL and for their unfailing cooperation in helping us over our occasional misuse of it. We congratulate our own assistant, Joan H. Rohrer, who coded Herzberger's ray-trace to within a single bug which she herself promptly exterminated!

FOOTNOTES

[1] John C. Holladay, Computer Applications-1960, B. Mittman and A. Ungar, eds. (MacMillan Co., New York, 1960) pp. 112-127.

[2] Donald P. Feder, Appl. Opt. 2, 1214 (1963).

[3] Richard E. von Holdt, J. Assoc. Comp. Mach. 3, 223-238 (1956).

[4] J. B. Rosser, C. Lanczos, M. R. Hestenes, W. Karush, J. Research Natl. Bur. Standards 47, 291-297 (1951).

[5] F. Wachendorf, Optik 12, 329-359 (1955).

[6] Berlyn Brixner, Appl. Opt. 2, 1281-1286 (1963).

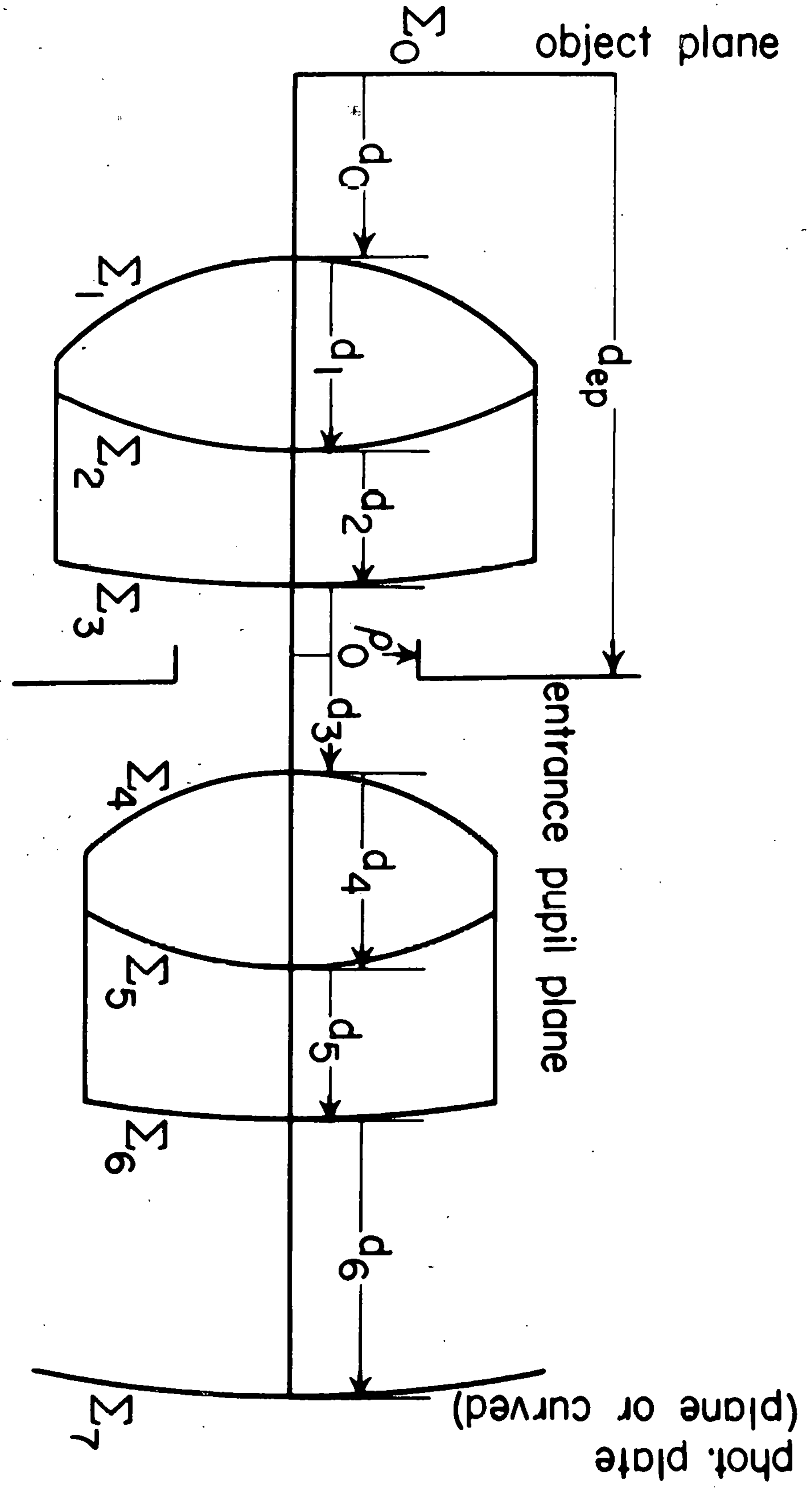
[7] Max Herzberger, Modern Geometrical Optics, (Interscience Publishers, Inc., New York, 1958), Part I, pp. 1-68.

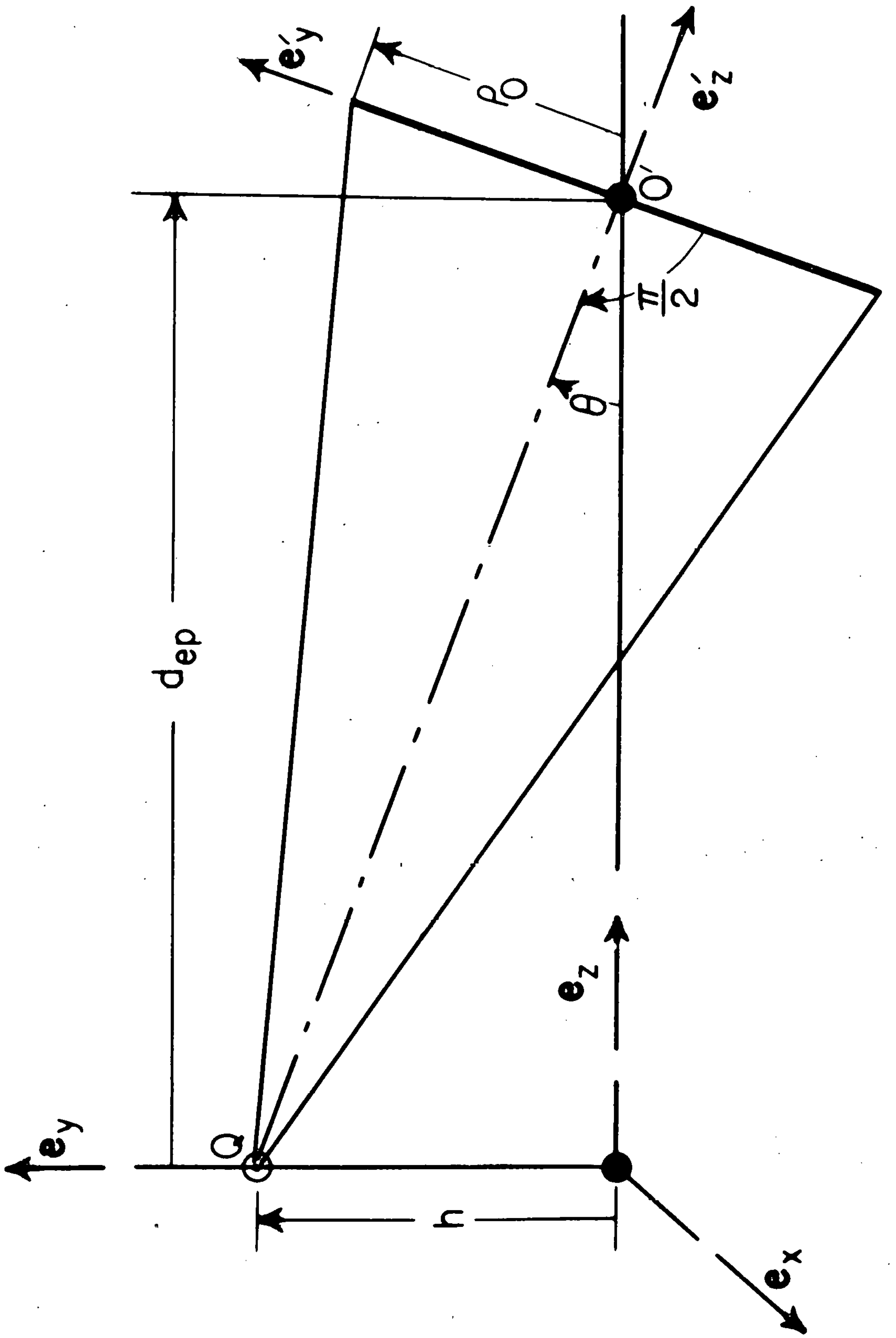
[8] R. Courant, Differential and Integral Calculus, (Interscience Publishers, Inc., New York, 1950), Vol. II, Chap. III, p. 188.

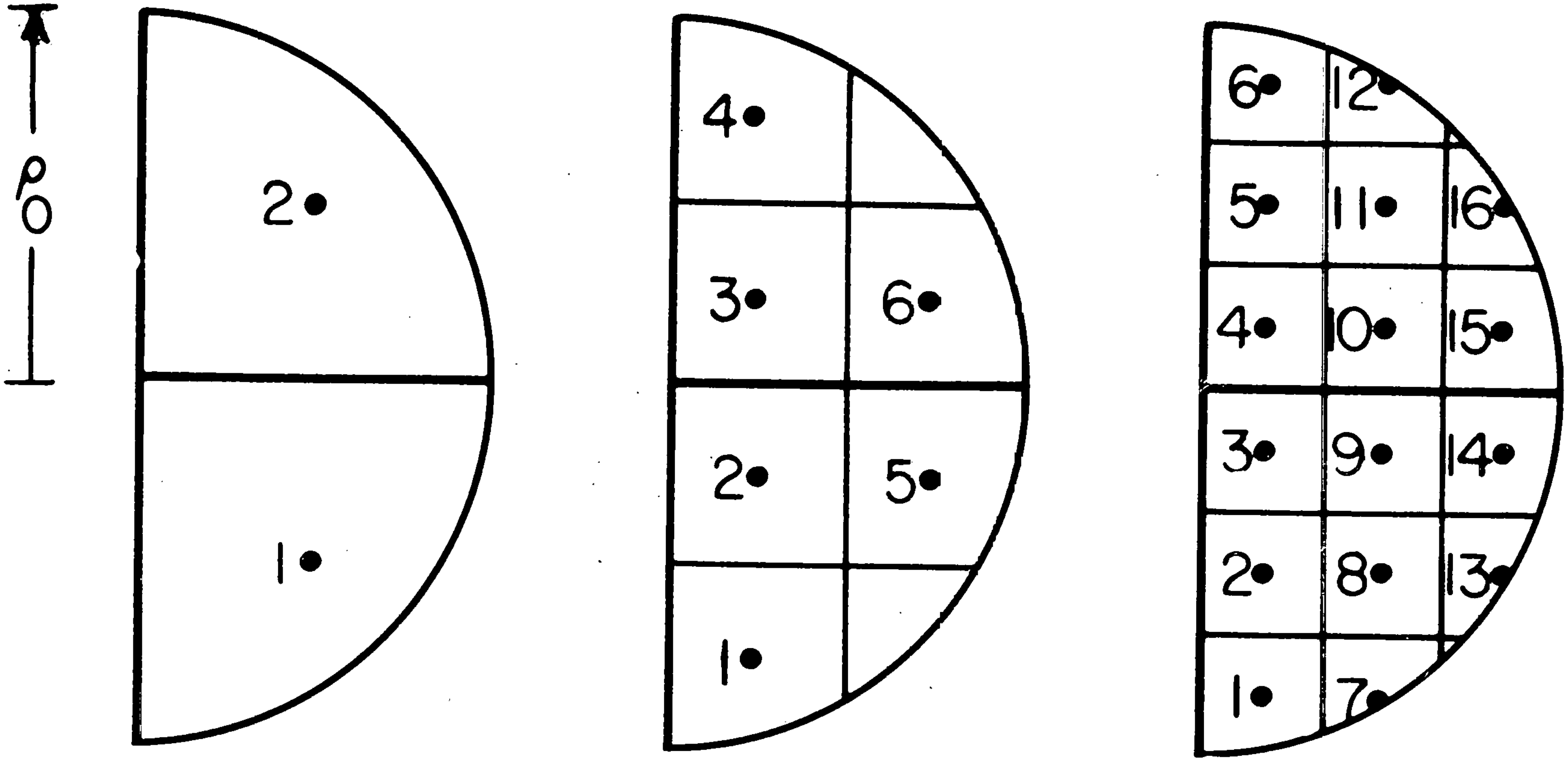
[9] A. L. Cauchy, Compt. rend. 25, 536-538 (1847).

[10] Charles A. Lehman, Los Alamos Scientific Laboratory Report LA-2837 (1963).

ILLUSTRATIONS


object plane
Σ_0
d_0
d_{ep}
d_1
d_2
ρ_0
d_3
d_4
d_5
d_6
entrance pupil plane
Σ_1
Σ_2
Σ_3
Σ_4
Σ_5
Σ_6
Σ_7
phot. plate
(plane or curved)


$\mathbf{e}'_y$
ρ_0
$\mathbf{e}'_z$
O'
$\frac{\pi}{2}$
θ
d_{ep}
$\mathbf{e}_z$
$\mathbf{e}_y$
Q
h
$\mathbf{e}_x$



$q_0 = 1$

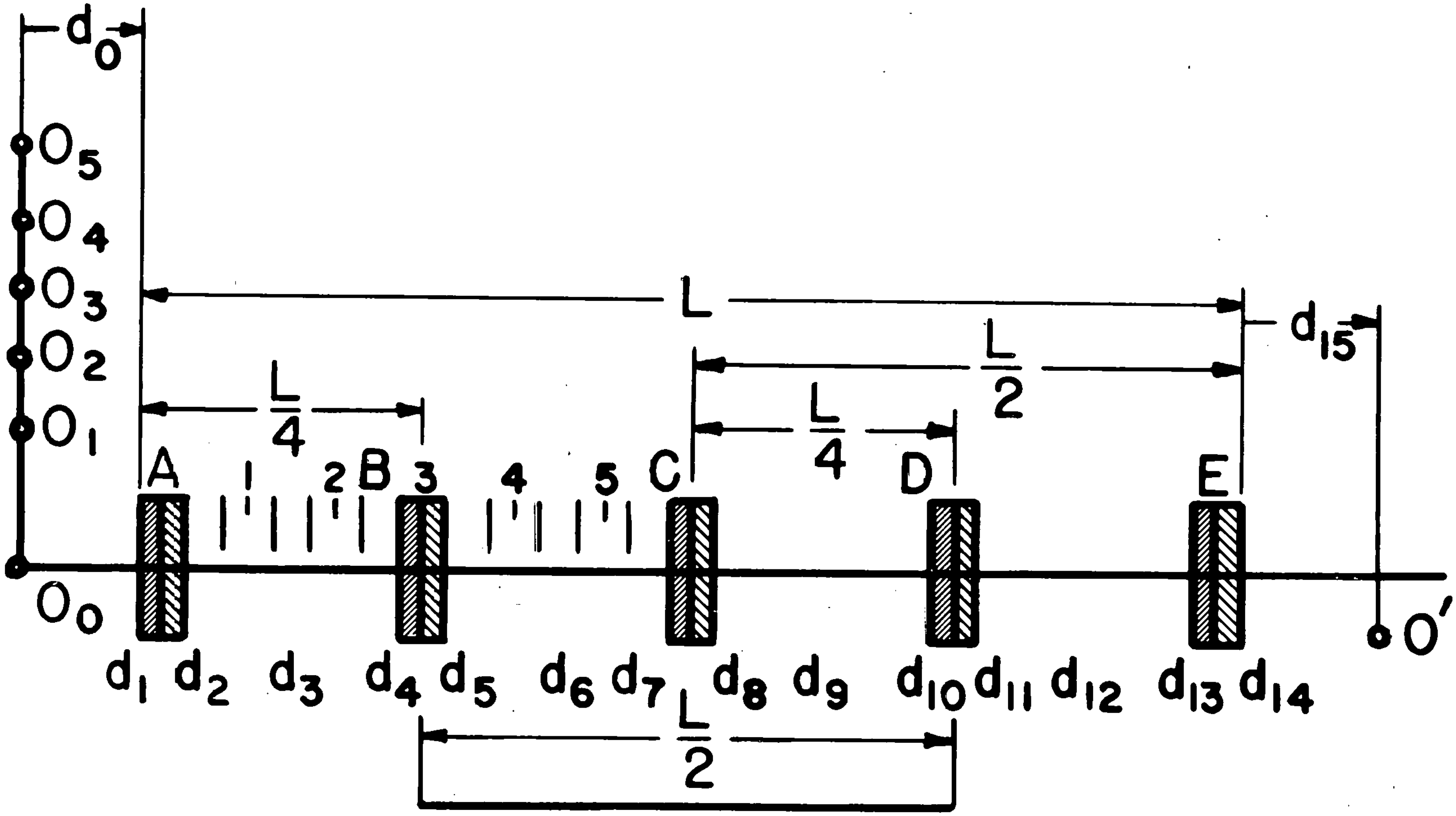
$\mathbf{q} \equiv (1)$

$q_0 = 2$

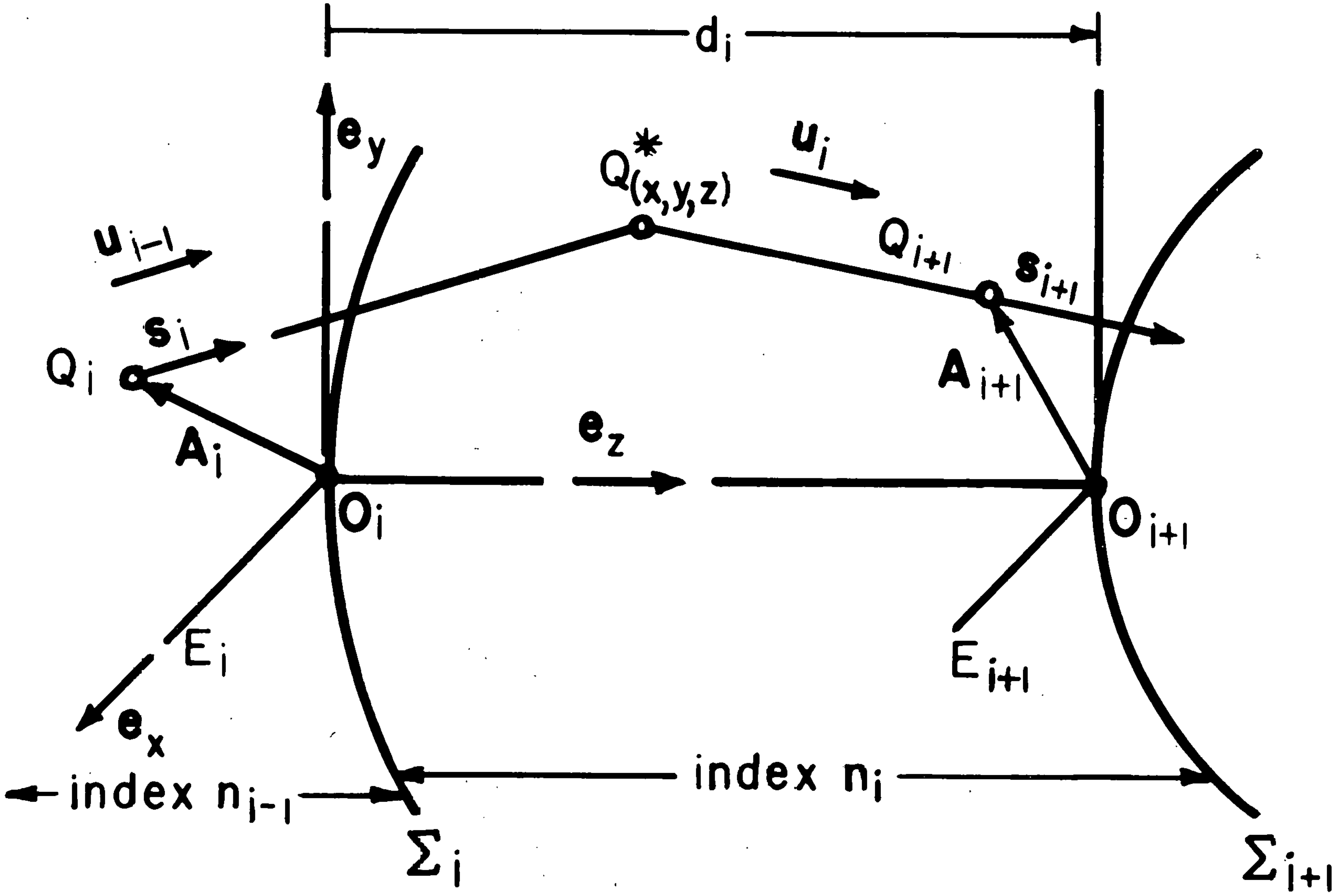
$\mathbf{q} \equiv (2,1)$

$q_0 = 3$

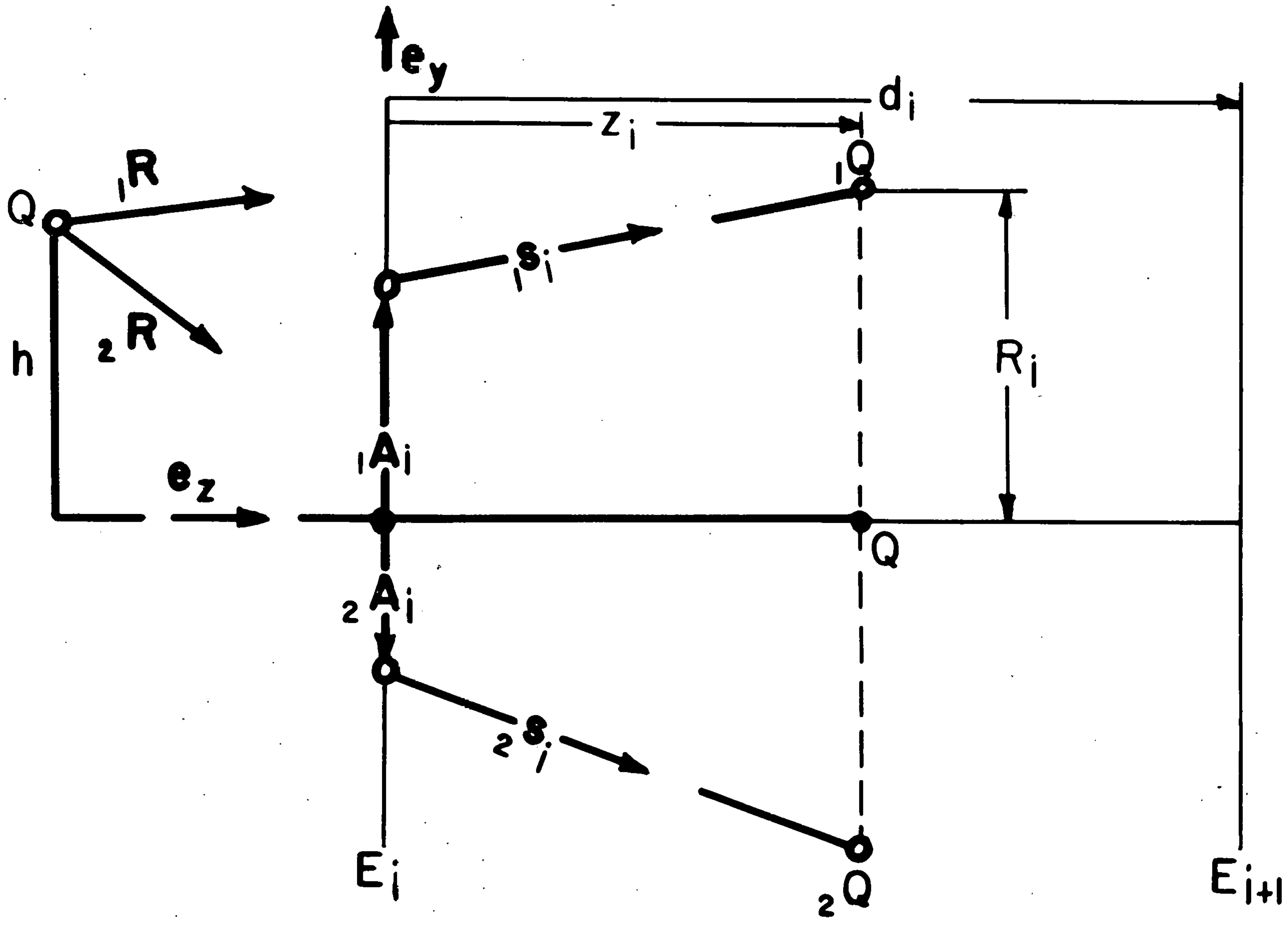
$\mathbf{q} \equiv (3,3,2)$

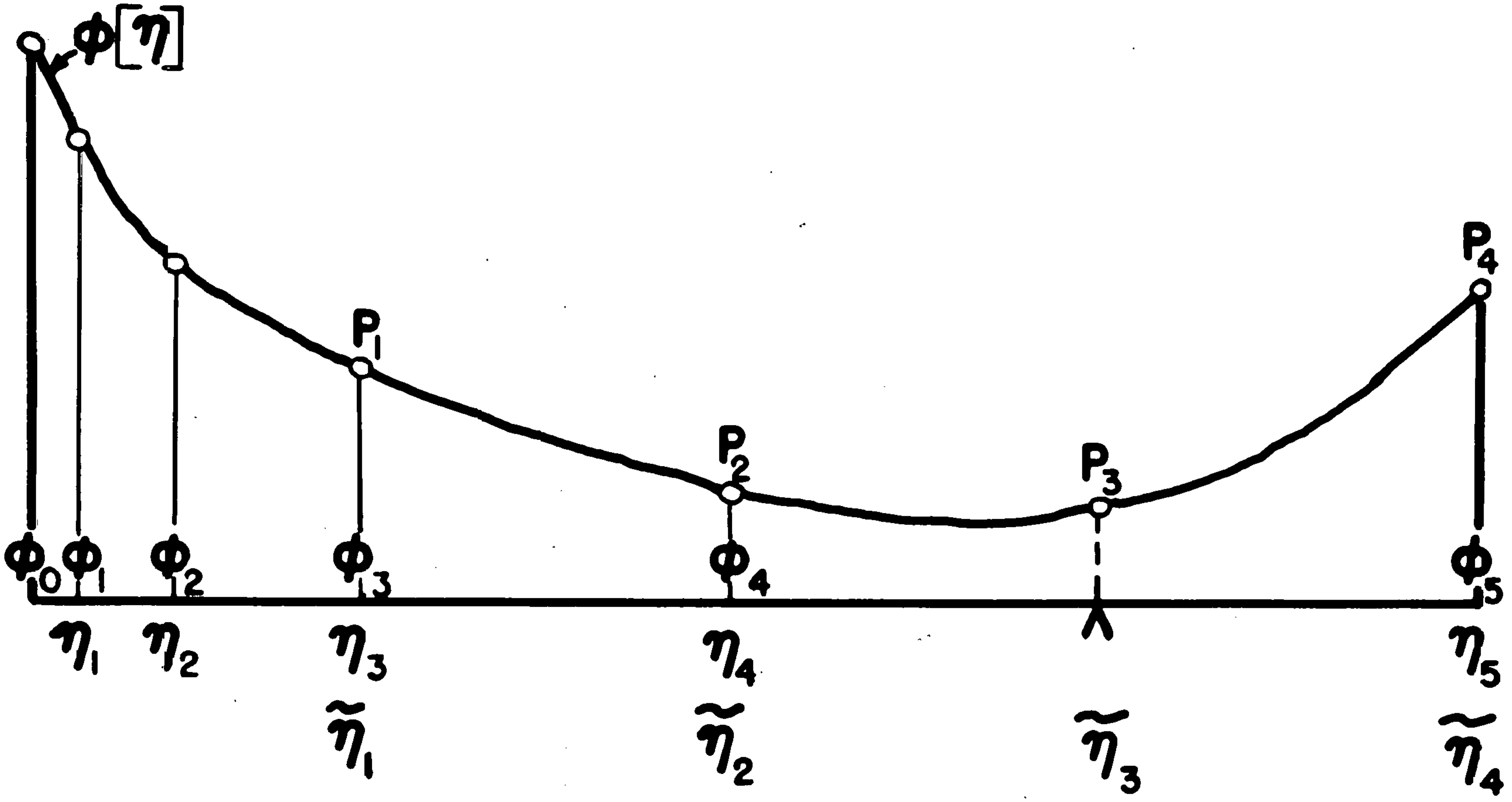


d_0
O_5
O_4
O_3
O_2
O_1
O_0
L
d_{15}
$\frac{L}{2}$
$\frac{L}{4}$
$\frac{L}{4}$
A
1
2
B
3
4
5
C
D
E
O'
d_1
d_2
d_3
d_4
d_5
d_6
d_7
d_8
d_9
d_{10}
d_{11}
d_{12}
d_{13}
d_{14}
$\frac{L}{2}$

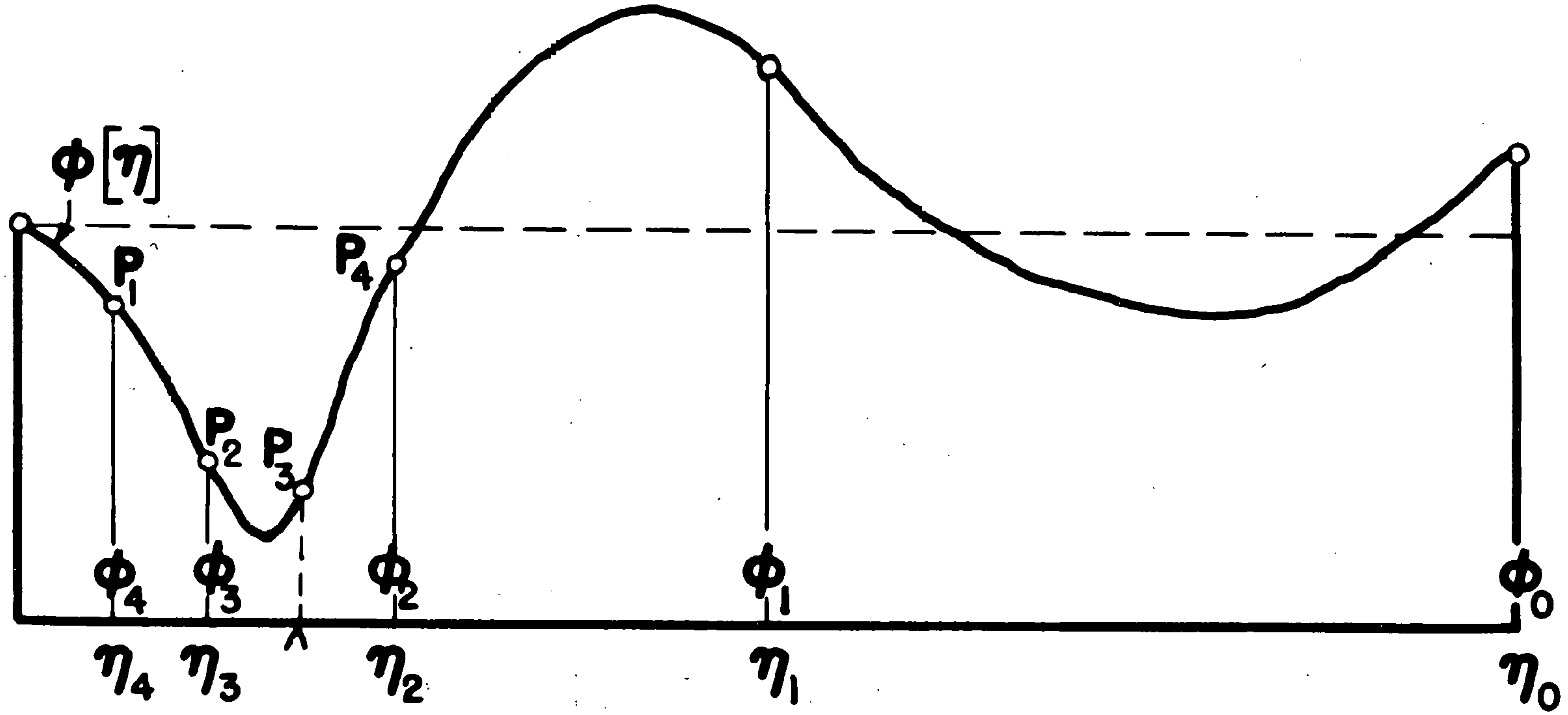


d_i
$\mathbf{e}_y$
$Q^*_{(x,y,z)}$
$\mathbf{u}_i$
$\mathbf{u}_{i-1}$
Q_{i+1}
$\mathbf{s}_{i+1}$
$\mathbf{s}_i$
Q_i
$\mathbf{A}_{i+1}$
$\mathbf{e}_z$
$\mathbf{A}_i$
O_i
O_{i+1}
E_i
E_{i+1}
$\mathbf{e}_x$
index n_i
index n_{i-1}
Σ_i
Σ_{i+1}


e_y
d_i
z_i
$_1Q$
Q
$_1R$
$_2R$
h
$_1s_i$
R_i
$_1A_i$
e_z
Q
$_2A_i$
$_2s_i$
E_i
$_2Q$
E_{i+1}



$\Phi[\eta]$
P_1
P_2
P_3
P_4
Φ_0
Φ_1
Φ_2
Φ_3
Φ_4
Φ_5
η_1
η_2
η_3
η_4
η_5
$\tilde{\eta}_1$
$\tilde{\eta}_2$
$\tilde{\eta}_3$
$\tilde{\eta}_4$


$\phi[\eta]$
P_1
P_4
P_2
P_3
ϕ_4
ϕ_3
ϕ_2
ϕ_1
ϕ_0
η_4
η_3
η_2
η_1
η_0