DOE/ER 13726-6

FUNDAMENTALS AND TECHNIQUES OF NONIMAGING OPTICS

Roland Winston

The University of Chicago
The Enrico Fermi Institute
Chicago, Illinois 60637

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or 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.

April 1993

Prepared for

THE U.S. DEPARTMENT OF ENERGY

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

A. HIGHLIGHTS

1992 (March) Sede Boqer Symposium in Nonimaging Optics Sede Boqer, Israel

1992 (November) Sacramento Municipal Utility District (SMUD) and NREL Workshop on the potential of evacuated integrated CPC solar thermal collectors for application to air conditioning. Sacramento, CA.
FUTURE CONFERENCES SCHEDULED:

1993(July) International Symposium on Nonimaging Optics: Maximum Efficiency Light Transfer, SPIE 1991 Annual Meeeting at San Diego Ca.

1993 (October) Symposium on Nonimaging Optics and Illumination Systems, Optical Society of America Annual Meeting in Toronto Canada.

B. NONIMAGING OPTICS AT OTHER INSTITUTIONS

At the Sede Boqer Symposium, a Nonimaging Optics bibliography of publications in peer-reviewed journals and books was presented. This contains some 450 articles and 20 books. The following is a partial list of active work:

Hewlett-Packard has constructed light emitting diodes encapsulated in CPC collectors, so as to direct the light forward over a well-defined range of directions. These are used in the '92 Ford Thunderbird rear illumination red applique.

Midway Labs, a Chicago based company, is manufacturing assemblies consisting of a Fresnel lens and a nonimaging secondary followed by a photovoltaic cell.

The original work on focal plane nonimaging astronomical systems with far infrared radiation carried out by Roger Hildebrand and co-workers has continued and this system is now in wide use. An example is the John Mather photometer for measuring the nominally 3° cosmic black body radiation in space (COBE sattelite launched by NASA in '89).

The University of Chicago has helped to set up a company to apply nonimaging optics to a wide variety of problems; the company is NiOptics Inc., Evanston, Illinois.

The New Energy Development Organization of Japan has selected a CPC as their single choice for mid-temperature solar heat utilization. This is being developed by Koto Electric Co. in Japan.

In the Weitzmann Institute of Science (Israel) second stage nonimaging systems are being used for solar pumping of large-scale lasers and other applications. Laser power has exceeded 300 Watts and is expected to approach a killowatt. A large energy company in Israel is using and marketing CPC solar heaters for space heating in institutions and companies.

At Ben Gurion University (Sede Boqer, Israel) Nonimaging Optics is being developed for solar collection and for illumination (J. Gordon and colleagues).

In Madrid at the Polytechnic Institute, as adjunct to the work on photovoltaics, significant theoretical and experimental work on nonimaging concentrators is in progress, by Luque, Minano and colleagues.

At the University of Sydney (Australia) nonimaging optics for solar collection and for illumination is being studied.

At Lockheed/Sanders (J. D. Kuppenheimer) CPC designs for infra-red countermeasure jammers are in use.

I. OVERVIEW

Nonimaging optics began in the mid 60's with the discovery that optical systems could be designed and built that approached the theoretical limit of light collection (the sine law of concentration). [A semi-popular account is given in R. Winston, Scientific American cover article, March, 1991.] Since its inception, the field has undergone three periods of rapid conceptual development. In the 70's the "string" or "edge-ray" method [W. T. Welford and R. Winston, High Collection Nonimaging Optics (Academic, New York, 1989)]. (the "Hottel string" is a useful construct for calculating radiative transfer between lambertian surfaces) [W. H. McAdams, Heat Transmission (McGraw-Hill, New York, 1964)] was formulated and elaborated for a large variety of geometries. This development was driven by the desire to design wide-angle solar concentrators. It may be succinctly characterized as: $\int n dl$ = constant along a string. [Notice that replacing "string" by "ray" (Fermat's principle) gives all of imaging optics.] In the early 80's, a second class of algorithms was found, driven by the desire to obtain ideally perfect solutions in three dimensions (3-D). (The "string" solutions are ideal only in 2-D, and as figures of revolution in 3-D are only appoximately ideal, though still very useful). This places reflectors along the lines of flow of a radiation field set up by a radiating lambertian source. In cases of high symmetry such as a sphere or disc, one obtains ideal solutions in *both* 2-D and 3-D. The third period of rapid development has taken place only in the past year; its implications and consequences have yet to be worked out. This was driven by the desire to address a wider class of problems in illumination that could not be solved by the old methods. (specifically an infra-red counter-measure beam). Here are two examples: It is well-known that the far-field illuminance from a lambertian source falls off with a power of the cosine of the radiating angle α. For example, strip radiators produce a $\cos^3\alpha$ illuminance on a distant plane, while circular disc radiators produce a $\cos^4\alpha$ illuminance. But suppose one desires a predetermined far-field illuminance pattern e.g., uniform illuminance? The old designs will not suffice; they simply transform a lambertian source radiating over 2π into a lambertian source radiating over a restricted set of angles. Another example is more technical. We recall that older nonimaging designs require that reflectors be positioned very close to the source (or receiver). Violating this rule introduces

undesirable structure in the radiating or angular acceptance pattern of the device, typically a dip in the forward direction. The limitation of the old designs is that they are too static and depend on a few parameters such as, the area of the beam A_1 and the divergence angle θ. One needs to introduce additional degrees of freedom into the nonimaging designs to solve a wider class of problems.

Nonimaging optics is the optics of extended sources. That is why the subject has more in common with radiative transfer than with conventional optical design and relies on such notions as "Hottel strings". In contrast, imaging optics in its geometrical considerations is the optics of point sources. But in considering extended sources, one is led to distributions in phase space and inevitably to the theory of Radiance. This is a mature subject: pioneered by Adrian Walther, beautifully developed by many workers, notably by Emil Wolf and his school. Our work in this area is driven by the desire to describe radiance in nonimaging optical systems. The systems currently studied are quasi-homogenous which means that the cross-spectral density $W(r_1,r_2)$ has the form $I\,[{}^1/_2(r_1 + r_2)]\,g\,(r_1 - r_2)$ so that the correlation is translationally invariant. [Carter, W.H. and E. Wolf JOSA 67,785-796,1977]. But boundaries are significant in nonimaging systems (after all, the edge-ray algorithm is one of the most useful in the subject) so that the quasi-homogenous model is not suitable . We are collaborating with the University of California/Berkeley group of Robert Littlejohn and Allan Kaufman to address this difficult problem. Our approach is to study the evolution of various distribution functions along rays, since in classical radiometry this evolution is null (the radiance is conserved along rays)

A. RECENT PROGRESS IN NONIMAGING OPTICS

Slightly over one year ago, we presented the proposal that nonimaging designs be regarded as functionals of the desired irradiance, rather than depending on a static parameter such as the "acceptance angle". [R. Winston, Nonimaging Optics: optical design at the thermodynamic limit. In Nonimaging Optics: Maximum Efficiency Light Transfer, pages 2-6. Proc. SPIE 1528, Roland Winston and Robert L. Holman, editors, July 1991.] The response of the "nonimaging optics community" was gratifying; (I have not witnessed comparable excitement since the solar collector developments of the '70's). Two journal articles

from colleagues are in press. Our own work is best described in a paper being prepared for publication in J.O.S.A. A in collaboration with Harald Ries. An earlier version elicited the following remark from an anonymous referee: "I find the material (in section 4) to be new and original. It forms an important solution to a fundamental problem in illumination optics that, to the best of my knowledge, has never been tackled successfully, namely, generating a uniform far-field illuminance from an extended radiation source. The method of solution also opens up a new approach to the more general problem, which the authors show, of producing almost any illuminance pattern from a given extended radiation source." It is reproduced here in its entirety. [The authors are Roland Winston and Harald Ries.]

Abstract

For many tasks in illumination and collection the acceptance angle is required to vary along the reflector. If the acceptance angle function is known, then the reflector profile can be calculated as a functional of it. The total flux seen by an observer from a source of uniform brightness (radiance) is proportional to the sum of the view factor of the source and its reflection. This allows one to calculate the acceptance angle function necessary to produce a certain flux distribution and thereby construct the reflector profile. We demonstrate the method for several examples, including finite size sources with reflectors directly joining the source.

1 Introduction

Nonimaging optics was developed to solve a well posed but narrow set of problems [1]. A prototypical example is the concentration of a light beam with divergence half-angle θ and cross-sectional area A_1 into the minimum possible area A_2 without loss of throughput or conversely, the design of illumination systems that convert a lambertian source into a beam with divergence half-angle θ and no stray light without loss of throughput. Two classes of algorithms have been found which solve these problems exactly or nearly so. These are summarized here; the details can be found in Ref.[2]. The first is the "string" or "edge-ray" method. The "Hottel string" is a useful concept for calculating radiative transfer between lambertian surfaces [3]. It may be succinctly characterized as: $\int n\, dl = constant$ along a string, where n denotes the index of refraction and dl the path length. Notice that replacing "string" by "ray" gives all of imaging optics. The second class of algorithms places reflectors along the lines of flow of a radiation field set up by a radiating source. In cases of high symmetry such as a sphere or disc, one obtains ideal solutions in both two and three dimensions. In either case, reflecting and sometimes refracting elements are shaped in specific ways in combination to solve the problem.

A wider class of problems can not be solved by the known methods. Here are a few examples:

It is well-known that the irradiance on a distant plane at an angle θ from a long, cylindric lambertian source of uniform brightness falls off with $\cos^2(\theta)$. Strip radiators and spherical sources produce a $\cos^3(\theta)$ irradiance on a distant plane, while circular disc radiators produce an irradiance proportional to $\cos^4(\theta)$. The angular power density of the flat sources (disc and strip) falls off as $\cos(\theta)$ while the power density of cylindric and spherical sources is constant. But suppose one desires a predetermined far-field power or irradiance pattern e.g. uniform irradiance? The classical designs will not suffice; they simply transform a lambertian source radiating over 2π into a lambertian source radiating over a restricted angular range.

The limitation of the old designs is that they are too static and depend only on a few parameters, such as the area of the beam A_1 and the divergence angle θ. One needs to introduce additional degrees of freedom into the nonimaging designs to solve a wider class of problems. The purpose of this communication is to indicate the lines along which this additional freedom can be introduced.

2 Determining the reflector profile for small sources

In the usual design methods the profile of the reflector is determined by the given constant acceptance angle θ and the geometry of the entrance and exit surfaces. Thus we can regard the reflector profile R as a function of θ, $R(\theta)$. However, in certain situations a "constant acceptance angle" design is unduly restrictive. But suppose θ is itself made a function of some other parameter of the problem say, ϕ. Then R is determined only after the functional relationship of θ and ϕ is known i.e., R is now a *functional* of θ, $R = R\{\theta\}$.

For illustrative and pedagogical reasons, we will consider first the simple case when the size of the source is much less than the closest distance of approach R_0 to any reflective or refractive component. Thus the angle subtended by the source at any reflective or refractive component may be regarded as small. Our approximation of small source dimension d and large observer distance D amounts to

$$d << R_0 << D. \tag{1}$$

In this limit the illumination problem has been solved earlier [4] We briefly review the classical solution before we introduce a novel approach capable of deriving in closed form the reflector for large sources.

Polar coordinates are used to represent the reflector profile by $R = R(\phi)$, with the source at the origin. The angle of the reflected ray with the optical axis is denoted by θ, and the incidence angle at the reflector with respect to its normal is denoted by α as depicted in Fig.1. The geometry shows that the following relation between the reflector profile and incidence angle holds:

$$\frac{d\log(R)}{d\phi} = \tan(\alpha). \tag{2}$$

Note, that the underlying assumption for this equation is, that the edge rays incident onto

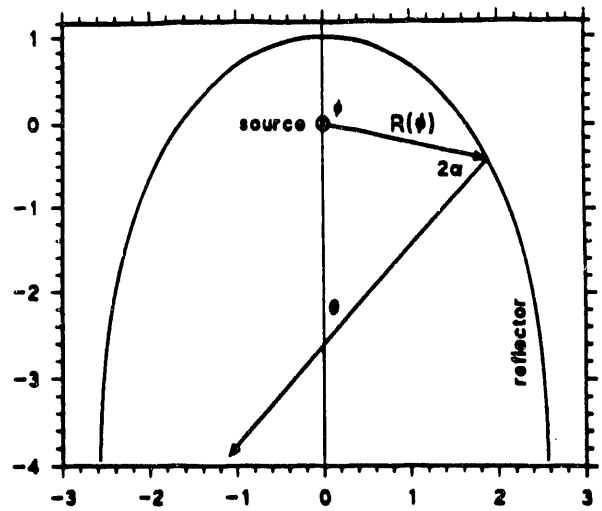

Figure 1: The reflector profile is represented in polar coordinates $R(\phi)$ with the source at the origin. The reflected radiation has an angle θ with the optical axis y and α with the normal to the reflector.

the reflector travel along the vector R. Clearly,

$$\alpha = \frac{\phi - \theta}{2}. \tag{3}$$

Equation 2 is readily integrated,

$$\log\left(\frac{R(\phi)}{R_0}\right) = \int_0^{\alpha(\phi)} \tan(\alpha(\phi))d\phi \tag{4}$$

so that,

$$R(\phi) = R_0 \exp\left(\int_0^{\alpha(\phi)} \tan(\alpha(\phi))d\phi\right). \tag{5}$$

This determines the reflector profile $R(\phi)$ for any desired acceptance angle function $\theta(\phi)$.

Suppose we wish to radiate power with a particular angular distribution $P^o(\theta)$ from a source which itself radiates with a power distribution $P^s(\phi)$. The angular characteristic of the source is the combined result of its shape, surface brightness, and surface angular emissivity at each point. A distant observer viewing the source fitted with the reflector under an angle θ will see a reflected image of the source in addition to the source itself. This image will be magnified by some factor $|M|$ if the reflector is curved. Ideally both the source and its reflected image have the same brightness, so the power each produces is proportional to the apparent size. The intensity perceived by the observer, $P^o(\theta)$ will be the sum of the two

$$P^o(\theta) = P^s(\theta) + |M|P^s(\phi). \tag{6}$$

The absolute value of the magnification has to be taken because, if the reflected image and the source are on different sides of the reflector and we therefore perceive the image as reversed or upside down, then the magnification is negative. Actually, at small angles, the

source and its reflection image may be aligned so that the observer perceives only the larger of the two. But if $|M|$ is large one can neglect the direct radiation from the source.

Thus one is concerned with the magnification of the reflector. A distant observer will see a thin source placed in the axis of a trough reflector magnified in width by a factor

$$M_m = \frac{d\phi}{d\theta}. \tag{7}$$

This can be proved from energy conservation. The power emitted by the source must be conserved upon reflection: $P^e d\phi = M\, P^e d\theta$.

For a rotationally symmetric reflector the magnification M_m as given in Eq.(7) refers to the meridional direction. In the sagittal direction the magnification is

$$M_s = \frac{d\mu_1}{d\mu_2} = \frac{\sin(\phi)}{\sin(\theta)}, \tag{8}$$

where now μ_1 and μ_2 are small angles in the sagittal plane, perpendicular to the cross section shown in Fig 1. Equation (8) can be easily verified by noting that the sagittal image of an object on the optical axis must also lie on the optical axis. The reason is, that because of symmetry, all reflected rays must be coplanar with the optical axis.

The total magnification M_t is the product of the sagittal and the meridional magnification

$$M_t = M_s M_m = \frac{d\cos(\phi)}{d\cos(\theta)}. \tag{9}$$

Actually Eq.(9) could also have been derived directly from energy conservation by noting that the differential solid angle is proportional to $d\cos(\theta)$ and $d\cos(\phi)$ respectively.

Thus inserting the magnification given in Eq.(9) or Eq.(7), as the case may be, into Eq.(6) yields the relationship between θ and ϕ which produces a desired power distribution $P^o(\theta)$ for a given angular power distribution of the source P^e. This relationship then can be integrated as outlined in Eq.5 to construct the shape of the reflector which solves that particular problem.

There are two qualitatively different solutions depending on whether we assume the magnification to be positive or negative. If $M_m > 0$ this leads to CEC-type devices, whereas $M_m < 0$ leads to CHC-type devices. The term CEC refers to Compound Elliptical Concentrator and CHC to the so called Compound Hyperbolic Concentrator [5, 6, 7, 8].

Now the question arises of how long we can extend the reflector or over what angular range we can specify the power distribution. From Eq.(5) one can see that if $\phi - \theta = \pi$ then R diverges. In the case of negative magnification this happens when the total power seen by the observer between $\theta = 0$ and $\theta = \theta^{max}$ approaches the total power radiated by the source between $\phi = 0$ and $\phi = \pi$. A similar limit applies to the opposite side and specifies θ^{min}. The reflector asymptotically approaches an infinite cone or V-trough. There is no power radiated or reflected outside the range $\theta^{min} < \theta < \theta^{max}$.

For positive magnification the reflected image is on the opposite side of the symmetry axis (plane) to the observer. In this case the limit of the reflector is reached as the reflector on the side of the observer starts to block the source and its reflection image. For symmetric devices this happens when $\phi + \theta = \pi$. In this case too one can show that the limit is actually imposed by the first law. However, the reflector remains finite in this limit. It always ends with a vertical tangent. For symmetric devices where $\theta^{max} = -\theta^{min}$ and $\phi^{max} = -\phi^{min}$ the extreme directions for both the CEC-type and the CHC-type solution are related by

$$\phi^{max} + \theta^{max} = \pi. \tag{10}$$

In general CEC-type devices tend to be more compact. The mirror area needed to reflect a certain beam of light is proportional to $1/\cos(\alpha)$. The functional dependence of θ and ϕ for symmetrical problems is similar except that they have opposite signs for CHC-type devices and equal signs for CEC-type solutions. Therefore α increases much more rapidly for the CHC-type solution which therefore requires a larger reflector, assuming the same initial value R_0. This is visualized in Fig.2 and where the acceptance angle function as well as the incidence angle α are plotted both for the negative magnification solution.

2.1 Simple Example: strip source

For a narrow, one-sided lambertian strip, the radiant power is proportional to the cosine of the angle. In order to produce a constant irradiance on a distant target the total radiation of source and reflection should therefore be proportional to $1/\cos^2(\theta)$. This yields

$$\cos\theta + \left|\cos(\phi)\frac{d\phi}{d\theta}\right| = \frac{a}{\cos^2(\theta)}. \tag{11}$$

The boundary condition is, in this case, $\theta = 0$ at $\phi = \pm\pi/2$ because we assume that the strip only radiates on one side, downward. Equation 11 can only be integrated for $a = 1$:

$$\sin(\phi) = 1 - |\tan(\theta) - \sin(\theta)|. \tag{12}$$

The acceptance angle function θ as well as the incidence angle for the CEC-type solution are shown in Fig.2. Integrating yields the reflector shapes plotted in Fig.3.

3 Reflector adjacent to a finite planar source

We have now developed the analytical tools to solve the real problems which involve reflectors close to the source. We do this by combining the above technique with the edge ray method which has proved so effective in nonimaging designs [2]. That is, we apply the above methods to edge rays. As a first example, we design a reflector for a planar, lambertian strip source so as to achieve a predetermined far-field irradiance. We design the reflector so that the

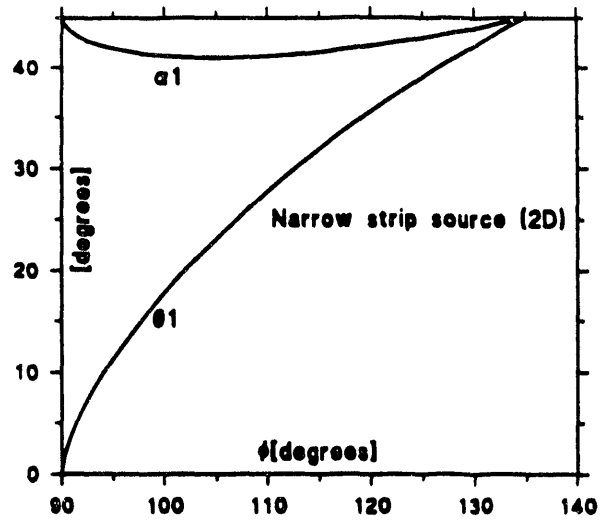

Figure 2: Acceptance angle function which produces a constant irradiance on a distant plane from a narrow one-sided lambertian strip source (2D) ; a=1.

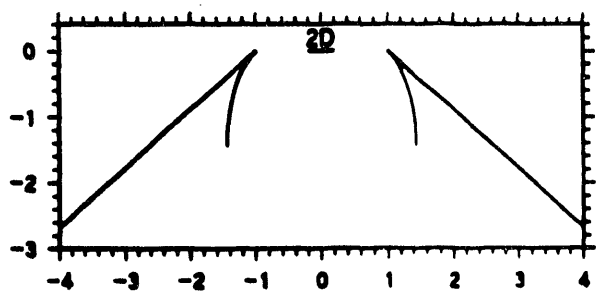

Figure 3: The reflector profile which produces a constant irradiance on a distant plane from a one-sided lambertian strip source (2D) at the origin, $R(\phi = \pi/2) = 1$, $a = 1$. CEC (inner curve) and CHC-type solutions (outer truncated curve) are shown.

reflected image is immediately adjacent to the source. This is only possible in a negative magnification arrangement. Then the combination of source and its mirror image is bound by two edge rays as indicated in Fig.4. The combined angular power density for a source of unit brightness radiated into a certain direction is given by the edge ray separation

$$R\sin(2\alpha) = P^o(\theta). \tag{13}$$

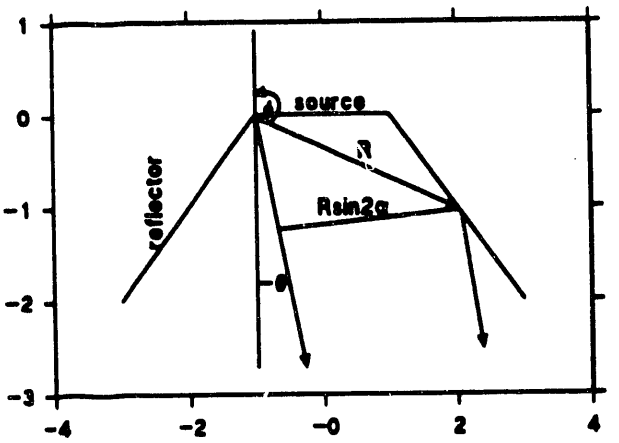

Figure 4: The reflector is designed to produce a reflected image adjacent to the source. The combined intensity radiated in the direction $-\theta$ is determined by the separation of the two edge rays: $R\sin 2\alpha$.

By taking the logarithmic derivative of Eq.(13) and substituting Eq.(2) we obtain

$$\frac{d\alpha}{d\theta} = \frac{\sin(2\alpha)}{2}\frac{d\log(P^o(\theta))}{d\theta} - \sin^2(\alpha). \tag{14}$$

This describes the right hand side, where $\theta < 0$. The other side is the mirror image.

3.1 Deriving the reflector shape directly for finite source

For $2\alpha = \pi$, R diverges just as in the case of the CHC-type solutions for small sources. Thus in general the full reflector extends to infinity. For practical reasons it will have to be truncated. Let's assume that the reflector is truncated at a point T from which the edge ray is reflected into the direction θ_T. For angles θ in between $\pm\theta_T$ the truncation has no effect because the outer parts of the reflector do not contribute radiation in that range. Therefore within this range the truncated reflector also produces strictly the desired illumination. Outside this range the combination of source plus reflector behaves like a flat source bounded by the point T and the opposite edge of the source. Its angular power density is given by Eq.(13) with $R = R_T = constant$. The total power P_T radiated beyond θ_T is thus

$$P_T = R(\theta_T)\int_{2\alpha_T}^{\pi} \sin\gamma d\gamma = R(\theta_T)(1 + \cos(2\alpha_T)). \tag{15}$$

In order to produce an intensity $P^o(\theta_T)$ at θ_T, $R(\theta_T)$ must be

$$R(\theta_T) = \frac{P^o(\theta_T)}{\sin(2\alpha_T)}. \tag{16}$$

The conservation of total energy implies that the truncated reflector radiates the same total power beyond θ_T as does the untruncated reflector.

$$\frac{1 + cos(2\alpha_T)}{\sin(2\alpha_T)} = \frac{1}{P^o(\theta_T)} \int_{\theta^{max}}^{\theta_T} P^o(\psi) d\psi =: B(\theta_T) \tag{17}$$

This equation must hold true for any truncation $\theta = \theta_T$. It allows to explicitly calculate α, and with it ϕ and R, in closed form as functions of θ, if $B(\theta)$, that is the integral of $P^o(\theta)$ is given in closed form. Solving Eq.(17) for α yields

$$2\alpha = \arccos\left(\frac{B^2 - 1}{B^2 + 1}\right). \tag{18}$$

Substituting Eq.(3) yields the acceptance angle function

$$\phi(\theta) = \theta + 2\alpha. \tag{19}$$

¿From Eq.(13) the radius is given by

$$R(\theta) = P^o(\theta)\frac{B^2 + 1}{2B}. \tag{20}$$

These equations specify the shape of the reflector in a parametric polar representation for any desired angular power distribution $P^o(\theta)$. A straight forward calculation shows that Eq.(18) is indeed the solution of the differential equation (14). In fact Eq.(14) was not needed for this derivation of the reflector shape. We have presented it only to show the consistency of the approach.

3.2 Example - constant irradiance

For example to produce a constant irradiance on a plane parallel to the source we must have $P^o(\theta) = 1/\cos^2(\theta)$ and thus $B(\theta) = \cos^2(\theta)(1 + \tan(\theta))$. The resulting acceptance angle function and the reflector profile are shown in Fig.5 and Fig.6 respectively. The reflector shape is close to a V-trough. Though, the acceptance angle function is only poorly approximated by a straight line, which characterizes the V-trough. In Fig.7 we show the deviation of the reflector shape depicted in Fig.6. from a true V-trough. Note, that a true V-trough produces a markedly non-constant irradiance distribution proportional to $\cos(\theta + \pi/4)\cos(\theta)$ for $0 < -\theta < \pi/4$.

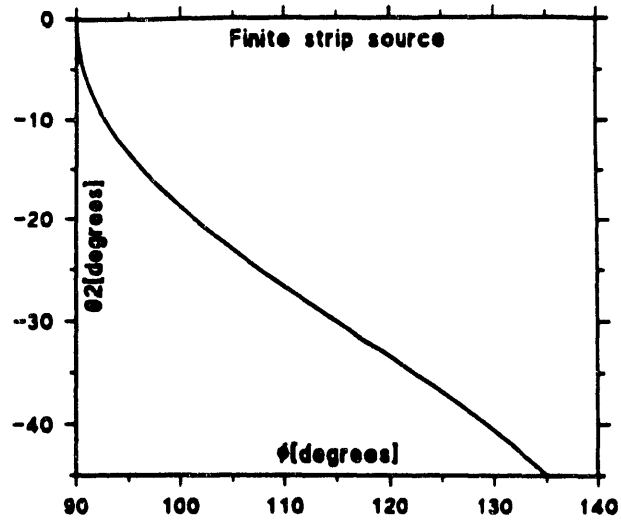

Figure 5: Acceptance angle function which produces a constant irradiance on a distant plane from a finite one-sided lambertian strip source. There is only a CHC-type solution.

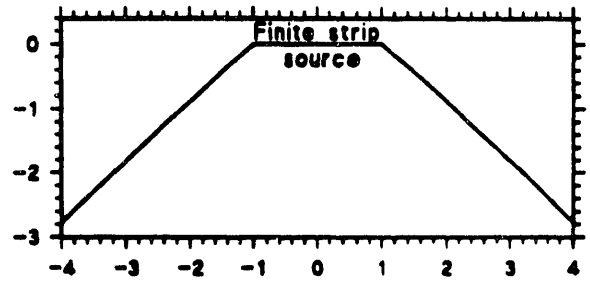

Figure 6: The reflector profile which produces a constant irradiance on a distant plane from a finite one-sided lambertian strip source of width two units. Note that there is only a CHC-type solution and it is truncated.

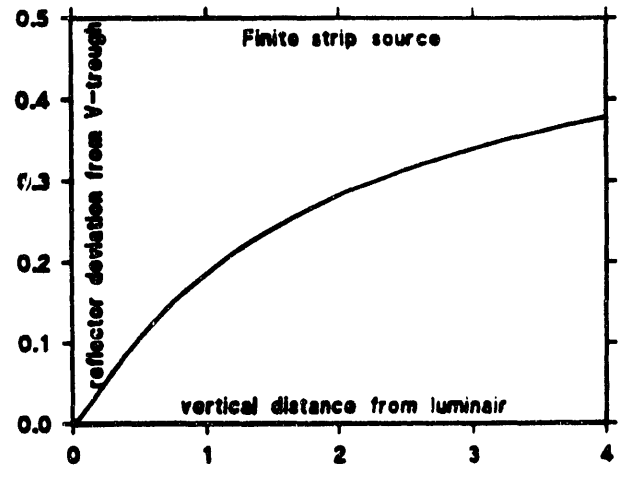

Figure 7: Deviation of the reflector depicted in Fig.6 from a true V-trough.

3.3 Example - specific non-constant irradiance

As a second example we design the reflector which produces the irradiance distribution on a plane shown in Fig.8. The corresponding angular power distribution is shown in Fig.9. The acceptance angle function according to Eq.(19) and (18) and the resulting reflector shape according to Eq.(20) are visualized in Fig.10 and Fig.11.

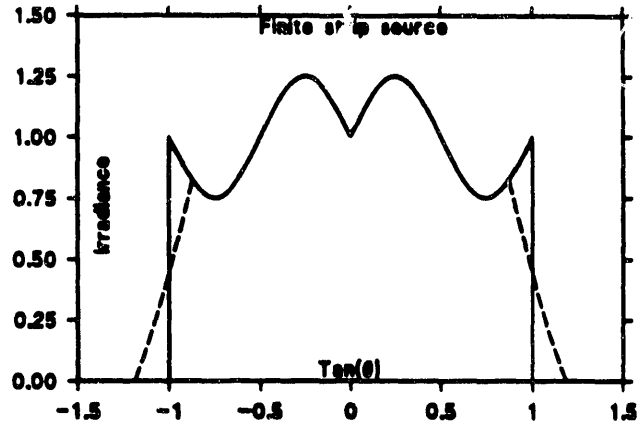

Figure 8: Desired irradiance distribution on a distant plane perpendicular to the optical plane divided by the irradiance produced along the axis by the source alone. Broken line shows the irradiance of a truncated device

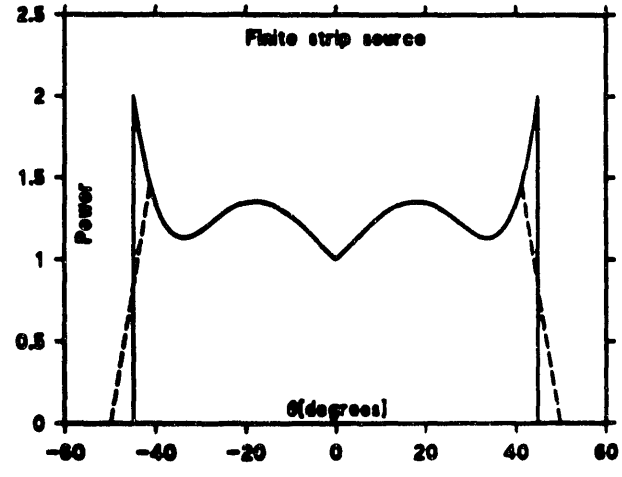

Figure 9: Angular power distribution corresponding to the irradiance distribution shown in Fig.8. Broken line refers to a truncated device.

Although the desired irradiance in this example is significantly different from the constant irradiance treated in the example before, the reflector shape again is astonishingly close to a V-trough and the reflector of the previous example. The subtle difference between the reflector shape of this example and a true V-trough are visualized in Fig.12 and Fig.13 where we plot the slope of our reflector and the distance to a true V-trough. Most structure is

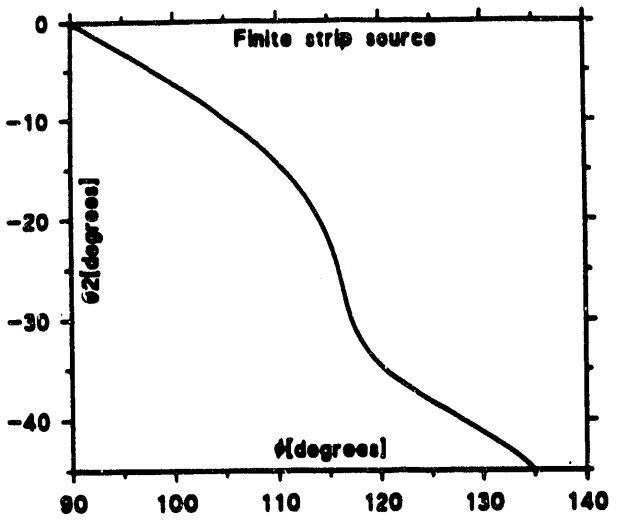

Figure 10: Acceptance angle function corresponding to the desired irradiance distribution plotted in Fig.8.

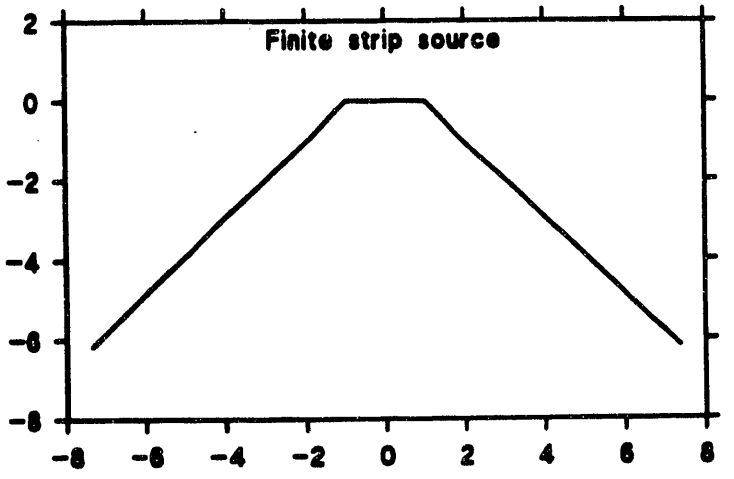

Figure 11: The reflector profile which produces the desired irradiance shown in Fig.8 on a distant plane from a finite one-sided lambertian strip source of width two units. Note that there is only a CHC-type solution and it is truncated.

confined to the region adjacent to the source. The fact that subtle variations in reflector shape have marked effects on the power and irradiance distribution of the device can be attributed to the large incidence angle with which the edge rays strike the outer parts of the reflector.

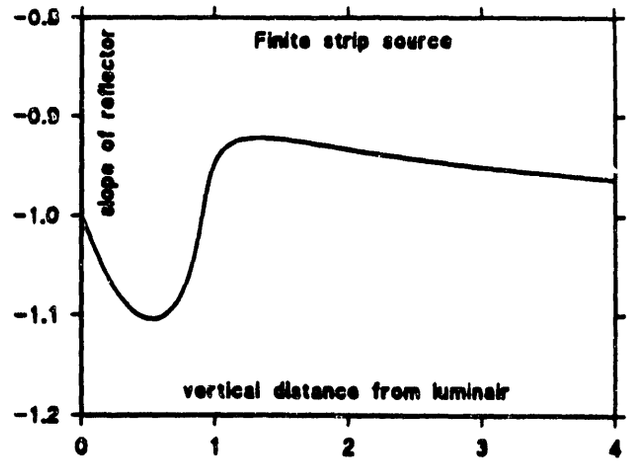

Figure 12: Slope of the reflector as a function of vertical distance from the source.

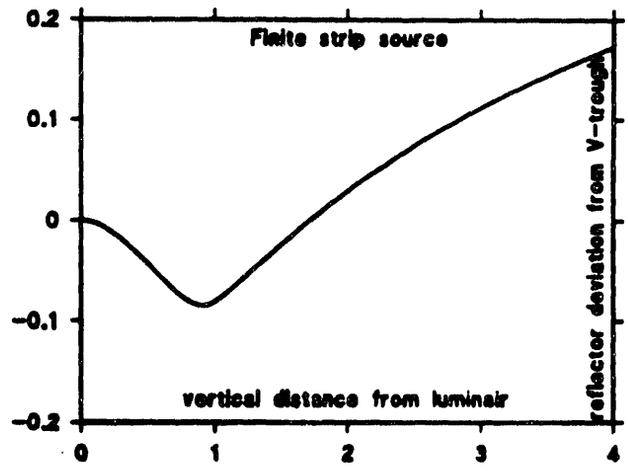

Figure 13: Deviation of the reflector depicted in Fig.11 from a true V-trough.

As mentioned before, in general the reflector is of infinite size. Truncation alters, however, only the distribution in the outer parts. To illustrate the effects of truncation for the reflector of this example, we plot in Fig.14 the angle up to which the truncated device matches the desired power distribution, as a function of the vertical length of the reflector. Thus for example the truncated device shown in Fig 11 has the irradiance distribution and power distribution shown in broken line in Fig.8 and Fig.9. Note that the reflector truncated to a vertical length of 3 times the source width covers more than 5/6 of the angular range.

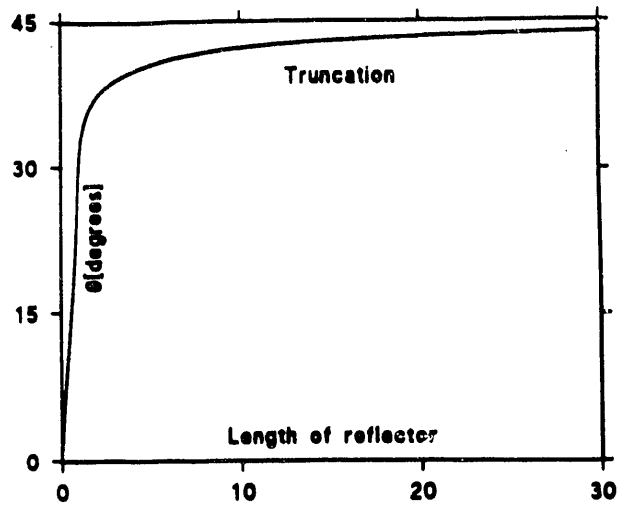

Figure 14: The effect of truncation is indicated by the angle up to which the truncated device matches the desired power distribution, plotted as a function of the vertical length of the reflector.

3.4 What power distributions can be produced?

First, evidently the total desired power must match the total power emitted by the source. Here we investigate what other conditions $P^o(\theta)$ must meet.

The intensity desired at the center cannot be less than that produced by the source alone: $P^o(0) \geq R_o$ because the reflector can only add radiation. If the intensity and irradiance desired at the center, at $\theta = 0$ is larger than that produced by the source alone, then the reflector shape starts with ϕ increasing at constant $\theta = 0$, thus $\alpha = \phi/2$. Equation(2) can then be directly integrated. The shape of the first section is a parabola with axis perpendicular to the reference plane.

$$R(\phi) = R_o \frac{1}{1 + \cos(\phi)} \tag{21}$$

The parabolic section extends until Eq.(13) is met. For larger angles a sudden, step-increase of the power density, proceeding away from the center, can be produced by adding parabolic sections. Note that the reflector remains continuous and smooth (differentiable).

The strongest decrease, that can be produced at any point, is that produced by truncation. As the incidence angle of the edge ray increases, this strongest decrease becomes more marked. Thus a step-decrease cannot be produced except in the limit at θ^{max}, where the reflector extends to infinity. Algebraically this is expressed by the condition that ϕ, as given in Eq.(19), is monotonous so that from each point on the reflector, the opposite edge of the source can be seen.

If the initial part of the reflector starts as a parabolic section then the view factor of the source is larger and thus the maximum angle up to which constant illumination can be achieved is

correspondingly smaller. However, the effects of truncation are the same: The radiation will be strictly as desired in the central part of the range and some of the radiation will be spread beyond the maximum angle of the untruncated device. The "smallest" possible maximum angle is $\theta^{max} = 0$. In this case both sides of the reflector are sections of a parabola with vertical axis and the opposite source edge as a focus.

Constant illumination over angles larger then $\pi/2$ cannot be achieved with a flat 2D source because this would imply that at $\theta = 0$ the combination of source and reflector radiates less than the source alone.

4 Conclusions

The classical nonimaging reflector shapes can be viewed as functions of an acceptance angle which is constant along the reflector profile. A variety of problems, however, require variable acceptance angles. In these cases the reflector profile is a functional of the acceptance angle function or the function describing the desired power density distribution. For the calculation of the reflector based on the variable magnification there are in general two different types of solution, depending on whether the meridional magnification is positive or negative: a CEC-type, characterized by positive magnification in which the reflection of the source appears on the side opposite to the observer, and a CHC-type, of negative magnification, where the reflection is on the same side. The CEC-type reflector is finite and always ends with a vertical tangent, while the CHC-type solution is infinite and approaches a constant slope. The end point of the CEC-type solution and the asymptotic slope of the CHC-type reflect the conservation of total radiant power.

For a finite size source, we have shown how to calculate a CHC-type reflector profile touching the source. For a flat source the solution can be given in closed form. The method presented here does not entail in any way an optimisation procedure. It yields the reflector profile which produces a desired irradiance distribution from a given source by straightforward calculation based on first principles.

The desired irradiance or power distribution in the approach presented in this paper for the finite sources was produced by designing the reflection of the source immediately joining the source. The price of this choice is that only CHC-type reflectors and no CEC-type reflectors result. But the benefit is that, unlike the classical reflectors designed in the small source approximation, the reflectors described here for finite sources can be readily adapted to the reverse problem, namely as ideal nonimaging concentrators for given radiation. For example the ideal secondary concentrator for a Fresnel primary described in a recent publication [9], can now be given in a closed form by equations.(19) and (20) with $P^o(\theta) = |1/\cos(\theta) - \alpha_s)|$ where $2\alpha_s$ is the angle subtended by the sun ($\simeq 0.01$ radian).

Subtle differences in the reflector shape have strong effects on the produced power and irradiance distribution. Therefore a high precision is needed for the manufacturing of such reflectors.

List of Figures

1 The reflector profile is represented in polar coordinates $R(\phi)$ with the source at the origin. The reflected radiation has an angle θ with the optical axis y and α with the normal to the reflector. 3

2 Acceptance angle function which produces a constant irradiance on a distant plane from a narrow one-sided lambertian strip source (2D) , a=1. 6

3 The reflector profile which produces a constant irradiance on a distant plane from a one-sided lambertian strip source (2D) at the origin, $R(\phi = \pi/2) = 1$, $a = 1$. CEC (inner curve) and CHC-type solutions (outer truncated curve) are shown. 6

4 The reflector is designed to produce a reflected image adjacent to the source. The combined intensity radiated in the direction $-\theta$ is determined by the separation of the two edge rays: $R \sin 2\alpha$. 7

5 Acceptance angle function which produces a constant irradiance on a distant plane from a finite one-sided lambertian strip source. There is only a CHC-type solution. 9

6 The reflector profile which produces a constant irradiance on a distant plane from a finite one-sided lambertian strip source of width two units. Note that there is only a CHC-type solution and it is truncated. 9

7 Deviation of the reflector depicted in Fig.6 from a true V-trough. 9

8 Desired irradiance distribution on a distant plane perpendicular to the optical plane divided by the irradiance produced along the axis by the source alone. Broken line shows the irradiance of a truncated device 10

9 Angular power distribution corresponding to the irradiance distribution shown in Fig.8. Broken line refers to a truncated device. 10

10 Acceptance angle function corresponding to the desired irradiance distribution plotted in Fig.8. 11

11 The reflector profile which produces the desired irradiance shown in Fig.8 on a distant plane from a finite one-sided lambertian strip source of width two units. Note that there is only a CHC-type solution and it is truncated. . . . 11

12 Slope of the reflector as a function of vertical distance from the source. . . . 12

13 Deviation of the reflector depicted in Fig.11 from a true V-trough. 12

14 The effect of truncation is indicated by the angle up to which the truncated device matches the desired power distribution, plotted as a function of the vertical length of the reflector. 13

References

[1] R. Winston. Nonimaging optics. *Scientific American*, 264(3):76–81, March 1991.

[2] W. T. Welford and R. Winston. *High Collection Non-Imaging Optics*, chapter 3. Academic Press, New York, 1989.

[3] W. H. McAdams. *Heat Transmission*, chapter 4 by Hoyt C. Hottel. McGraw-Hill, New York, third edition, 1954.

[4] William B. Elmer. *The Optical Design of Reflectors*, chapter 4.4. John Wiley & Sons, New York, second edition, 1980.

[5] R. Winston. Cone collectors for finite sources. *Appl. Opt.*, 17:688–1689, 1978.

[6] M. Collares-Pereira, A. Rabl, and R. Winston. Lens-mirror combinations with maximal concentration. *Applied Optics*, 16(10):2677–2683, 1977.

[7] Ari Rabl and Roland Winston. Ideal concentrators for finite sources and restricted exit angles. *Applied Optics*, 15(11):2880–2883, 1976.

[8] J. M. Gordon and A. Rabl. Nonimaging CPC-type reflectors with variable extreme directions. *to be published in Appl. Optics*, 1993.

[9] J. M. Gordon and H. Ries. Tailored edge-ray concentrators (TERC's) as ideal second stages for Fresnel reflectors. *to be published in Appl. Optics*, 1993.

[10] R. Winston. Nonimaging optics: optical design at the thermodynamic limit. In Roland Winston and Robert L. Holman, editors, *Nonimaging Optics: Maximum Efficiency Light Transfer*, pages 2–6. Proc. SPIE 1528, July 1991.

B. RECENT PROGRESS IN PHYSICAL OPTICS

Slightly over one year ago we started a collaboration with the Univesity of California/Berkeley group of Robert Littlejohn and Allan Kaufman with the goal of developing a theory of radiance applicable to nonimaging optical systems. Our results are presented in a paper submitted to J.O.S.A. A which is reproduced here in its entirety. [The authors are Robert Littlejohn and Roland Winston.]

1. Introduction

The relationship between classical radiometry and the electromagnetic theory of light is now a well developed subject. Reviews have been give by Wolf[1] and Apresyan and Kravtsov.[2] A basic problem in this area, first posed by Walther,[3,4] is how to define the radiance or brightness in terms of the electromagnetic fields and their statistical properties. It is now understood that there are many possible definitions, none of which has exactly all the properties expected of the radiance in classical radiometry, but all of which acquire these properties in the limits of short wavelength and sufficient incoherence. The various possible definitions of the brightness are quite similar to the various phase space distributions used in quantum mechanics, among which the Wigner function[5] is one of the better known.

This paper is concerned with the deviations from classical radiometry, i.e., effects of physical optics which occur when the wavelength is not negligable and effects which occur only for substantially coherent light. We are particularly interested in the evolution of various distribution functions along rays, since in classical radiometry this evolution is null (the radiance is conserved along rays). In this paper we are concerned only with propagation in a homogeneous medium. This is the simplest case in which to make a systematic exploration of the corrections to classical radiometry, and we have chosen to work on it first. Since there are many ways to define a distribution function representing the radiance, our examination of the corrections to classical radiometry is different for the different functions, and we are able to make comparisons.

In Sec. 2 we lay out our physical assumptions and the mathematical formalism we will use to describe them. The material in this section is standard in the literature on diffraction and radiometry and coherence. In Sec. 3 we discuss Walther's first proposed definition of radiance,[3] which is essentially the Wigner function,[5] and express its properties in terms of the Weyl correspondence[6] and the product formulas of Moyal.[7] This material is standard in the literature on the Wigner function. Then we present some new results, namely infinite series representations for the evolution of the Wigner function along rays and for the components of the energy flux in terms of the Wigner function. In Sec. 4 we discuss Walther's second definition

of radiance,[4] which we study in both its real and complex versions. We develop various identities connecting the real and imaginary parts of Walther's complex function, including infinite series. We also develop both integral and infinite series formulas connecting Walther's function with the Wigner function, and we illustrate these series explicitly for a Gaussian-Schell model. Then we develop infinite series representations for the evolution of Walther's function along rays, in both its real and complex forms. Finally, in Sec. 5 we present a new distribution function which has the property that it is exactly conserved along rays.

Our results allow us to draw conclusions about which function is better conserved along rays, by estimating the order of magnitude of the first correction term. In the case of Walther's complex function, such an estimate was made by Walther himself,[3] and our result, obtained by different means, agrees with his. This same estimate has also been examined by Jannson.[8] In the case of Walter's real function, we find that the conservation of brightness along rays is better than in the case of Walther's complex function, and we provide the appropriate estimates. As for the Wigner function, we find that it is even better conserved along rays than Walther's real function, especially for paraxial rays. Finally, our new distribution function introduced in Sec. 5 is exactly conserved along rays.

2. The Physical Model and Its Mathematical Formulation

In this section we describe the physical model we will adopt for our study of the propagation of optical radiation, and the mathematical formalism we will use to represent it. The same physical model is common in work on diffraction, so we will just quickly summarize our assumptions. The mathematics we use is basically the Hilbert space formalism of quantum mechanics, which has also been used by other authors in applications to optics.[9] One reason for using this formalism is the strong analogy which exists between the correlation functions in optics and the properties of the density operator in quantum mechanics.

We represent the optical wave field by a scalar function $\psi(\mathbf{r})$, which can be loosely identified with one component of the electric field. We consider only monochromatic radiation of frequency ω, and we suppress the factor $e^{-i\omega t}$ in ψ. The

field ψ is complex. The radiation is generated in the region $z < 0$ by currents and charges which need not be specified, and propagates into the source-free region $z > 0$ which is assumed to be homogeneous and isotropic with a constant, real index of refraction n. Thus, in the region $z > 0$, ψ satisfies the Helmholtz equation,

$$\nabla^2\psi + k_0^2\psi = 0, \tag{2.1}$$

where $k_0 = n\omega/c$. We assume that in the region $z > 0$ the wave field consists only of waves propagating or damping in the direction of increasing z. We take the energy density to be $n^2|\psi|^2$, and the energy flux to be

$$\mathbf{J} = \frac{nc}{k_0}\,\mathrm{Im}\,\psi^*\nabla\psi. \tag{2.2}$$

The modifications required to accout for the true vectorial nature of light are straightforward, since the medium in the region $z > 0$ is uniform. All the assumptions which go into this model are standard in studies of diffraction and in the literature on radiometry and coherence.[1-4,8-11]

We will write $\mathbf{r}_\perp = (x, y)$ and $\mathbf{r} = (x, y, z) = (\mathbf{r}_\perp, z)$. We will usually regard z as a parameter and think of $\psi(\mathbf{r}_\perp, z)$ as a wave function in the Hilbert space of wave functions defined over the xy-plane. We use the Dirac notation to write $|\psi\rangle$ or $|\psi(z)\rangle$ for the state of the optical field in a given plane $z =$ const., regarded as an abstract vector in the Hilbert space, and we write

$$\psi(\mathbf{r}_\perp, z) = \langle \mathbf{r}_\perp|\psi(z)\rangle \tag{2.3}$$

to show the relation between the abstract Hilbert space vector $|\psi(z)\rangle$ and the usual wavefunction $\psi(\mathbf{r}_\perp, z)$ (which is the $\mathbf{r}_\perp$-representation of that abstract vector). Similarly, we introduce the $\mathbf{k}_\perp$-representation by the Fourier transforms,

$$\tilde{\psi}(\mathbf{k}_\perp, z) = \int \frac{d^2\mathbf{r}_\perp}{2\pi}\, e^{-i\mathbf{k}_\perp\cdot\mathbf{r}_\perp}\, \psi(\mathbf{r}_\perp, z), \tag{2.4}$$

$$\psi(\mathbf{r}_\perp, z) = \int \frac{d^2\mathbf{k}_\perp}{2\pi}\, e^{+i\mathbf{k}_\perp\cdot\mathbf{r}_\perp}\, \tilde{\psi}(\mathbf{k}_\perp, z), \tag{2.5}$$

and we write

$$\tilde{\psi}(\mathbf{k}_\perp, z) = \langle \mathbf{k}_\perp|\psi(z)\rangle. \tag{2.6}$$

In these equations and below, we use tildes to represent quantities referred to the $\mathbf{k}_\perp$-representation, and we set $\mathbf{k}_\perp = (k_x, k_y)$. The normalization conventions used in Eqs. (2.4) and (2.5) make the transformation between the $\mathbf{r}_\perp$- and $\mathbf{k}_\perp$-representation unitary, so that

$$\int d^2\mathbf{r}_\perp \, |\psi(\mathbf{r}_\perp, z)|^2 = \int d^2\mathbf{k}_\perp \, |\tilde{\psi}(\mathbf{k}_\perp, z)|^2. \tag{2.7}$$

To find the optical field on some plane $z = \text{const.} > 0$, given the optical field at $z = 0$, we solve the Helmholtz equation in the $\mathbf{k}_\perp$-representation. The solution is

$$\tilde{\psi}(\mathbf{k}_\perp, z) = e^{ik_z z} \, \tilde{\psi}_0(\mathbf{k}_\perp), \tag{2.8}$$

where we write $\tilde{\psi}_0(\mathbf{k}_\perp)$ for $\tilde{\psi}(\mathbf{k}_\perp, 0)$, and where

$$k_z = \begin{cases} \sqrt{k_0^2 - k_\perp^2}, & \text{if } k_\perp \leq k_0, \\ i\sqrt{k_\perp^2 - k_0^2}, & \text{if } k_\perp \geq k_0. \end{cases} \tag{2.9}$$

Here and throughout this paper it will be necessary to regard k_z, not as an independent variable like k_x and k_y, but as a function of $\mathbf{k}_\perp$. Exceptions to this rule will be noted explicitly. The waves for which $k_\perp < k_0$ are travelling waves, and those for which $k_\perp > k_0$ are evanescent waves. Now we combine Eqs. (2.5) and (2.8) to obtain,

$$\psi(\mathbf{r}_\perp, z) = \int \frac{d^2\mathbf{k}_\perp}{2\pi} \, e^{i\mathbf{k}\cdot\mathbf{r}} \, \tilde{\psi}_0(\mathbf{k}_\perp). \tag{2.10}$$

Here we write $\mathbf{k} = (\mathbf{k}_\perp, k_z)$, so that

$$\mathbf{k}\cdot\mathbf{r} = \mathbf{k}_\perp \cdot \mathbf{r}_\perp + k_z z \tag{2.11}$$

with k_z regarded as a function of $\mathbf{k}_\perp$. Next we use Eq. (2.4) evaluated at $z = 0$ to express the result completely in the $\mathbf{r}_\perp$-representation. We find

$$\psi(\mathbf{r}_\perp, z) = \int d^2\mathbf{r}'_\perp \, K(\mathbf{r}_\perp - \mathbf{r}'_\perp, z)\psi_0(\mathbf{r}'_\perp), \tag{2.12}$$

where we write $\psi_0(\mathbf{r}_\perp)$ for $\psi(\mathbf{r}_\perp, 0)$, and where

$$K(\mathbf{r}_\perp, z) = \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} \, e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{2.13}$$

Equations (2.8) and (2.12) express a kind of "z-evolution," which is specified by a certain operator $\hat{K}(z)$, parameterized by z, which we call the "z-propagator." The matrix elements of this propagator in both the $\mathbf{r}_\perp$-representation and the $\mathbf{k}_\perp$-representation can be read off from these equations; we have

$$\langle \mathbf{r}_\perp | \hat{K}(z) | \mathbf{r}'_\perp \rangle = K(\mathbf{r}_\perp - \mathbf{r}'_\perp, z) \tag{2.14}$$

and

$$\langle \mathbf{k}_\perp | \hat{K}(z) | \mathbf{k}'_\perp \rangle = e^{ik_z z} \delta(\mathbf{k}_\perp - \mathbf{k}'_\perp). \tag{2.15}$$

Thus, we can write Eqs. (2.8) and (2.12) in the form,

$$|\psi(z)\rangle = \hat{K}(z) |\psi(0)\rangle. \tag{2.16}$$

The matrix elements of $\hat{K}(z)$ in the $\mathbf{r}_\perp$-representation can be expressed in terms of the free-space Green's function G for the Helmholtz equation,

$$G(\mathbf{r}) = -\frac{e^{ik_0 r}}{4\pi r} = -\frac{i}{2} \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{k_z} = -\int \frac{d^3\mathbf{k}}{(2\pi)^3} \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{k^2 - k_0^2}. \tag{2.17}$$

In the second of these integrals, k_z is an independent variable of integration, not a function of $\mathbf{k}_\perp$. The Green's function G satisfies

$$\nabla^2 G + k_0^2 G = \delta(\mathbf{r}). \tag{2.18}$$

To obtain the relation connecting G and K, we differentiate Eq. (2.17) with respect to z and use Eq. (2.13) to obtain

$$K(\mathbf{r}_\perp, z) = 2\frac{\partial G(\mathbf{r})}{\partial z} = \frac{z e^{ik_0 r}}{2\pi} \left(\frac{1}{r^3} - \frac{ik_0}{r^2} \right). \tag{2.19}$$

The operator $\hat{K}(z)$ is not unitary, due to the evanescent waves, for we have

$$\begin{aligned} \frac{d}{dz} \int d^2\mathbf{r}_\perp \, |\psi(\mathbf{r}_\perp, z)|^2 &= \frac{d}{dz} \int d^2\mathbf{k}_\perp \, |\psi(\mathbf{k}_\perp, z)|^2 \\ &= \frac{d}{dz} \int\limits_{k_\perp > k_0} d^2\mathbf{k}_\perp \exp\left(-2z\sqrt{k_\perp^2 - k_0^2}\right) |\psi_0(\mathbf{k}_\perp)|^2 \leq 0. \end{aligned} \tag{2.20}$$

The final integral is carried out over evanescent waves only. If we should have a wave field with negligable contribution from evanescent waves, then the norm of the

wave function ψ in the sense of Eq. (2.7) or (2.20) is conserved by the z-evolution, and $\hat{K}(z)$ behaves as if it were unitary.

A differential equation like the Schrödinger equation can be written down for the z-evolution. It is

$$i\frac{\partial}{\partial z}|\psi(z)\rangle = \hat{H}|\psi(z)\rangle, \tag{2.21}$$

where the "Hamiltonian" operator $\hat{H}$ is given by its (diagonal) matrix elements in the $\mathbf{k}_\perp$-representation,

$$\langle \mathbf{k}_\perp|\hat{H}|\mathbf{k}'_\perp\rangle = -k_z\,\delta(\mathbf{k}_\perp - \mathbf{k}'_\perp), \tag{2.22}$$

or in the $\mathbf{r}_\perp$-representation,

$$\langle \mathbf{r}_\perp|\hat{H}|\mathbf{r}'_\perp\rangle = -\frac{e^{ik_0 r}}{2\pi}\left(\frac{1}{\rho^3} - \frac{ik_0}{\rho^2}\right), \tag{2.23}$$

where $\rho = |\mathbf{r}_\perp - \mathbf{r}'_\perp|$. These equations follow from

$$\hat{K}(z) = e^{-i\hat{H}z}. \tag{2.24}$$

The Hamiltonian $\hat{H}$ is not Hermitian for the same reason that $\hat{K}$ is not unitary (the evanescent waves). It is sometimes suggestive to write $\hat{H}$ in terms of the operator $\hat{\mathbf{k}}_\perp$ (which is $-i\nabla_\perp$ in the $\mathbf{r}_\perp$-representation, or multiplication by $\mathbf{k}_\perp$ in the $\mathbf{k}_\perp$-representation). That is,

$$\hat{H} = -\sqrt{k_0^2 - \hat{\mathbf{k}}_\perp^2} = -\sqrt{k_0^2 + \nabla_\perp^2}, \tag{2.25}$$

where the square root is defined as in Eq. (2.9).

Now we introduce the angular spectrum. We write $\mathbf{s} = (s_x, s_y, s_z) = \mathbf{k}/k_0$, so that $\mathbf{s}$ is a unit vector. Then we write

$$\begin{aligned} s_x &= \sin\theta\cos\phi, \\ s_y &= \sin\theta\sin\phi, \\ s_z &= \cos\theta, \end{aligned} \tag{2.26}$$

so that

$$d^2\mathbf{k}_\perp = k_0^2\,d^2\mathbf{s}_\perp = s_z k_0^2\,d\Omega, \tag{2.27}$$

where $d\Omega$ is the element of solid angle. When $\mathbf{k}_\perp$ lies outside the circle $k_\perp = k_0$ (or $\mathbf{s}_\perp$ lies outside the unit circle), then the angle θ (like $s_z = k_z/k_0$) takes on complex values. Now we transform Eq. (2.10) into an angular integral,

$$\psi(\mathbf{r}_\perp, z) = \int d\Omega\, \tilde{a}(\mathbf{s}) e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{2.28}$$

where

$$\tilde{a}(\mathbf{s}) = \frac{k_0^2}{2\pi} s_z \tilde{\psi}_0(\mathbf{k}_\perp). \tag{2.29}$$

The quantity $\tilde{a}(\mathbf{s})$ is the angular spectrum. It actually depends only on $\mathbf{s}_\perp$, although it is convenient to imagine it as defined over a hemisphere on which the unit vector $\mathbf{s}$ lies. The hemisphere should properly be extended so as to include complex angles, on account of the evanescent waves.

Next we introduce a statistical ensemble of wave fields $\{\psi_\alpha(\mathbf{r})\}$, $\alpha = 1, \ldots, N$, with corresponding weights c_α satisfying $c_\alpha \geq 0$ and

$$\sum_{\alpha=0}^{N} c_\alpha = 1. \tag{2.30}$$

In some cases, we may promote α into a continuous index and make appropriate changes to our formulas. The statistical averages we will be interested in can be expressed in terms of the mutual intensity,

$$\Gamma(\mathbf{r}, \mathbf{r}') = \sum_\alpha c_\alpha \psi_\alpha(\mathbf{r}) \psi_\alpha^*(\mathbf{r}') = \overline{\psi(\mathbf{r})\psi^*(\mathbf{r}')}, \tag{2.31}$$

where the overbar indicates the statistical average. Usually we will be interested in $\Gamma(\mathbf{r}, \mathbf{r}')$ only when $z = z'$; in this case we will write $\Gamma(\mathbf{r}_\perp, \mathbf{r}'_\perp; z)$, which is the $\mathbf{r}_\perp$-space matrix element of an operator $\hat{\Gamma}(z)$ (the density operator),

$$\Gamma(\mathbf{r}_\perp, \mathbf{r}'_\perp; z) = \langle \mathbf{r}_\perp | \hat{\Gamma}(z) | \mathbf{r}'_\perp \rangle = \overline{\psi(\mathbf{r}_\perp, z)\psi^*(\mathbf{r}'_\perp, z)}, \tag{2.32}$$

where

$$\hat{\Gamma}(z) = \sum_\alpha |\psi_\alpha(z)\rangle c_\alpha \langle \psi_\alpha(z)|. \tag{2.33}$$

We will also write

$$\tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}'_\perp; z) = \langle \mathbf{k}_\perp | \hat{\Gamma}(z) | \mathbf{k}'_\perp \rangle \tag{2.34}$$

for the $\mathbf{k}_\perp$-space matrix elements of $\hat{\Gamma}(z)$ (with a tilde to indicate the $\mathbf{k}_\perp$-representation).

The z-evolution of $\hat{\Gamma}(z)$ is straightforward. The basic formula is

$$\hat{\Gamma}(z) = \hat{K}(z)\hat{\Gamma}(0)\hat{K}(z)^\dagger, \tag{2.35}$$

which is especially simple in the $\mathbf{k}_\perp$-representation:

$$\tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}'_\perp; z) = e^{[i(k_z - k_z'^*)z]}\, \tilde{\Gamma}_0(\mathbf{k}_\perp, \mathbf{k}'_\perp), \tag{2.36}$$

where k'_z is given by a primed version of Eq. (2.9) and where we define $\tilde{\Gamma}_0(\mathbf{k}_\perp, \mathbf{k}'_\perp) = \tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}'_\perp; 0)$. A differential equation for the evolution of $\hat{\Gamma}(z)$ follows by combining Eqs. (2.24) and (2.30) and differentiating with respect to z. The result is

$$i\frac{d\hat{\Gamma}(z)}{dz} = \hat{H}\hat{\Gamma}(z) - \hat{\Gamma}(z)\hat{H}^\dagger. \tag{2.37}$$

If the statistical ensemble does not contain any evanescent waves (a condition which is independent of z), then the right hand side of this equation can be replaced by the commutator $[\hat{\Gamma}(z), \hat{H}]$.

3. The Wigner Function in Radiometry

In this section we consider the Wigner function as a candidate distribution function in terms of which the brightness or radiance can be defined. The Wigner function has certain advantages and disadvantages in comparison to other candidate functions in this role. The principal disadvantage seems to be that the exact expression for the z-component of the energy flux is not identical to the formula expected on the basis of classical radiometry, although the two formulas do agree in the limits of short wavelengths and sufficient incoherence. This disadvantage is offset by a number of advantages. First, in the paraxial approximation (or any approximation which leads to quadratic phase factors for the $\mathbf{r}_\perp$-space kernel of the z-propagator), the Wigner function is exactly conserved along rays, not only in a homogeneous medium (which is our main interest in this paper), but also when the rays pass through lenses, etc. This exact conservation is independent of the degree of coherence, and applies even to completely coherent light. Second, the Wigner

function has elegant analytical properties, which allow us to provide explicit forms for the corrections to classical radiometry, out to all orders in the appropriate small parameters. Indeed, we have found it convenient, in exploring the properties of other distribution functions, to express them first in terms of the Wigner function, so we could invoke the analytical properties of the latter. Third, we find that the Wigner function is conserved along rays to a higher degree of approximation than other distribution functions, even for rays which are substantially off-axis. We begin this section with a summary of the usual properties of the Weyl transform and the Wigner function, transcribed to the optical context in which we are interested. Reviews and other articles of interest on the Wigner function and the Weyl correspondence include Refs. 12–16.

Let $\hat{A}$ be any operator which acts on the Hilbert space of wave functions defined over the xy-plane, and let $A_w(\mathbf{r}_\perp, \mathbf{k}_\perp)$ be its Weyl transform, which is defined by

$$\begin{aligned} A_w(\mathbf{r}_\perp, \mathbf{k}_\perp) &= \int d^2\mathbf{a}_\perp\, e^{-i\mathbf{k}_\perp\cdot\mathbf{a}_\perp}\, \langle \mathbf{r}_\perp + \tfrac{1}{2}\mathbf{a}_\perp|\hat{A}|\mathbf{r}_\perp - \tfrac{1}{2}\mathbf{a}_\perp\rangle \\ &= \int d^2\mathbf{q}_\perp\, e^{+i\mathbf{q}_\perp\cdot\mathbf{r}_\perp}\, \langle \mathbf{k}_\perp + \tfrac{1}{2}\mathbf{q}_\perp|\hat{A}|\mathbf{k}_\perp - \tfrac{1}{2}\mathbf{q}_\perp\rangle. \end{aligned} \tag{3.1}$$

The Weyl transform has the following properties. First, if A_w is the Weyl transform of operator $\hat{A}$, then the Weyl transform of operator $\hat{A}^\dagger$ is A_w^*. In particular, the Weyl transform of a Hermitian operator is real. Next, if $\hat{A}$ and $\hat{B}$ are two operators and A_w and B_w the corresponding Weyl transforms, then

$$\mathrm{Tr}(\hat{A}^\dagger\hat{B}) = \int \frac{d^2\mathbf{r}_\perp\, d^2\mathbf{k}_\perp}{(2\pi)^2}\, A_w(\mathbf{r}_\perp, \mathbf{k}_\perp)^* B_w(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{3.2}$$

In particular, since the Weyl transform of the identity operator is unity, we have

$$\mathrm{Tr}(\hat{A}) = \int \frac{d^2\mathbf{r}_\perp\, d^2\mathbf{k}_\perp}{(2\pi)^2}\, A_w(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{3.3}$$

The third property of the Weyl transform is given by the Moyal formula. Let $\hat{A}$, $\hat{B}$, $\hat{C}$ be operators with Weyl transforms A_w, B_w, C_w, and let $\hat{C} = \hat{A}\hat{B}$. Then C_w is given in terms of A_w and B_w by

$$C_w(\mathbf{r}_\perp, \mathbf{k}_\perp) = A_w(\mathbf{r}_\perp, \mathbf{k}_\perp) \exp\Big[\frac{i}{2}\Big(\frac{\overleftarrow{\partial}}{\partial\mathbf{r}_\perp}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{k}_\perp} - \frac{\overleftarrow{\partial}}{\partial\mathbf{k}_\perp}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{r}_\perp}\Big)\Big] B_w(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{3.4}$$

In this expression, the arrows over the partial derivatives indicate the direction in which the derivatives act, i.e., those with a left arrow act on A_w, and those with a right arrow act on B_w. The exponential in this expression can be expanded out, and the first few terms give

$$C_w = A_w B_w + \frac{i}{2}\{A_w, B_w\} + \ldots, \tag{3.5}$$

where the curly bracket represents the Poisson bracket in the variables $\mathbf{r}_\perp$, $\mathbf{k}_\perp$,

$$\{A_w, B_w\} = \frac{\partial A_w}{\partial \mathbf{r}_\perp} \cdot \frac{\partial B_w}{\partial \mathbf{k}_\perp} - \frac{\partial A_w}{\partial \mathbf{k}_\perp} \cdot \frac{\partial B_w}{\partial \mathbf{r}_\perp}. \tag{3.6}$$

A variation on the Moyal formula is obtained if we let $\hat{C} = \hat{A}\hat{B} - \hat{B}\hat{A} = [\hat{A}, \hat{B}]$. Then the Weyl transform C_w can be written,

$$C_w = 2iA_w \sin\Big[\frac{1}{2}\Big(\frac{\overleftarrow{\partial}}{\partial \mathbf{r}_\perp} \cdot \frac{\overrightarrow{\partial}}{\partial \mathbf{k}_\perp} - \frac{\overleftarrow{\partial}}{\partial \mathbf{k}_\perp} \cdot \frac{\overrightarrow{\partial}}{\partial \mathbf{r}_\perp}\Big)\Big] B_w. \tag{3.7}$$

We will find the Moyal formula useful in developing the corrections to classical radiometry.

We now tabulate some Weyl transforms of various operators which will be of use to us later. If $\hat{A}$ is an operator, we will use a two-sided arrow to show the correspondence with its Weyl transform, a function of $(\mathbf{r}_\perp, \mathbf{k}_\perp)$. First, the Weyl transform of the identity operator, denoted 1, is unity:

$$1 \longleftrightarrow 1. \tag{3.8}$$

Next, the operators $\hat{\mathbf{r}}_\perp$, $\hat{\mathbf{k}}_\perp$ are defined respectively by multiplication by $\mathbf{r}_\perp$ and $-i\nabla_\perp$ in the $\mathbf{r}_\perp$-representation, or $+i\partial/\partial \mathbf{k}_\perp$ and multiplication by $\mathbf{k}_\perp$ in the $\mathbf{k}_\perp$-representation. The Weyl transforms of these operators are given by

$$\hat{\mathbf{r}}_\perp \longleftrightarrow \mathbf{r}_\perp, \qquad \hat{\mathbf{k}}_\perp \longleftrightarrow \mathbf{k}_\perp. \tag{3.9}$$

The Moyal formula can be used to compute the Weyl transforms of higher order polynomials in $\hat{\mathbf{r}}_\perp$, $\hat{\mathbf{k}}_\perp$. More generally, if f and g are any two functions, then we have

$$f(\hat{\mathbf{r}}_\perp) \longleftrightarrow f(\mathbf{r}_\perp), \qquad g(\hat{\mathbf{k}}_\perp) \longleftrightarrow g(\mathbf{k}_\perp). \tag{3.10}$$

Next, we have the Weyl transforms of two projection operators,

$$|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}| \longleftrightarrow \delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0}), \tag{3.11}$$

$$|\mathbf{k}_{\perp 0}\rangle\langle\mathbf{k}_{\perp 0}| \longleftrightarrow \delta(\mathbf{k}_\perp - \mathbf{k}_{\perp 0}). \tag{3.12}$$

In these formulas, $\mathbf{r}_{\perp 0}$ and $\mathbf{k}_{\perp 0}$ are the parameters of the projection operators, which are distinguished from the variables $\mathbf{r}_\perp$, $\mathbf{k}_\perp$ upon which the Weyl transforms depend.

Finally, we have the Wigner function itself, which is the Weyl transform of the density operator $\hat{\Gamma}(z)$:

$$\hat{\Gamma}(z) \longleftrightarrow W(\mathbf{r}_\perp, \mathbf{k}_\perp; z). \tag{3.13}$$

This can also be written,

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int d^2\mathbf{a}_\perp\, e^{-i\mathbf{k}_\perp\cdot\mathbf{a}_\perp}\, \Gamma(\mathbf{r}_\perp + \tfrac{1}{2}\mathbf{a}_\perp, \mathbf{r}_\perp - \tfrac{1}{2}\mathbf{a}_\perp; z). \tag{3.14}$$

We regard the Wigner function as a distribution function defined on the 4-dimensional phase space $(\mathbf{r}_\perp, \mathbf{k}_\perp)$, and parameterized by z. By writing the definition in terms of a $\mathbf{q}_\perp$-integral as in Eq. (3.1) and using Eq. (2.36), it is easy to express W in the plane $z = \text{const.} > 0$ in terms of Γ evaluated in the plane $z = 0$. We find

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int d^2\mathbf{q}_\perp\, e^{i[\mathbf{q}_\perp\cdot\mathbf{r}_\perp + (\kappa_+ - \kappa_-^*)z]}\, \tilde{\Gamma}_0(\mathbf{k}_\perp + \tfrac{1}{2}\mathbf{q}_\perp, \mathbf{k}_\perp - \tfrac{1}{2}\mathbf{q}_\perp), \tag{3.15}$$

where

$$\kappa_\pm = \sqrt{k_0^2 - (\mathbf{k}_\perp \pm \tfrac{1}{2}\mathbf{q}_\perp)^2}, \tag{3.16}$$

and with the square root of negative numbers interpreted as in Eq. (2.9). An alternative version of Eq. (3.15) is

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp\, \delta\Big(\mathbf{k}_\perp - \frac{\mathbf{k}'_\perp + \mathbf{k}''_\perp}{2}\Big) e^{i(\mathbf{k}' - \mathbf{k}''^*)\cdot\mathbf{r}}\, \tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp), \tag{3.17}$$

which is stated in a more symmetrical way.

Basic properties of the Wigner function include its marginal distributions and lowest order moments. If the Wigner function is integrated over $\mathbf{k}_\perp$, it gives the average value of the intensity $I(\mathbf{r})$,

$$\int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\, W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \Gamma(\mathbf{r}_\perp, \mathbf{r}_\perp, z) = \overline{|\psi(\mathbf{r}_\perp, z)|^2} = \overline{I(\mathbf{r})}. \tag{3.18}$$

If the Wigner function is integrated over $\mathbf{r}_\perp$, it gives an analogous result in $\mathbf{k}_\perp$-space:

$$\int \frac{d^2\mathbf{r}_\perp}{(2\pi)^2}\, W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}_\perp; z) = \overline{|\tilde{\psi}(\mathbf{k}_\perp, z)|^2}. \tag{3.19}$$

Finally, if the Wigner function is integrated over all of the $(\mathbf{r}_\perp, \mathbf{k}_\perp)$ phase space, it gives the average of the norm of the wave function, in the same sense as in Eqs. (2.7) and (2.9):

$$\int \frac{d^2\mathbf{r}_\perp\, d^2\mathbf{k}_\perp}{(2\pi)^2}\, W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int d^2\mathbf{r}_\perp\, \overline{|\psi(\mathbf{r}_\perp, z)|^2}, \tag{3.20}$$

which is also the trace of the density operator $\hat{\Gamma}(z)$.

In these formulas, the Wigner function would look more like an ordinary probability density on phase space if the factors of 2π were absorbed into the definition of W. If this were done, however, the Wigner function would no longer be the Weyl transform of the density operator. We prefer to retain the latter property.

Now we propose a definition of brightness or radiance $B(\mathbf{r}, \mathbf{s})$ in terms of the Wigner function $W(\mathbf{r}, \mathbf{k}; z)$, in which we use the classical definition of radiance as a guide in order to obtain the correct factors of proportionality. To recall the classical definition, we define $B(\mathbf{r}, \mathbf{s})$ by saying that the energy flux dJ passing through an area element dA at position $\mathbf{r}$ into solid angle $d\Omega$ centered on direction $\mathbf{s}$ is given by

$$dJ = (\mathbf{n} \cdot \mathbf{s}) B(\mathbf{r}, \mathbf{s})\, dA\, d\Omega, \tag{3.21}$$

where $\mathbf{n}$ is the unit normal to dA. On the other hand, the energy density of the radiation at position $\mathbf{r}$ is $n^2|\psi(\mathbf{r})|^2$, and $|\psi(\mathbf{r})|^2$ is given in terms of the Wigner function by Eq. (3.18). Since the energy of a photon is $\hbar\omega$, we can interpret the quantity

$$dN = \frac{n^2}{(2\pi)^2 \hbar\omega} W(\mathbf{r}_\perp, \mathbf{k}_\perp; z)\, d^3\mathbf{r}\, d^2\mathbf{k}_\perp \tag{3.22}$$

as the number of photons in volume element $d^3\mathbf{r}$ centered at $\mathbf{r}$ with wavevectors $\mathbf{k}_\perp$ lying in element $d^2\mathbf{k}_\perp$. Thus, with the multiplicative constants shown in Eq. (3.22), the Wigner function can be interpreted as a photon number density in the 5-dimensional $(\mathbf{r}, \mathbf{k}_\perp)$ phase space. Next, since the photons have velocity $c\mathbf{k}/nk_0$ and energy $\hbar\omega$, the energy flux crossing area element dA lying in the xy-plane in

the given $\mathbf{k}_\perp$-interval is

$$dJ = \frac{ncs_z}{(2\pi)^2} W(\mathbf{r}_\perp, \mathbf{k}_\perp; z)\, dA\, d^2\mathbf{k}_\perp, \tag{3.23}$$

which can be combined with Eqs. (2.27) and (3.21) to obtain the desired formula connecting W and B:

$$B(\mathbf{r}, \mathbf{s}) = cns_z \left(\frac{k_0}{2\pi}\right)^2 W(\mathbf{r}_\perp, \mathbf{k}_\perp; z). \tag{3.24}$$

We will take this formula as our definition of radiance, and investigate to what extent it has the properties expected from classical radiometry. Notice that the factor s_z is not a constant, but depends on $\mathbf{k}_\perp$. We could have absorbed this factor into the definition of W, as other authors have done, but we prefer to leave things as shown so that we can interpret W as a Weyl transform and use the various properties which follow from this fact.

Let us now compute various moments of our radiance function and compare the results with the expectations of classical radiometry. First, we expect the integral of B over all solid angles to be the photon velocity c/n times the average energy density. Indeed, with our definition (3.24) we have

$$\frac{n}{c} \int B(\mathbf{r}, \mathbf{s})\, d\Omega = n^2 \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = n^2 \overline{|\psi(\mathbf{r})|^2}, \tag{3.25}$$

where we use Eqs. (3.14), (2.27) and (3.18). This result is exactly what we expect.

Next, in classical radiometry the average energy flux $\mathbf{J}(\mathbf{r})$ is the integral of $\mathbf{s}B(\mathbf{r}, \mathbf{s})$ over all angles. To see whether this relation is fulfilled by our definition (3.24), we begin with the perpendicular components of $\mathbf{J}$. From Eq. (2.2), appropriately averaged, we have

$$\begin{aligned} \mathbf{J}_\perp(\mathbf{r}) &= \frac{nc}{k_0} \operatorname{Im} \overline{\nabla_\perp \psi(\mathbf{r})\, \psi(\mathbf{r})^*} \\ &= \frac{nc}{k_0} \operatorname{Re} \int \frac{d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp}{(2\pi)^2}\, \mathbf{k}'_\perp\, e^{i(\mathbf{k}' - \mathbf{k}''^*)\cdot\mathbf{r}}\, \tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp), \end{aligned} \tag{3.26}$$

where we use Eq. (2.10). By swapping $\mathbf{k}'_\perp$, $\mathbf{k}''_\perp$ and noting the Hermiticity of $\hat{\Gamma}$,

$$\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp)^* = \tilde{\Gamma}_0(\mathbf{k}''_\perp, \mathbf{k}'_\perp), \tag{3.27}$$

we can symmetrize Eq. (3.26) and remove the Re, obtaining

$$\mathbf{J}_\perp(\mathbf{r}) = \frac{nc}{k_0}\int \frac{d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp}{(2\pi)^2}\left(\frac{\mathbf{k}'_\perp + \mathbf{k}''_\perp}{2}\right) e^{i(\mathbf{k}'-\mathbf{k}''^*)\cdot\mathbf{r}}\,\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp). \tag{3.28}$$

But with the help of Eqs. (3.17) and (3.24), this is easily expressed in terms of the radiance. We find

$$\mathbf{J}_\perp(\mathbf{r}) = \frac{nc}{k_0}\int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\,\mathbf{k}_\perp\, W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int \mathbf{s}_\perp B(\mathbf{r}, \mathbf{s})\, d\Omega. \tag{3.29}$$

Thus we see that the perpendicular components of the energy flux are exactly given by the classical formula.

It is instructive to derive this same result purely by means of the Weyl transform, since the derivation provides an illustration of the application of these properties in a context in which the answer is known. The basic strategy is to express the quantity we wish to evaluate (in this case, the perpendicular components of the energy flux) in terms of a trace of a product of operators, and then to use Eq. (3.2) to compute the trace. In the present case we evaluate $\mathbf{J}_\perp$ at $(\mathbf{r}_{\perp 0}, z)$ and write

$$\begin{aligned}\frac{k_0}{nc}\mathbf{J}_\perp(\mathbf{r}_{\perp 0}, z) &= \mathrm{Im}\,\overline{\nabla_\perp\psi(\mathbf{r}_{\perp 0}, z)\,\psi(\mathbf{r}_{\perp 0}, z)^*} = \mathrm{Re}\,\overline{\langle\mathbf{r}_{\perp 0}|\hat{\mathbf{k}}_\perp|\psi\rangle\langle\psi|\mathbf{r}_{\perp 0}\rangle} \\ &= \mathrm{Re}\langle\mathbf{r}_{\perp 0}|\hat{\mathbf{k}}_\perp\hat{\Gamma}(z)|\mathbf{r}_{\perp 0}\rangle = \mathrm{Re}\,\mathrm{Tr}\Big[|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\,\hat{\mathbf{k}}_\perp\hat{\Gamma}(z)\Big] \\ &= \tfrac{1}{2}\,\mathrm{Tr}\Big[\Big(|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\hat{\mathbf{k}}_\perp + \hat{\mathbf{k}}_\perp|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\Big)\hat{\Gamma}(z)\Big].\end{aligned} \tag{3.30}$$

The operator in parentheses in the final expression is Hermitian and has a real Weyl transform, but the Weyl transforms of its two constituent terms are not real. To compute the Weyl transform of the first term in the parentheses we use Eqs. (3.9), (3.11), and the Moyal formula, Eq. (3.4). The Moyal formula terminates after two terms, and we find

$$|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\hat{\mathbf{k}}_\perp \longleftrightarrow \mathbf{k}_\perp\delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0}) + \frac{i}{2}\nabla_\perp\delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0}). \tag{3.31}$$

The second operator in the parentheses in Eq. (3.30) is the Hermitian conjugate of the first, with complex conjugate Weyl transform. Adding these together, the imaginary terms cancel, and we have

$$\tfrac{1}{2}\Big(|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\hat{\mathbf{k}}_\perp + \hat{\mathbf{k}}_\perp|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\Big) \longleftrightarrow \mathbf{k}_\perp\delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0}). \tag{3.32}$$

Finally, since the Weyl transform of $\hat{\Gamma}(z)$ is just the Wigner function, we can use Eq. (3.2) to obtain

$$\mathbf{J}_\perp(\mathbf{r}_{\perp 0}, z) = \frac{nc}{k_0} \int \frac{d^2\mathbf{r}_\perp\, d^2\mathbf{k}_\perp}{(2\pi)^2}\, \mathbf{k}_\perp\, \delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0}) W(\mathbf{r}_\perp, \mathbf{k}_\perp; z), \tag{3.33}$$

which agrees with our earlier result in Eq. (3.29).

The z-component of the energy flux is more complicated. We begin as in Eqs. (3.26)–(3.28), obtaining

$$\begin{aligned} J_z(\mathbf{r}) &= \frac{nc}{k_0} \operatorname{Im} \overline{\frac{\partial \psi(\mathbf{r})}{\partial z} \psi(\mathbf{r})^*} \\ &= \frac{nc}{k_0} \int \frac{d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp}{(2\pi)^2} \left(\frac{k'_z + k''^*_z}{2} \right) e^{i(\mathbf{k}' - \mathbf{k}''^*)\cdot\mathbf{r}}\, \tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp). \end{aligned} \tag{3.34}$$

On the other hand, if we follow the formulas of classical radiometry, we expect J_z to be given by

$$\begin{aligned} \int s_z\, B(\mathbf{r}, \mathbf{s})\, d\Omega &= \frac{nc}{k_0} \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\, k_z\, W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) \\ &= \frac{nc}{k_0} \int \frac{d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp}{(2\pi)^2} \sqrt{k_0^2 - \left(\frac{\mathbf{k}'_\perp + \mathbf{k}''_\perp}{2} \right)^2}\, e^{i(\mathbf{k}' - \mathbf{k}''^*)\cdot\mathbf{r}}\, \tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp), \end{aligned} \tag{3.35}$$

where we have used Eq. (3.17) in the final step. The expressions in Eqs. (3.34) and (3.35) are not equal, because

$$\frac{1}{2} \left(\sqrt{k_0^2 - k'^2_\perp} + \sqrt{k_0^2 - k''^2_\perp} \right) \neq \sqrt{k_0^2 - \left(\frac{\mathbf{k}'_\perp + \mathbf{k}''_\perp}{2} \right)^2}. \tag{3.36}$$

But if the function $\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp)$ is sharply peaked about $\mathbf{k}'_\perp = \mathbf{k}''_\perp$, i.e., if it has the form

$$\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp) \approx \tilde{F}(\mathbf{k}'_\perp)\delta(\mathbf{k}'_\perp - \mathbf{k}''_\perp), \tag{3.37}$$

where $\tilde{F}$ is a slow function of its argument, and if we ignore evanescent waves so that k''_z is real, then both the expressions in Eqs. (3.34) and (3.35) reduce to

$$J_z(\mathbf{r}) = \frac{nc}{k_0} \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\, k_z\, \tilde{F}(\mathbf{k}_\perp). \tag{3.38}$$

(Notice that the expression in Eq. (3.35) is not even real if evanescent waves are allowed.) The condition (3.37) is essentially that of quasihomogeneity.[1,17] Thus we see that when the source is quasihomogeneous, the z-component of the energy flux can be computed in terms of B by the formula expected from classical radiometry, although the calculation is not exact and there are corrections involving an appropriate small parameter. We will examine these corrections in more detail later.

For now, however, let us consider another expectation of classical radiometry, namely the conservation of brightness along rays. Since s_z is constant along rays, B will be constant if and only if W is constant. Let us therefore examine the quantity

$$W\Big(\mathbf{r}_\perp + \frac{s\mathbf{k}_\perp}{k_0}, \mathbf{k}_\perp; z + \frac{sk_z}{k_0}\Big) = \int d^2\mathbf{q}_\perp\, e^{i[\mathbf{q}_\perp\cdot\mathbf{r}_\perp + (\kappa_+ - \kappa_-)z]}\, \tilde{\Gamma}_0(\mathbf{k}_\perp + \tfrac{1}{2}\mathbf{q}_\perp, \mathbf{k}_\perp - \tfrac{1}{2}\mathbf{q}_\perp)$$
$$\times\, e^{is[\mathbf{q}_\perp\cdot\mathbf{k}_\perp + (\kappa_+ - \kappa_-)k_z]/k_0}, \tag{3.39}$$

where we have invoked Eq. (3.15) and denoted the distance along the ray by s. We are considering only real rays in this expression, so k_z is real, and we also assume κ_- is real. This expression will equal $W(\mathbf{r}_\perp, \mathbf{k}_\perp; z)$ if the final exponential factor in the integrand is unity. In fact, it is not exactly unity, so W is not exactly conserved along rays, but if the source is quasihomogeneous, so that the integral is dominated by small values of $\mathbf{q}_\perp$, then we can expand the final exponent out. We find

$$\mathbf{q}_\perp\cdot\mathbf{k}_\perp + (\kappa_+ - \kappa_-)k_z = -\frac{1}{8k_z^4}\left[k_z^2(\mathbf{k}_\perp\cdot\mathbf{q}_\perp)q_\perp^2 + (\mathbf{k}_\perp\cdot\mathbf{q}_\perp)^3\right] + O(q_\perp^5). \tag{3.40}$$

In other words, the corrections are only of order $q_\perp^3$, which are small for a quasihomogeneous source. These corrections are smaller than the ones which arise in a similar treatment of Walther's (second) definition of radiance, which gives correction terms of order $q_\perp^2$. (The difference is that the Wigner function is a kind of centered Fourier transform, whereas Walther's definition is one-sided.) We will return later to the fact that the Wigner function is better conserved along rays than Walther's function.

Now, however, we will develop a systematic series of corrections to the conservation of W along rays. We begin by assuming that the wave field of interest contains no evanescent waves; the necessity for this assumption will become apparent in a moment. This means that the right hand side of Eq. (2.37) can be replaced

by a commutator, and we can write

$$i\frac{d\hat{\Gamma}(z)}{dz} = [\hat{H}, \hat{\Gamma}(z)]. \tag{3.41}$$

Next we take the Weyl transform of both sides. On the left hand side, we get simply $i\partial W/\partial z$. On the right hand side, we call on the Moyal formula in the form shown in Eq. (3.7). To use this formula, we need the Weyl transform of $\hat{H}$, defined in Eq. (2.25); but by Eq. (3.10), this is just

$$H = -\sqrt{k_0^2 - k_\perp^2}. \tag{3.42}$$

H is the ray Hamiltonian, and is just another notation for $-k_z$. Then we see that the operator in the sine function in Eq. (3.7) simplifies, because H depends only on $\mathbf{k}_\perp$ and all $\mathbf{r}_\perp$-derivatives acting on it vanish. Altogether, we find

$$\frac{\partial W(\mathbf{r}_\perp, \mathbf{k}_\perp; z)}{\partial z} = 2\sqrt{k_0^2 - k_\perp^2}\sin\left[\frac{1}{2}\left(\frac{\overleftarrow{\partial}}{\partial \mathbf{k}_\perp}\cdot\frac{\overrightarrow{\partial}}{\partial \mathbf{r}_\perp}\right)\right] W(\mathbf{r}_\perp, \mathbf{k}_\perp; z). \tag{3.43}$$

Except for the neglect of the evanescent waves and the question of the convergence of the series implied by the sine function, this is exact. The first term of the sine series gives

$$\frac{-1}{\sqrt{k_0^2 - k_\perp^2}}\mathbf{k}_\perp \cdot \nabla_\perp W, \tag{3.44}$$

which can be brought over to the left side to give the total convective derivative along a ray with respect to z,

$$\frac{dW}{dz} = \frac{\partial W}{\partial z} + \frac{1}{k_z}\mathbf{k}_\perp \cdot \nabla_\perp W = \sum_{m=1}^{\infty} \frac{(-1)^m}{2^{2m}(2m+1)!}\frac{\partial^{2m+1}k_z}{\partial \mathbf{k}_\perp^{2m+1}}\cdot\frac{\partial^{2m+1}W}{\partial \mathbf{r}_\perp^{2m+1}}. \tag{3.45}$$

By multiplying this by k_z/k_0 we obtain dW/ds, the convective derivative of W along a ray with respect to distance s. It is the right hand side of this expression which would vanish in classical radiometry; the series we see here gives us the correction terms. In these correction terms, we have used a single dot to represent the complete contraction of the two tensors involved; for example, the $m = 1$ term is more explicitly written

$$\frac{-1}{2^2 3!}\sum_{ij\ell}\frac{\partial^3 k_z}{\partial k_{\perp i}\partial k_{\perp j}\partial k_{\perp \ell}}\frac{\partial^3 W}{\partial r_{\perp i}\partial r_{\perp j}\partial r_{\perp \ell}}. \tag{3.46}$$

The first correction term, shown in Eq. (3.46), gives an estimate of the error committed in neglecting the right hand side of Eq. (3.45). The estimate depends on whether the ray is paraxial ($k_\perp \ll k_z \approx k_0$), substantially off-axis ($k_\perp \sim k_z \sim k_0$), or nearly tangent to the reference plane ($k_z \ll k_\perp \approx k_0$). In the paraxial case, we have the estimate

$$\frac{dW}{ds} \sim \frac{k_z}{k_0}\frac{k_\perp}{k_z^3}\frac{W}{L^3} \sim \frac{\theta}{\lambda}\left(\frac{\lambda}{L}\right)^3 W, \tag{3.47}$$

where θ is the paraxial angle, λ is the wavelength, and L is the spatial scale length of the Wigner function. This estimate may be constrasted with the analogous estimate given by Walther[4] for his function A, which may be written,

$$\frac{dA}{ds} \sim \frac{1}{\lambda}\left(\frac{\lambda}{L}\right)^2 A. \tag{3.48}$$

This is one order of λ/L worse than the estimate for the Wigner function, and does not contain the paraxial factor.

For substantially off-axis rays the estimate for the rate of change of the Wigner function along rays is

$$\frac{dW}{ds} \sim \frac{k_z}{k_0}\left(\frac{k_\perp}{k_z^3} + \frac{k_\perp^3}{k_z^5}\right)\frac{W}{L^3} \sim \frac{1}{\lambda}\left(\frac{\lambda}{L}\right)^3 W, \tag{3.49}$$

which is the same as Eq. (3.47) but without the paraxial factor. The factor in parentheses in this expression is a schematic indication of the third derivative of H with respect to $\mathbf{k}_\perp$, with indices suppressed since we are interested only in order of magnitude. The second term in the parentheses was neglected in the paraxial case; here both are comparable.

One reason we are interested in rays which are substantially off-axis is that such rays occur in nonimaging concentrators.[18] Indeed, the rays in nonimaging concentrators typically cover 2π steradians at the exit aperture. Geometric optics has proven satisfactory for most analysis to date of nonimaging concentrators, but in newer applications diffraction effects are important. We have such applications in mind throughout this paper.

For rays which are nearly tangent to the reference plane, the estimate becomes

$$\frac{dW}{ds} \sim \frac{k_z}{k_0}\frac{k_\perp^3}{k_z^5}\frac{W}{L^3} \sim \frac{1}{\lambda\alpha^4}\left(\frac{\lambda}{L}\right)^3 W, \tag{3.50}$$

where α is the angle between the ray and the xy-plane. This expression obviously diverges when $\alpha \to 0$, but since λ/L is presumably also small, the Wigner function will in many circumstances be conserved along rays even when α is reasonably small. However, the series (3.45) does break down when α becomes small enough, and it becomes meaningless for evanescent waves.

The divergence of the estimate (3.50) as $\alpha \to 0$ is caused mathematically by the square root branch point of the ray Hamiltonian H at $k_\perp = k_0$. Obviously there is nothing physical about this divergence, since the ray does not know that it is nearly tangent to an imaginary reference plane. The reference plane used in all constructions of this sort, not only in this paper but in the vast literature on diffraction and radiometry and coherence, is essentially a surface of section in the mechanical sense. There is no reason why other surfaces of section could not be used, such as spheres, and these might have some advantages. More fundamentally, it is a defect of the entire approach usually taken in treatments of radiometry and coherence that a quantity such as radiance, which is supposed to have a physical meaning if only it could be defined properly, should depend on the reference plane.

For rays which are not too close to $k_\perp = k_0$, the series in Eq. (3.45) will converge rapidly (or start to converge rapidly; the series may be asymptotic) if λ/L is small. The quantity L in certain cases has the significance of the spatial scale length of the average intensity of the light. To show this, we write the mutual intensity in terms of sum and difference variables,

$$\Gamma(\mathbf{r}_\perp, \mathbf{r}'_\perp) = F\Big(\frac{\mathbf{r}_\perp + \mathbf{r}'_\perp}{2}, \mathbf{r}_\perp - \mathbf{r}'_\perp\Big), \tag{3.51}$$

where F is a new function and where we suppress the z-dependence for simplicity. Then we have

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp) = \int d^2\mathbf{a}_\perp \, e^{i\mathbf{k}_\perp \cdot \mathbf{a}_\perp} \, F(\mathbf{r}_\perp, \mathbf{a}_\perp), \tag{3.52}$$

so that L is the scale length of F with respect to its first argument. If L is independent of the value of the second argument of F, then L is also the scale length of the average intensity of the radiation, since

$$\overline{|\psi|^2} = \Gamma(\mathbf{r}_\perp, \mathbf{r}_\perp) = F(\mathbf{r}_\perp, 0). \tag{3.53}$$

Under these assumptions, we can say that the Wigner function is conserved along rays when λ is much less than the scale length of the intensity.

In drawing these conclusions we have not had to make any assumptions about the scale length of F with respect to its second (difference) argument, but in many applications this latter length, which is essentially the correlation length, is small compared to L. In such a case the radiation is quasihomogeneous on the given plane $z = \text{const}$. (Sometimes quasihomogeneity is defined by demanding that the right hand side of Eq. (3.51) factor into the product of a slow function of the sum variable times a fast function of the difference variable. But this is too restrictive; for example, it precludes the case in which the correlation length is a slow function of position.) We denote the correlation length by ℓ. It is never much less than λ, and for quasihomogeneous sources satisfies

$$\lambda \leq \ell \ll L. \tag{3.54}$$

Notice that for paraxial rays the Wigner function is conserved along rays even if λ/L is not small. In fact, if we return to Eq. (3.43) and approximate H in accordance with the paraxial condition,

$$H \approx -k_0 + \frac{k_\perp^2}{2k_0}, \tag{3.55}$$

then the Moyal series terminates after one term with no assumptions on W. In this case we obtain

$$\frac{dW}{dz} = \frac{\partial W}{\partial z} + \{W, H\} = 0. \tag{3.56}$$

For example, in some cases of coherent light W has a spatial scale L which is comparable to a wavelength, but along paraxial rays W is still conserved. This result is a special case of a well-known result in quantum mechanics, that the Wigner function is exactly conserved along classical orbits in the case that the Hamiltonian is at most a quadratic function of q's and p's.[16]

The evolution of the Wigner function along rays has been considered previously by Kim and Wolf,[19] but those authors did not derive explicit formulas for the correction terms nor did they estimate their order of magnitude.

As a final calculation involving the Wigner function, we will work out the correction terms in the z-component of the energy flux, i.e., the terms which express

the difference between Eqs. (3.34) and (3.35). We begin by treating J_z much as we treated $\mathbf{J}_\perp$ in Eq. (3.30). We write

$$\begin{aligned}\frac{k_0}{nc}J_z(\mathbf{r}_{\perp 0},z) &= \mathrm{Im}\,\overline{\frac{\partial\psi(\mathbf{r}_{\perp 0},z)}{\partial z}\,\psi(\mathbf{r}_{\perp 0},z)^*} = -\,\mathrm{Re}\,\overline{\langle\mathbf{r}_{\perp 0}|\hat{H}|\psi\rangle\langle\psi|\mathbf{r}_{\perp 0}\rangle} \\ &= -\,\mathrm{Re}\langle\mathbf{r}_{\perp 0}|\hat{H}\hat{\Gamma}(z)|\mathbf{r}_{\perp 0}\rangle = -\,\mathrm{Re}\,\mathrm{Tr}\Big[|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\,\hat{H}\hat{\Gamma}(z)\Big] \\ &= -\tfrac{1}{2}\,\mathrm{Tr}\Big[\Big(|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\hat{H} + \hat{H}^\dagger|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\Big)\hat{\Gamma}(z)\Big]. \end{aligned} \tag{3.57}$$

Next we need the Weyl transform of the operator in parentheses in the final expression. We use the Moyal formula in the form (3.4) to compute the Weyl transforms of the two terms in the parentheses, and add the results. In doing this we simply treat $\hat{H}$ as Hermitian (and H as real), and ignore evanescent waves. Such waves do not contribute to J_z anyway, and the Moyal formula will lead to divergences near $k_\perp = k_0$, just like it did in the calculation of the evolution of W along rays. The result is

$$\tfrac{1}{2}\Big(|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\hat{H} + \hat{H}|\mathbf{r}_{\perp 0}\rangle\langle\mathbf{r}_{\perp 0}|\Big) \longleftrightarrow \delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0})\cos\Big[\frac{1}{2}\Big(\frac{\overleftarrow{\partial}}{\partial\mathbf{r}_\perp}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{k}_\perp}\Big)\Big]H. \tag{3.58}$$

Next we use Eq. (3.2) to compute the trace, obtaining

$$\begin{aligned}J_z(\mathbf{r}_{\perp 0},z) &= \frac{nc}{k_0}\int\frac{d^2\mathbf{r}_\perp\,d^2\mathbf{k}_\perp}{(2\pi)^2}\,W(\mathbf{r}_\perp,\mathbf{k}_\perp;z) \\ &\quad\times\Big\{\delta(\mathbf{r}_\perp - \mathbf{r}_{\perp 0})\cos\Big[\frac{1}{2}\Big(\frac{\overleftarrow{\partial}}{\partial\mathbf{r}_\perp}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{k}_\perp}\Big)\Big]\sqrt{k_0^2 - k_\perp^2}\Big\}. \end{aligned} \tag{3.59}$$

The first term of the cosine series is the term expected by classical radiometry. In the higher order terms the $\mathbf{r}_\perp$-derivatives acting on the δ-function can be transferred to W by integration by parts, and then the $\mathbf{r}_\perp$-integral can be done. The result is

$$\begin{aligned}J_z(\mathbf{r}) &= \frac{nc}{k_0}\int\frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\,k_z\,W(\mathbf{r}_\perp,\mathbf{k}_\perp;z) \\ &\quad+\frac{nc}{k_0}\sum_{m=1}^{\infty}\frac{(-1)^m}{2^{2m}(2m)!}\int\frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\,\frac{\partial^{2m}W}{\partial\mathbf{r}_\perp^{2m}}\cdot\frac{\partial^{2m}k_z}{\partial\mathbf{k}_\perp^{2m}}. \end{aligned} \tag{3.60}$$

This series is similar to that in Eq. (3.45), with the terms decreasing in powers of $(\lambda/L)^2$.

4. The Walther Distribution Function

In this section we examine Walther's (second) proposed distribution function,[4] in terms of which the brightness or radiance can be defined, and subject it to some of the same kinds of analysis we applied to the Wigner function in Sec. 3. (Walther's first proposed distribution function[3] is essentially the Wigner function itself, which we have already examined.) We will also develop integral and infinite series formulas connecting Walther's function with the Wigner function, and use them to find corrections to the conservation of Walther's function along rays. We find explicit formulas for these corrections, which are different depending on whether Walther's function is used in its complex form or real form. In the case of the complex form, an estimate of the order of magnitude of the first of these correction terms was made by Walther himself[4]; this estimate was repeated above in Eq. (3.48). In the case of the real form, we find a different estimate, which is more favorable than the one given by Walther, but which is still not as favorable as the one given in Eq. (3.47) for the Wigner function.

We begin by providing several equivalent definitions of a distribution function A, which we will call the "Walther function":

$$\begin{aligned} A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) &= (2\pi) e^{i\mathbf{k}_\perp \cdot \mathbf{r}_\perp} \, \overline{\tilde{\psi}(\mathbf{k}_\perp, z) \psi(\mathbf{r}_\perp, z)^*} \\ &= \int d^2\mathbf{k}'_\perp \, e^{i(\mathbf{k}_\perp - \mathbf{k}'_\perp)\cdot \mathbf{r}_\perp} \, \tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}'_\perp; z) \\ &= \int d^2\mathbf{k}'_\perp \, e^{i(\mathbf{k} - \mathbf{k}'^*)\cdot \mathbf{r}} \, \tilde{\Gamma}_0(\mathbf{k}_\perp, \mathbf{k}'_\perp) \\ &= \int d^2\mathbf{r}'_\perp \, e^{i\mathbf{k}_\perp \cdot (\mathbf{r}_\perp - \mathbf{r}'_\perp)} \, \Gamma(\mathbf{r}'_\perp, \mathbf{r}_\perp; z). \end{aligned} \tag{4.1}$$

This function is complex; often we will be interested only in its real part. This definition parallels our earlier definition of the Wigner function in its dimensions and normalization, so that Walther's proposed definition of the radiance can be written,

$$B(\mathbf{r}, \mathbf{s}) = cns_z \left(\frac{k_0}{2\pi}\right)^2 \operatorname{Re} A(\mathbf{r}_\perp, \mathbf{k}_\perp; z), \tag{4.2}$$

just as in Eq. (3.24), with $\operatorname{Re} A$ replacing W.

The Walther function has the following marginal distributions,

$$\int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \Gamma(\mathbf{r}_\perp, \mathbf{r}_\perp; z) = \overline{I(\mathbf{r})}, \tag{4.3}$$

$$\int \frac{d^2\mathbf{r}_\perp}{(2\pi)^2} A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \tilde{\Gamma}(\mathbf{k}_\perp, \mathbf{k}_\perp; z), \tag{4.4}$$

in analogy with Eqs. (3.18) and (3.19) for the Wigner function. In the case of both functions, the two formulas are exact. Also, since the right hand sides of these equations are real, we can replace A on the left hand sides by $\mathrm{Re}\, A$; and we have the following identities for the imaginary part of A,

$$\int d^2\mathbf{r}_\perp \, \mathrm{Im}\, A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int d^2\mathbf{k}_\perp \, \mathrm{Im}\, A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = 0. \tag{4.5}$$

As for the energy flux, it is straightforward to combine Eq. (4.1) with the identities of Sec. 2 to obtain

$$\mathbf{J}(\mathbf{r}) = \frac{nc}{k_0} \mathrm{Re} \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} \mathbf{k}\, A(\mathbf{r}_\perp, \mathbf{k}_\perp; z), \tag{4.6}$$

which can be compared to Eqs. (3.29) and (3.60). For the perpendicular component of this equation, the Re operator can be pulled through the integrand adjacent to A, so that the (exact) formula looks just like Eq. (3.29), with the Wigner function W replaced by $\mathrm{Re}\, A$. For the z-component, the same can only be done if evanescent waves can be ignored, so that $k_z^* = k_z$. In that case, the Walther function gives a formula for the energy flux which is simpler than in the case of the Wigner function,

$$\mathbf{J} = \frac{nc}{k_0} \int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2} \mathbf{k}\, \mathrm{Re}\, A, \tag{4.7}$$

although in both cases it is the z-component which gives trouble. Walther's original definition of his function was designed to make this equation come out as shown.

Now we turn to the evolution of the Walther function along rays. To study this question, it is convenient first to make a connection between the Walther function and the Wigner function, so that the properties of the latter can be invoked. Since both the Wigner function and the Walther function are related to $\Gamma(\mathbf{r}_\perp, \mathbf{r}'_\perp; z)$ by

invertible integral transforms, it is straightforward to find an integral transform connecting W and A. The desired transform and its inverse are

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int \frac{d^2\mathbf{r}'_\perp\, d^2\mathbf{k}'_\perp}{\pi^2}\, e^{-2i(\mathbf{r}_\perp - \mathbf{r}'_\perp)\cdot(\mathbf{k}_\perp - \mathbf{k}'_\perp)}\, A(\mathbf{r}'_\perp, \mathbf{k}'_\perp; z), \tag{4.8}$$

$$A(\mathbf{r}_\perp, \mathbf{k}_\perp; z) = \int \frac{d^2\mathbf{r}'_\perp\, d^2\mathbf{k}'_\perp}{\pi^2}\, e^{+2i(\mathbf{r}_\perp - \mathbf{r}'_\perp)\cdot(\mathbf{k}_\perp - \mathbf{k}'_\perp)}\, W(\mathbf{r}'_\perp, \mathbf{k}'_\perp; z). \tag{4.9}$$

These integral formulas are exact, but it is convenient to transform them into another (differential) form in which the effects of short wavelength and/or short correlation length will be manifest. We begin with Eq. (4.9), making the changes of variable, $\mathbf{r}'_\perp = \mathbf{r}_\perp + \mathbf{a}_\perp$, $\mathbf{k}'_\perp = \mathbf{k}_\perp + \mathbf{q}_\perp$, and then expanding W in a Taylor series about $\mathbf{q}_\perp = 0$. For convenience, we also suppress the z-dependence. We find

$$\begin{aligned} A(\mathbf{r}_\perp, \mathbf{k}_\perp) &= \int \frac{d^2\mathbf{a}_\perp\, d^2\mathbf{q}_\perp}{\pi^2}\, e^{2i\mathbf{a}_\perp\cdot\mathbf{q}_\perp}\, W(\mathbf{r}_\perp + \mathbf{a}_\perp, \mathbf{k}_\perp + \mathbf{q}_\perp) \\ &= \int \frac{d^2\mathbf{a}_\perp\, d^2\mathbf{q}_\perp}{\pi^2}\, e^{2i\mathbf{a}_\perp\cdot\mathbf{q}_\perp} \sum_{m=0}^{\infty} \frac{1}{m!}\Big(\mathbf{q}_\perp \cdot \frac{\partial}{\partial \mathbf{k}_\perp}\Big)^m W(\mathbf{r}_\perp + \mathbf{a}_\perp, \mathbf{k}_\perp). \end{aligned} \tag{4.10}$$

Next we note that the $\mathbf{q}_\perp$-integral can be done in terms of δ-functions and their derivatives, which in turn allow the $\mathbf{a}_\perp$-integral to be done:

$$\begin{aligned} A(\mathbf{r}_\perp, \mathbf{k}_\perp) &= \int \frac{d^2\mathbf{a}_\perp\, d^2\mathbf{q}_\perp}{\pi^2} \sum_{m=0}^{\infty} \frac{1}{m!(2i)^m} \Big[e^{2i\mathbf{a}_\perp\cdot\mathbf{q}_\perp} \Big(\frac{\overleftarrow{\partial}}{\partial \mathbf{a}_\perp} \cdot \frac{\overrightarrow{\partial}}{\partial \mathbf{k}_\perp}\Big)^m W(\mathbf{r}_\perp + \mathbf{a}_\perp, \mathbf{k}_\perp)\Big] \\ &= \int d^2\mathbf{a}_\perp \sum_{m=0}^{\infty} \frac{1}{m!(2i)^m} \Big[\delta(\mathbf{a}_\perp)\Big(\frac{\overleftarrow{\partial}}{\partial \mathbf{a}_\perp} \cdot \frac{\overrightarrow{\partial}}{\partial \mathbf{k}_\perp}\Big)^m W(\mathbf{r}_\perp + \mathbf{a}_\perp, \mathbf{k}_\perp)\Big] \\ &= \sum_{m=0}^{\infty} \frac{1}{m!}\Big(\frac{i}{2}\Big)^m \Big(\frac{\partial^2}{\partial \mathbf{r}_\perp \cdot \partial \mathbf{k}_\perp}\Big)^m W(\mathbf{r}_\perp, \mathbf{k}_\perp). \end{aligned} \tag{4.11}$$

In the the first and second of these formulas, we use a notation as in the Moyal formula, Eq. (3.4), in which the arrow shows the direction in which the operand of an operator lies. When no arrow is present, it is assumed that the operand lies to the right.

Thus we have found an infinite series connecting A with W. It is convenient to write this series in terms of an operator,

$$\hat{D} = \frac{\partial^2}{\partial \mathbf{r}_\perp \cdot \partial \mathbf{k}_\perp}, \tag{4.12}$$

so that

$$A(\mathbf{r}_\perp, \mathbf{k}_\perp) = e^{(i/2)\hat{D}}\, W(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{4.13}$$

We use a hat on $\hat{D}$ to denote an operator, but we must remember that this operator and others to be introduced momentarily act on phase space distributions, i.e., functions of $(\mathbf{r}_\perp, \mathbf{k}_\perp)$, and not on wave functions defined over $\mathbf{r}_\perp$-space, upon which the operators introduced earlier, $\hat{K}$, $\hat{H}$, $\hat{\Gamma}$, etc., acted.

The operator notation shows its power when we ask for the inverse of Eq. (4.13); it is simply

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp) = e^{-(i/2)\hat{D}}\, A(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{4.14}$$

Furthermore, since W is real, we can take the real and imaginary parts of Eq. (4.13) to find

$$\operatorname{Re} A(\mathbf{r}_\perp, \mathbf{k}_\perp) = \cos(\tfrac{1}{2}\hat{D}) W(\mathbf{r}_\perp, \mathbf{k}_\perp), \tag{4.15}$$

$$\operatorname{Im} A(\mathbf{r}_\perp, \mathbf{k}_\perp) = \sin(\tfrac{1}{2}\hat{D}) W(\mathbf{r}_\perp, \mathbf{k}_\perp) = \tan(\tfrac{1}{2}\hat{D}) \operatorname{Re} A(\mathbf{r}_\perp, \mathbf{k}_\perp), \tag{4.16}$$

$$W(\mathbf{r}_\perp, \mathbf{k}_\perp) = \sec(\tfrac{1}{2}\hat{D}) \operatorname{Re} A(\mathbf{r}_\perp, \mathbf{k}_\perp). \tag{4.17}$$

We see that the real and imaginary parts of Walther's function A are not independent of one another, but rather the imaginary part can be derived from the real one. This fact was noted previously by Walther,[20] who derived the relation in a somewhat different form. (Evidently, the real part cannot be derived uniquely from the imaginary part without specifying boundary conditions, since the operator $\hat{D}^{-1}$ is an integral, not a differential, operator. For this reason, we avoid the cotangent and cosecant functions in the formulas above.)

In these formulas, the transcedental functions of the operator $\hat{D}$ are understood as shorthand for the corresponding infinite series, such as shown in the final expression of Eq. (4.11). The series may be convergent or asymptotic; in the latter case, the terms may either start out decreasing rapidly in magnitude, or not. If the terms do not decrease rapidly in magnitude, then the series will not be of much use.

To find out when these terms do decrease rapidly in magnitude, we make the estimate that the $\mathbf{r}_\perp$-derivative acting on W is of the order of $1/L$, as in Eqs. (3.47)–(3.50), where L is the spatial scale length of the average intensity, and we estimate

the effect of the $\mathbf{k}_\perp$-derivative on W by a factor of the correlation length ℓ. The reason for the latter estimate is seen in Eq. (3.52), which shows that the $\mathbf{k}_\perp$-derivative acting on W is of the order of the spatial scale length of the function F, defined in Eq. (3.51), with respect to its second argument. This is the correlation length ℓ. Altogether, the estimate of the effect of the operator $\hat{D}$ is

$$\hat{D} \sim \ell/L, \tag{4.18}$$

which is $\ll 1$ in the case of quasihomogeneous sources, as shown by Eq. (3.54). Thus we see that the terms of the series in Eqs. (4.13)–(4.17) decrease by powers of ℓ/L, and that the convergence is good for quasihomogeneous sources. For coherent or nearly coherent sources, however, we will have $\ell \sim L$, and the series will converge slowly if at all.

An example will illustrate these features. For the purpose of this example, we suppress the z-dependence of A and W, and we replace the 2-dimensional xy-plane by the 1-dimensional x-axis, writing x and k instead of $\mathbf{r}_\perp$ and $\mathbf{k}_\perp$. To have a model which is analytically tractible and sometimes even physically relevant, we assume that $\Gamma(x, x')$ has the Gaussian-Schell form,[21]

$$\Gamma(x, x') = \frac{I_0}{L\sqrt{2\pi}} \exp\left[-\frac{1}{2L^2}\left(\frac{x+x'}{2}\right)^2 - \frac{(x-x')^2}{2\ell^2}\right], \tag{4.19}$$

where I_0 is a reference intensity, and where L and ℓ are the intensity scale length and correlation length as above, defined precisely in a r.m.s. sense. The normalization is chosen so that

$$\int \Gamma(x, x)\, dx = I_0. \tag{4.20}$$

The Gaussian-Schell model shown is not physically realizable if $\ell/L > 2$, because in that case the density operator $\hat{\Gamma}$ has negative eigenvalues; it corresponds to coherent light if $\ell/L = 2$, in which case the wave field is given by

$$\psi(x) = \sqrt{\frac{I_0}{L\sqrt{2\pi}}}\, e^{-x^2/4L^2}; \tag{4.21}$$

and it corresponds to incoherent light if $\ell/L < 2$, becoming quasihomogeneous as $\ell/L \ll 2$.

Now we compute both the Wigner and Walther functions for the Gaussian-Schell model, using 1-dimensional versions of Eqs. (3.14) and (4.1). We find,

$$W(x,k) = \frac{I_0 \ell}{L} \exp\left[-\frac{x^2}{2L^2} - \frac{\ell^2 k^2}{2}\right], \tag{4.22}$$

$$A(x,k) = \frac{2I_0\ell}{\sqrt{4L^2+\ell^2}} \exp\left[\frac{2(-x^2 - L^2\ell^2k^2 + i\ell^2 xk)}{4L^2+\ell^2}\right]. \tag{4.23}$$

We notice that for coherent light ($\ell = 2L$) the two functions do not approximate one another well at all, since A has a significant complex part and the two spatial scale lengths, while of the same order of magnitude, differ by the factor $\sqrt{2}$. If, however, we hold the dimensionless parameters x/L and ℓk fixed while letting ℓ/L become small, then we see that the imaginary part of A becomes small and the real part of A approximates the Wigner function W better and better. These conclusions are a special case of a general observation made previously by Walther,[20] who noted that in the quasihomogeneous limit, A becomes real and equal to W. Walther's observation was more general than ours because he considered lenses and other complications, and we are restricted to a homogeneous medium. The same conclusions were drawn for a homogeneous medium by Carter and Wolf.[17]

Now we will test a 1-dimensional version of the series shown in Eq. (4.13). First we note that

$$\frac{d^m}{dz^m} e^{-z^2} = (-1)^m H_m(z) e^{-z^2}, \tag{4.24}$$

where H_m is the m-th Hermite polynomial, so that

$$\left(\frac{\partial^2}{\partial x \partial k}\right)^m W(x,k) = \left(\frac{\ell}{2L}\right)^m H_m\left(\frac{x}{L\sqrt{2}}\right) H_m\left(\frac{\ell k}{\sqrt{2}}\right) W(x,k), \tag{4.25}$$

and

$$e^{(i/2)\hat{D}} W(x,k) = \sum_{m=0}^{\infty} \frac{1}{m!}\left(\frac{i\ell}{4L}\right)^m H_m\left(\frac{x}{L\sqrt{2}}\right) H_m\left(\frac{\ell k}{\sqrt{2}}\right) W(x,k). \tag{4.26}$$

Next we use the identity,

$$H_m(z) = \frac{2^m}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (z+it)^m e^{-t^2}\, dt, \tag{4.27}$$

to replace the two Hermite polynomials in Eq. (4.26) and to make the series a summable exponential series. This gives

$$e^{(i/2)\hat{D}}\,W(x,k) = \frac{\ell}{\pi L}\int\int ds\,dt\,\exp\Big[-\frac{x^2}{2L^2} - \frac{\ell^2 k^2}{2} + \frac{i\ell}{L}\Big(\frac{x}{L\sqrt{2}} + is\Big)\Big(\frac{\ell k}{\sqrt{2}} + it\Big) - s^2 - t^2\Big]. \quad (4.28)$$

Finally, we do the s and t integrals, and find precisely the function $A(x,k)$ of Eq. (4.23). In this Gaussian-Schell example, the infinite series is actually convergent for all values of $\ell/L \leq 2$ (it diverges for the nonphysical values $\ell/L > 2$). In general we must not expect such luck.

To return to the problem of finding the rate of evolution of the Walther function along a ray, we differentiate Eq. (4.13) to obtain

$$\frac{\partial A}{\partial z} = e^{(i/2)\hat{D}}\,\frac{\partial W}{\partial z}, \quad (4.29)$$

where we can pull the operator $\partial/\partial z$ through the exponential since $\partial/\partial z$ commutes with $\hat{D}$. Next, for notational convenience we introduce the operators

$$\hat{S}_n = \frac{\partial^n k_z}{\partial \mathbf{k}_\perp^n}\cdot\frac{\partial^n}{\partial \mathbf{r}_\perp^n}, \quad (4.30)$$

using the same convention for the contraction of indices illustrated in Eq. (3.46). In terms of these operators, we write Eq. (3.45), the equation for the evolution of the Wigner function, in the form,

$$\frac{\partial W}{\partial z} = \sum_{m=0}^{\infty}\frac{(-1)^m}{2^{2m}(2m+1)!}\hat{S}_{2m+1}W. \quad (4.31)$$

The $m = 0$ term of the right hand side is $\hat{S}_1 W = -(\mathbf{k}_\perp\cdot\nabla_\perp W)/k_z$, which, when brought over to the left hand side, gives dW/dz, the rate of change of W along a ray with respect to z. The terms $m > 0$ are the corrections which go beyond classical radiometry, with $m = 1$ being dominant for short wavelengths. Substituting Eq. (4.31) into Eq. (4.29) and using Eq. (4.14), we obtain an equation of evolution purely in terms of the Walther function A,

$$\frac{\partial A}{\partial z} = \sum_{m=0}^{\infty}\frac{(-1)^m}{2^{2m}(2m+1)!}\Big[e^{(i/2)\hat{D}}\hat{S}_{2m+1}e^{-(i/2)\hat{D}}\Big]A. \quad (4.32)$$

If the exponentials on the right hand side are expanded out, we get a triply infinite series for the rate of change of A along rays; but it turns out that this can be simplified to a singly infinite series. The first step in the simplification is to invoke the identity,

$$e^{\hat{A}}\hat{B}e^{-\hat{A}} = \hat{B} + [\hat{A},\hat{B}] + \frac{1}{2!}[\hat{A},[\hat{A},\hat{B}]] + \ldots, \tag{4.33}$$

where $\hat{A}$ and $\hat{B}$ are any operators and where the right hand side is an exponential series of iterated commutators. This formula is standard in the theory of Lie algebras, and is commonly used in quantum mechanics. It allows us to express Eq. (4.32) in terms of the iterated commutators of $\hat{D}$ and $\hat{S}_n$. But a direct computation shows that

$$[\hat{D},\hat{S}_n] = \hat{S}_{n+1}, \qquad [\hat{D},[\hat{D},\hat{S}_n]] = \hat{S}_{n+2}, \tag{4.34}$$

etc., so that Eq. (4.32) can be reexpressed in terms of a doubly infinite series of the $\hat{S}$ operators:

$$\frac{\partial A}{\partial z} = \sum_{m=0}^{\infty} \frac{(-1)^m}{2^{2m}(2m+1)!} \sum_{p=0}^{\infty} \frac{1}{p!}\left(\frac{i}{2}\right)^p \hat{S}_{2m+p+1}A. \tag{4.35}$$

Now this doubly infinite series can be simplified further to a singly infinite series. First we collect all the terms on the right hand side for which the index p is even or odd, calling the corresponding operators $\hat{E}$ and $i\hat{O}$ respectively. Then, with a redefinition of the index p, we have

$$\hat{E} = \sum_{m=0}^{\infty}\sum_{p=0}^{\infty} \frac{(-1)^{m+p}}{2^{2m+2p}(2m+1)!(2p)!}\hat{S}_{2m+2p+1}, \tag{4.36}$$

$$\hat{O} = \sum_{m=0}^{\infty}\sum_{p=0}^{\infty} \frac{(-1)^{m+p}}{2^{2m+2p+1}(2m+1)!(2p+1)!}\hat{S}_{2m+2p+2}. \tag{4.37}$$

Working first on $\hat{E}$, we write $m' = m + p$, rearrange the summation, and drop the prime to obtain

$$\hat{E} = \sum_{m=0}^{\infty} \frac{(-1)^m \hat{S}_{2m+1}}{2^{2m}} \sum_{p=0}^{m} \frac{1}{(2m-2p+1)!(2p)!}. \tag{4.38}$$

The finite p-series on the right is easily expressed in terms of binomial coefficients and summed; it turns out to be $2^{2m}/(2m+1)!$. Therefore we have a rather simple result,

$$\hat{E} = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)!} \hat{S}_{2m+1}. \tag{4.39}$$

Similarly, we find

$$\hat{O} = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+2)!} \hat{S}_{2m+2}. \tag{4.40}$$

Altogether, we find the equation for evolution of the Walther function along a ray in the form,

$$\frac{\partial A}{\partial z} = (\hat{E} + i\hat{O})A = \sum_{m=0}^{\infty} (-1)^m \left[\frac{\hat{S}_{2m+1}}{(2m+1)!} + i \frac{\hat{S}_{2m+2}}{(2m+2)!} \right] A. \tag{4.41}$$

The $r = 0$ term of the $\hat{E}$ series is $\hat{S}_1 A$, which, when brought over to the left hand side and combined with $\partial A/\partial z$, gives dA/dz, the rate of change of A along rays with respect to z. The leading correction term is the $r = 0$ term of the $\hat{O}$ series; thus, when the terms are rapidly decreasing, we can estimate the deviations from classical radiometry by

$$\frac{dA}{dz} \approx \frac{i}{2} \hat{S}_2 A = \frac{i}{2} \frac{\partial^2 k_z}{\partial \mathbf{k}_\perp^2} \cdot \frac{\partial^2 A}{\partial \mathbf{r}_\perp^2} \sim \frac{A}{k_z L^2} \sim \frac{1}{\lambda} \left(\frac{\lambda}{L} \right)^2 A. \tag{4.42}$$

Here we are assuming that the rays are either paraxial or substantially off axis, but not nearly tangent to the reference plane. This is the same estimate made by Walther[4] and repeated in Eq. (3.48), except that here we have not just an estimate, but an explicit formula for the first and all higher order correction terms.

In the quasihomogeneous limit, the Walther function A is almost real, but we see from Eq. (4.42) that its rate of change along a ray is dominantly pure imaginary. This suggests that the real part of A is better conserved along rays than A itself. Indeed, taking the real part of Eq. (4.41), we have

$$\frac{\partial \operatorname{Re} A}{\partial z} = \hat{E}(\operatorname{Re} A) - \hat{O}(\operatorname{Im} A) = \left[\hat{E} - \hat{O} \tan\left(\tfrac{1}{2}\hat{D}\right) \right] (\operatorname{Re} A), \tag{4.43}$$

where we have called on Eq. (4.16) to express the answer purely in terms of $\operatorname{Re} A$. This is the equation of evolution of $\operatorname{Re} A$ along a ray. Again, it is the $m = 0$ term of

the $\hat{E}$ series which, when brought to the left hand side, gives $d(\mathrm{Re}\,A)/dz$; and the dominant correction is the leading term of the product, $-\hat{O}\tan(\hat{D}/2)$. Therefore in the quasihomogeneous limit we can approximate and estimate, finding

$$\frac{d\,\mathrm{Re}\,A}{dz} \approx -\frac{1}{4}\hat{S}_2\hat{D}(\mathrm{Re}\,A) = -\frac{1}{4}\frac{\partial^2 k_z}{\partial \mathbf{k}_\perp^2}\cdot\frac{\partial^4(\mathrm{Re}\,A)}{\partial \mathbf{r}_\perp^2(\partial \mathbf{r}_\perp\cdot\partial \mathbf{k}_\perp)}$$

$$\sim \frac{1}{k_z}\frac{\ell}{L^3}(\mathrm{Re}\,A) \sim \frac{1}{\lambda}\left(\frac{\ell}{L}\right)\left(\frac{\lambda}{L}\right)^3(\mathrm{Re}\,A). \tag{4.44}$$

This is indeed better than the estimate for the complex function A, but it is not as good as the estimate (3.47) for the Wigner function, since it does not contain the paraxial factor and since it does contain the factor ℓ/λ, which can be large.

It is interesting that the equations of evolution for both W and A involve only the $\hat{S}$ operators, so that the terms of the series can be estimated in terms of the single dimensionless parameter λ/L, which does not contain the correlation length. But the equation of evolution for $\mathrm{Re}\,A$ involves the additional parameter ℓ/λ.

Apart from other considerations, one would have to conclude that the Wigner function would be preferable to the Walther function, if the errors committed in using the classical rules of propagation are a concern. On the other hand, the Walther function may have other advantages over the Wigner function, e.g., it may be easier to compute in some circumstances, in which case the results of this section can be used to control or compensate for the errors.

5. A New Distribution Function

We now introduce a new distribution function, which has the property that it is exactly conserved along rays. We follow Walther[3] in motivating this definition. First we write down a formula for the energy flux,

$$\mathbf{J}(\mathbf{r}) = \frac{nc}{k_0}\int \frac{d^2\mathbf{k}'_\perp\, d^2\mathbf{k}''_\perp}{(2\pi)^2}\left(\frac{\mathbf{k}'+\mathbf{k}''}{2}\right)e^{i(\mathbf{k}'-\mathbf{k}'')\cdot\mathbf{r}}\,\tilde{\Gamma}_0(\mathbf{k}'_\perp,\mathbf{k}''_\perp), \tag{5.1}$$

which is a combination of Eqs. (3.28) and (3.34). The same formula was used previously by Winston and Ning,[22] who used it to construct a conserved flux from plane waves. We have made the z-components of $\mathbf{k}'$ and $\mathbf{k}''$ real in this equation,

assuming for simplicity that the wave field specified by $\tilde{\Gamma}_0$ contains only travelling waves. Next we perform a change of variables in the integral, $(\mathbf{k}'_\perp, \mathbf{k}''_\perp) \leftrightarrow (\mathbf{k}_\perp, \mathbf{q}_\perp)$, where the new variables of integration are defined by

$$\mathbf{q} = \mathbf{k}' - \mathbf{k}'', \tag{5.2}$$

$$\mathbf{k} = \frac{\mathbf{k}' + \mathbf{k}''}{2D}, \tag{5.3}$$

$$D = \sqrt{\frac{1}{2}\left(1 + \frac{\mathbf{k}' \cdot \mathbf{k}''}{k_0^2}\right)} = \frac{|\mathbf{k}' + \mathbf{k}''|}{2k_0}. \tag{5.4}$$

Since the independent variables of integration are the 2-vectors $(\mathbf{k}'_\perp, \mathbf{k}''_\perp)$ or $(\mathbf{k}_\perp, \mathbf{q}_\perp)$, the z-components of Eqs. (5.2) and (5.3) must be understood as functions of the perpendicular components. We do this as follows. First, if $(\mathbf{k}'_\perp, \mathbf{k}''_\perp)$ are regarded as the independent variables, then k'_z, k''_z are defined as in Eq. (2.9). This gives meaning to D, q_z, and k_z as functions of $(\mathbf{k}'_\perp, \mathbf{k}''_\perp)$, as well as to $\mathbf{q}_\perp$, $\mathbf{k}_\perp$. Notice that we have the important identities,

$$|\mathbf{k}|^2 = k_0^2, \tag{5.5}$$

$$\mathbf{q} \cdot \mathbf{k} = 0. \tag{5.6}$$

Next, if $(\mathbf{k}_\perp, \mathbf{q}_\perp)$ are regarded as independent variables, i.e., if we want the inverse transformation, then we define k_z as a function of $\mathbf{k}_\perp$ as in Eq. (2.9), we define q_z and D by

$$q_z = -\frac{\mathbf{k}_\perp \cdot \mathbf{q}_\perp}{k_z}, \tag{5.7}$$

$$D = \sqrt{1 - |\mathbf{q}_\perp|^2/4k_0^2}, \tag{5.8}$$

and finally we write

$$\begin{aligned} \mathbf{k}' &= D\mathbf{k} + \mathbf{q}/2, \\ \mathbf{k}'' &= D\mathbf{k} - \mathbf{q}/2. \end{aligned} \tag{5.9}$$

Next, to transform the integral (5.1), we need the Jacobian connecting $(\mathbf{k}'_\perp, \mathbf{k}''_\perp)$ and $(\mathbf{k}_\perp, \mathbf{q}_\perp)$. The calculation of this Jacobian takes some effort; we were unable to

find a clever way of doing it, so we report here the result of a brute-force approach. It is

$$\left|\frac{\partial(\mathbf{k}'_\perp, \mathbf{k}''_\perp)}{\partial(\mathbf{k}_\perp, \mathbf{q}_\perp)}\right| = \frac{k'_z k''_z}{k_z^2}. \tag{5.10}$$

Thus, we are able to write,

$$\mathbf{J}(\mathbf{r}) = \frac{nc}{k_0}\int \frac{d^2\mathbf{k}_\perp}{(2\pi)^2}\,\mathbf{k}\,R(\mathbf{r}, \mathbf{k}_\perp), \tag{5.11}$$

where the new distribution function R is given by

$$R(\mathbf{r}, \mathbf{k}_\perp) = \int d^2\mathbf{q}_\perp \frac{k'_z k''_z D}{k_z^2} e^{i\mathbf{q}\cdot\mathbf{r}}\,\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp). \tag{5.11}$$

Here the vectors $\mathbf{k}'$, $\mathbf{k}''$ are regarded as functions of $(\mathbf{k}_\perp, \mathbf{q}_\perp)$ as in Eqs. (5.7)–(5.9). Thus, by construction, the distribution function R reproduces exactly the formula for the energy flux which is expected from classical radiometry.

But the most notable property of this function is its exact conservation along rays. This follows by replacing $\mathbf{r}$ in Eq. (5.11) by $\mathbf{r} + s\mathbf{k}/k_0$, where s is the distance along a ray. Because of Eq. (5.6), we have

$$R(\mathbf{r} + s\mathbf{k}/k_0, \mathbf{k}_\perp) = R(\mathbf{r}, \mathbf{k}_\perp), \tag{5.12}$$

exactly.

As an example, let us compute the R-function for a plane wave,

$$\psi(\mathbf{r}) = \frac{\sqrt{I_0}}{2\pi} e^{i\mathbf{K}\cdot\mathbf{r}}, \tag{5.13}$$

where I_0 is a normalization intensity and $\mathbf{K}$ is a real wave vector satisfying $|\mathbf{K}|^2 = k_0^2$. Then we have

$$\tilde{\Gamma}_0(\mathbf{k}'_\perp, \mathbf{k}''_\perp) = I_0\delta(\mathbf{k}'_\perp - \mathbf{K}_\perp)\delta(\mathbf{k}''_\perp - \mathbf{K}_\perp) = I_0\delta(\mathbf{q}_\perp)\delta(\mathbf{k}'_\perp - \mathbf{K}_\perp). \tag{5.14}$$

Substituting this into Eq. (5.11), we easily find

$$R(\mathbf{r}, \mathbf{k}_\perp) = I_0\delta(\mathbf{k}_\perp - \mathbf{K}_\perp). \tag{5.15}$$

This function is indeed exactly conserved along rays, because it is independent of $\mathbf{r}$; it is also exactly what we would expect for the phase space distribution function representing a plane wave.

If we try to compute R for a less trivial example, say, a Gaussian-Schell model, then we find that the integral in Eq. (5.11) cannot be done in terms of elementary functions. This is as it must be, since a Gaussian-Schell model on some plane $z =$ const does not remain Gaussian under the exact free space propagation discussed in Sec. 2, and since the R-function is exactly conserved along rays. In other words, doing the integral of Eq. (5.11) necessarily includes all the complications involved in an exact propagation. For a plane wave these complications are not serious, so we are not surprised that we are able to do the integral in this case. Of course, if we are doing numerical integrations, then computing the R-function is no more difficult than computing any other distribution function, and it makes the z-propagation much easier.

6. Conclusions

We will conclude this analysis with some proposals for further study. First, the conservation of the Wigner function along rays is remarkably good, especially when the paraxial factor and the numerical factor of $1/2^2 3! = 1/24$ are taken into account. One suspects therefore that it might be possible to come rather close, say, to an edge, and still obtain good results by ray tracing. It would be interesting to consider this question from a practical point of view. Next, it is easy to develop various perturbation schemes which use the rays for a zeroth order approximation, and which take the correction terms developed in this paper as perturbations. It would be interesting to examine these questions more closely. Third, the results developed here should be extended to more complicated optical systems, such as those including lenses. Again, the Wigner function can be expected to be a useful place to start, although one must proceed carefully when discontinuitites such as the transition from air to glass are present. Finally, the new distribution function R presented in this paper should be understood better. For example, we would like to know what the most general distribution function is which is exactly conserved along rays, and whether they can be generalized to take care of lenses, etc. We hope to report on some of these questions in the future.

References

1. E. Wolf, J. Opt. Soc. Am. **68**, 6(1978).
2. L. A. Apresyan and Yu. A. Kravtsov, Usp. Fiz. Nauk **142**, 689(1984) [Sov. Phys. Usp. **27**, 301(1984)].
3. A. Walther, J. Opt. Soc. Am. **58**, 1256(1968).
4. A. Walther, J. Opt. Soc. Am. **63**, 1622(1973).
5. E. Wigner, Phys. Rev. **40**, 749(1932).
6. H. Weyl, Z. Phys. **46**, 1(1927).
7. J. E. Moyal, Proc. Camb. Phil. Soc. **45**, 99(1949).
8. T. Jannson, J. Opt. Soc. Am. **70**, 1544(1980).
9. T. Jannson and R. Janicki, Optik **56**, 429(1980).
10. M. Born and E. Wolf, *Principles of Optics* (Pergamon Press, Oxford, 6th ed., 1980).
11. S. Solimeno, B. Crosignani, P. Di Porto, *Guiding, Diffraction, and Confinement of Optical Radiation* (Academic Press, New York, 1986).
12. G. S. Agarwal and E. Wolf, Phys. Rev. D **2**, 2161(1970); **2**, 2187(1970); **2**, 2206(1970).
13. N. L. Balazs and B. K. Jennings, Phys. Rep. **104**, 347(1984).
14. M. Hillery, R. F. O'Connell, M. O. Scully, and E. P. Wigner, Phys. Rep. **106**, 123(1984).
15. S. W. McDonald, Phys. Rep. **158**, 337(1988).
16. R. G. Littlejohn, Phys. Rep. **138**, 193(1986).
17. W. H. Carter and E. Wolf, J. Opt. Soc. Am. **67**, 785(1977).
18. W. T. Welford and R. Winston, *High Collection Nonimaging Optics* (Academic
19. K. Kim and E. Wolf, J. Opt. Soc. Am. A **4**, 1233(1987).
20. A. Walther, J. Opt. Soc. Am. **68**, 1606(1978).
21. R. Simon, E. C. G. Sudarshan and N. Mukunda, Phys. Rev. A **29**, 3273(1984).
22. R. Winston and X. Ning, J. Opt. Soc. Am. A **3**, 1629(1986).

END

DATE
FILMED
6/24/93