The Interrelation of Physics and Mathematics in Ion-Neutralization Spectroscopy

H. D. Hagstrum and G. E. Becker
Bell Telephone Laboratories, Murray Hill, New Jersey 07974
(Received 9 March 1971)

Because the Auger neutralization of an ion at a solid surface involves two electrons, the kinetic-energy distribution of electrons ejected in the process is related to the convolution of two transition-density functions involving electronic-state–density and transition-probability factors. This fact interrelates the physics and the mathematics of the problem in an important and interesting way. Knowledge of how the physics puts limitations on the functions involved which make deconvolution mathematically tractable and how the mathematical procedures provide information concerning the physics is absolutely essential to the successful operation of the ion-neutralization spectroscopy. The nature of possible spurious structure introduced by the mathematics has led to the devising of tests for such features. It is shown that these test procedures are entirely adequate for the "peaked" functions encountered. Some discussion of inversion procedures is provided in the appendices.

I. INTRODUCTION

We discuss in this paper one of the central questions concerning the feasibility of the ion-neutralization spectroscopy (INS), the development and use of which has been reported in other papers.[1–4] INS is an electron spectroscopy based on a two-electron Auger-type ejection process. The method determines from the measured electron-energy distribution a so-called transition-density function which is essentially the *local* density of states in the surface region of the solid upon which the slowly moving ions are incident. Because two electrons are involved in the Auger neutralization process, the energy distribution of the ejected electrons will, in the one-electron statement of the problem, have the form of the convolution product

$$g(x) = \int_0^x v(x-t)w(t)\,dt = v * w \qquad (1)$$

of the transition densities $v(x)$ and $w(x)$ of the participating electrons. Each transition density is the product of initial-state–density and transition-probability factors which, in general, could be different for the two electrons. It is evident that if we must consider the electron kinetic-energy distribution to have the form of Eq. (1), the determination of v and w, which is the goal of INS, is impossible given g alone. In the work which we have done we have treated $g(x)$ as though it were a convolution square

$$f(x) = \int_0^x u(x-t)u(t)\,dt = u * u = u^{(2)}\,, \qquad (2)$$

which can be inverted accurately to obtain u from f by a simple digital sequential procedure. It is a primary purpose of this paper to discuss the physical and mathematical justifications for this assumption. We shall show how the local density of states at the surface of a solid limits us to a class of functions which makes the replacement of Eq. (1) with Eq. (2) possible. The u function obtained from Eq. (2) is an average of the v and w functions of Eq. (1) having the same peaked structures due to surface electron resonances as are present with possibly different magnitudes in both v and w. We also discuss the mathematics of deconvolution of this class of function sufficiently to indicate the nature of possible errors which could be introduced in the inversion procedure and devise tests for the identification of such errors if present. We show how errors in u introduced by errors in finding the origin of f, obscured by physical broadening, can be removed by the application of a simple origin-shifting procedure.

After a brief discussion in Sec. II of INS and its assumptions we state in Sec. III the deconvolution procedure used in this work. Other sequential procedures we have investigated are mentioned in Appendix A. The relative merits of sequential-vs-global-inversion methods for the INS problem are discussed briefly in Appendix B. Functional limitations imposed by the physics of the problem are discussed in Sec. IV. The fold and unfold characteristics of the physical functions we determine experimentally are investigated in Sec. V and the problems encountered in treating the convolution product g as though it were a convolution square are discussed in Sec. VI. Tests devised for the genuineness of peaked features in the convolution square root are presented in Sec. VII. Effects of amplitude distortions of f are presented in Sec. VIII, and the effects of origin error, tests for its presence, and means for its elimination, in Sec. IX.

It is evident that the material of this paper is central to the successful execution of the method of INS and the demonstration of its viability. This is both a physical and a mathematical problem. We think it particularly interesting how the physics of a two-electron transition process is so inextricably interrelated to the mathematics of the convolution

integral. INS has shown itself to be an informative tool for the study of electronic states and band structure, particularly at the surfaces of solids. We are particularly anxious to document properly the sound physical and mathematical basis of the method. Previous discussions of this aspect of our work have been mentioned in earlier papers[1-4] but these presentations are at best fragmentary and some of the earliest possibly misleading. This makes the material of this present paper an essential part of the development of the INS method.

II. INS METHOD AND ITS ASSUMPTIONS

We shall discuss briefly the method and assumptions of INS. INS is an Auger-type electron-emission spectroscopy based on the two-electron transition process shown in Fig. 1. In this section we shall use the notation used in previous publications indicating later its relation to that used elsewhere in this paper. Two band electrons initially at energies $\zeta - \Delta$ and $\zeta + \Delta$ participate in the process. One electron drops to the initially vacant ground state of an ion (normally He^+) presented immediately outside the solid surface. The second electron becomes, at energy E, an excited electron or secondary which may leave the solid. The external-energy-distribution function $X(E)$, which is measured for these electrons, contains the "spectroscopic" information we wish to extract.

The complication in the case of a two-electron process, in comparison with a one-electron process such as photoemission, arises because the final energy level at E corresponds not to a single initial energy level in the filled band of the solid, but to an infinity of paired levels. As can be seen from Fig. 1, all pairs of electrons in initial levels symmetrically disposed with respect to the level ζ, which lies halfway between the level E in question and the ground level in the atom, will yield an excited electron at E. If we assume transition probability to be independent of energy ζ, then the probability of involvement of each electron in the process is proportional to the density of states $N(\zeta)$, the probability of the elemental process is $N(\zeta - \Delta)\,N(\zeta + \Delta)$, and the total probability for final energy E is[5]

$$F_c(\zeta) = \int_{-\zeta}^{\zeta} N(\zeta - \Delta) N(\zeta + \Delta)\, d\Delta \ . \tag{3}$$

If we change variables from ζ to E, using the expression

$$E = E_i'(s_t) - 2(\zeta + \varphi) \tag{4}$$

derived by equating the energy changes of the up and down transitions in Fig. 1, we obtain $F(E)$, the distribution of excited electrons inside the solid. The measured external distribtuion $X(E)$ is obtained by inclusion of broadening and by multiplying by the escape probability $P(E)$. Thus, the broadened internal distribution is

$$F_b(E) = \int_{-\infty}^{\infty} B(t,\, L)\, F(E - t)\, dt \,, \tag{5}$$

in which $B(t, L)$ is the broadening function, whose breadth is specified by the parameter L. Finally,

$$X(E) = F_b(E) P(E) \ . \tag{6}$$

INS reverses the above procedure, as discussed in detail elsewhere.[1] The effect of broadening is essentially removed by extrapolating two measured kinetic-energy distributions $X_{K_1}(E)$ and $X_{K_2}(E)$, taken at two ion kinetic energies K_1 and K_2, to a function $X_0(E)$ on the basis that the broadening parameter L of Eq. (5) varies as the velocity of the ion.[6] $F(E)$ is then obtained as

$$F(E) = X_0(E)/P(E) \ . \tag{7}$$

Change of variable via Eq. (4) gives us $F(\zeta)$ which is the function to be deconvolved.

If, in fact, transition probability was energy independent, we see from Eq. (3) that inversion of $F_c(\zeta)$ would yield $N(\zeta)$, the density of states in the solid. Inclusion of the all-important nonzero transition-probability factors means that Eq. (3) must be rewritten as

$$F(\zeta) \propto \int_{-\zeta}^{\zeta} |\, H_{fi} \,|^2 N(\zeta - \Delta)\, N(\zeta + \Delta)\, d\Delta \ . \tag{8}$$

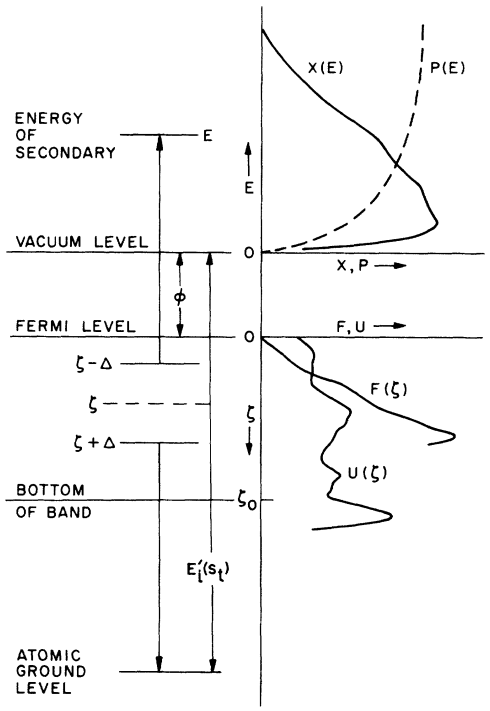

FIG. 1. Electron-energy diagram illustrating energy levels, electronic transitions, and distribution functions of the two-electron process upon which INS is based.

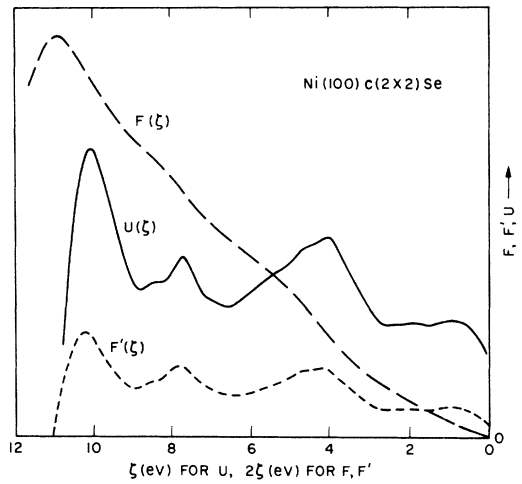

FIG. 2. Plots of $F(\zeta)$, $U(\zeta)$, and $F'(\zeta)$ obtained from data for He^+ ions incident on a Ni $(100)c(2\times 2)$Se surface by the method of INS (Ref. 4).

The basic assumption concerning the matrix element made in the INS procedure is that the energy dependences of $|H_{fi}|^2$ are included essentially correctly when it is written as

$$|H_{fi}|^2 = H_1(\zeta - \Delta)H_2(\zeta + \Delta)\,, \tag{9}$$

the product of two factors, not necessarily equal in magnitude, which depend on the initial energies of the two participating electrons. If $H_1 = H_2$, the H factors may be subsumed with the state-density factors of Eq. (8) to give $F(\zeta)$ as a convolution square,

$$F(\zeta) = \int_{-\zeta}^{\zeta} U(\zeta - \Delta)\,U(\zeta + \Delta)\,d\Delta = U * U = U^{(2)}\,, \tag{10}$$

of a function $U(\zeta)$ which we have termed the transition density. Equation (10) has formed the basis of the INS procedure. If, however, $H_1 \neq H_2$, as might be expected in general, $F(\zeta)$ must be replaced by the convolution product

$$G(\zeta) = \int_{-\zeta}^{\zeta} V(\zeta - \Delta)\,W(\zeta + \Delta)\,d\Delta = V * W\,, \tag{11}$$

which cannot be inverted given $G(\zeta)$ only.

Actual data obtained in the INS experiment are presented in Figs. 2 and 3. In each figure we show the $F(\zeta)$ function derived as discussed above from measured kinetic-energy distributions of ejected electrons. Also in each figure we show unfold functions $U(\zeta)$ and the derivative $F'(\zeta) = dF(\zeta)/d\zeta$ of the fold function. We observe, in Fig. 2 in particular, how faithfully F' reproduces the peak structure in U. This is the more surprising when one recalls that whereas F' is a local function depending on F only in a local region, U is a nonlocal function depending at given ζ on all F values in the range zero to ζ. In Fig. 3, F' and U differ in a significant but completely understood manner. Comparison of F' and U is the cornerstone of our testing for errors in U.

The symmetrical forms for F and G given in Eqs. (10) and (11) are those which arise most naturally in the physical discussion of the ion-neutralization process. However, we find it somewhat more convenient to use the mathematically more common and equivalent asymetrical forms of Eqs. (1) and (2). If $u(x) \equiv U(\zeta)$, $f(x) = F(\frac{1}{2}\zeta)$; or if $f(x) \equiv F(\zeta)$, $u(x) = U(2\zeta)$, with similar expressions among g, G, v, V, w, and W. Since $\zeta < 0$ corresponds to unoccupied states above the Fermi level, all of the functions F, G, U, V, W are zero there and positive definite for $\zeta > 0$, thus meeting the conditions required of g, f, u, v, and w if Eqs. (1) and (2) are to be valid.

III. DECONVOLUTION PROCEDURE USED

In our work with INS we have used a simple digital-inversion formulation based on writing $g(x)$

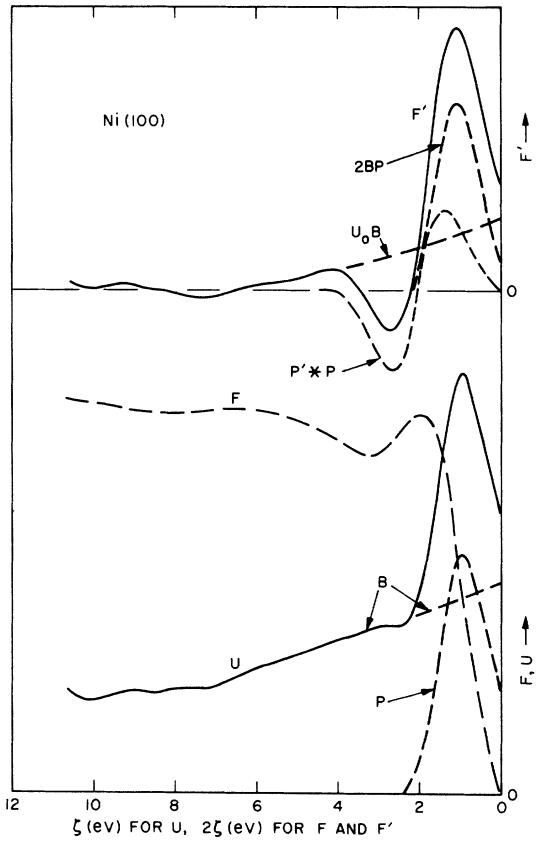

FIG. 3. Curves for $F(\zeta)$, $U(\zeta)$, and $F'(\zeta)$ for the He^+ ions incident on a clean Ni (100) surface as determined by INS (Ref. 2). We illustrate resolutions of the F' curve into U_0B, $2BP$, and $P'*P$ components and the U curve into B and P components.

TABLE I. Step-midpoint formulas for folding and unfolding (f_0, $g_0 = 0$ in each case).[a]

(A) Convolution square

$$f_n = h \sum_{p=0,\ n-1} u_{n-p-1} u_p, \qquad n \geq 1$$

$$F_n = 2\Delta\zeta \sum_{p=0,\ n-1} U_{2n-2p-2}\, U_{2p}, \qquad n \geq 1$$

(B) Convolution square root

$$u_0 = (f_1/h)^{1/2}$$

$$u_1 = (1/2u_0)(f_2/h)$$

$$u_{n-1} = \frac{1}{2u_0}\left(\frac{f_n}{h} - \sum_{p=1,\ n-2} u_{n-p-1} u_p\right), \qquad n \geq 2$$

$$U_0 = (F_1/2\Delta\zeta)^{1/2}$$

$$U_2 = (1/2U_0)\ (F_2/2\Delta\zeta)$$

$$U_{2n-2} = \frac{1}{2U_0}\left(\frac{F_n}{2\Delta\zeta} - \sum_{p=1,\ n-2} U_{2n-2p-2} U_{2p}\right), \qquad n \geq 2$$

(C) Convolution product

$$g_n = h \sum_{p=0,\ n-1} v_{n-p-1} w_p, \qquad n \geq 1$$

$$G_n = 2\Delta\zeta \sum_{p=0,\ n-1} V_{2n-2p-2} W_{2p}, \qquad n \geq 1$$

(D) Convolution factor (given w, W)

$$v_0 = g_1/hw_0$$

$$v_1 = (1/w_0)\ (g_2/h - v_0 w_1)$$

$$v_{n-1} = \frac{1}{w_0}\left(\frac{g_n}{h} - v_0 w_{n-1} - \sum_{p=1,\ n-2} v_{n-p-1} w_p\right), \qquad n \geq 2$$

$$V_0 = G_1/2\Delta\zeta W_0$$

$$V_2 = (1/W_0)\ (G_2/2\Delta\zeta - V_0 W_2)$$

$$V_{2n-2} = \frac{1}{W_0}\left(\frac{G_n}{2\Delta\zeta} + V_0 W_{2n-2} - \sum_{p=1,\ n-2} V_{2n-2p-2} W_{2p}\right), \qquad n \geq 2$$

[a]The equations given here for the F and U functions of the symmetrical formulation [Eqs. (3) and (4)] differ from those given in Ref. 1 for two reasons. First, there is the difference in the limits of the defining integral making the F of this paper twice that of Ref. 1. The form used here is preferred because it is the one which transforms directly into the symmetrical convolution product, where the limits $-\zeta$, ζ rather than 0, ζ are essential, and into the asymmetrical forms of Eqs. (1) and (2). Second, we have designated the digital U values here as U_0, U_2, ..., U_{2n-2}, rather than U_1, U_3, ..., U_{2n-1} as in Ref. 1, again for ease in shifting between the asymmetrical and symmetrical forms.

in digital form g_n, $n = 0, m$ ($g_0 = 0$), as the Cauchy product of the digital representations v_n, $n = 0, m$, and w_n, $n = 0, m$. Thus,

$$g_n = h \sum_{p=0,\ n-1} v_{n-p-1} w_p, \qquad n \geq 1 . \tag{12}$$

This is the form obtained by approximating $v(x)$ and $w(x)$ in Eq. (1) by step functions centered on the digital values. We have called it the "step-midpoint" formulation to distinguish it from other possible formulations discussed in Appendix A. The digital expressions for folds and unfolds are given in Table I. In general, we are given an equally spaced digital representation of a portion of a function $f(x)$ in the range $0 \leq x \leq x_m$. Thus we are given f_n, $n = 0, m$, at the points $x = nh$, $n = 0, m$, with $x_m = mh$.

In Eqs. (1) and (2) we have used the common shorthand notations $g = v * w$ and $f = u * u$ and have introduced $f = u^{(2)}$ as an alternative. We also suggest the convenient forms

$$u = f^{(1/2)}, \tag{13}$$

$$v = g /\!\!/ w \tag{14}$$

as expressions for the convolution square root of a convolution square and the convolution factor of a convolution product, respectively. In this shorthand the function obtained by inverting a convolution product $v * w$ as though it were a convolution square is written $(v * w)^{(1/2)}$. For smoothly varying functions v and w, $(v * w)^{(1/2)}$ is a sort of "convolution mean" of v and w and we shall call it that.

IV. FUNCTIONAL LIMITATIONS IMPOSED BY PHYSICS OF AUGER NEUTRALIZATION PROCESS

In common with many problems in physics, particularly those involving an inversion, the problem of interpretation in INS is "incorrectly posed."[7] According to the definition of this term used by the Soviet authors this is so because the experimentally determined $g(x)$, the input to Eq. (1), is distorted due to noise, kinetic-energy analyzer characteristics, and the effects of the first steps in the data-reduction procedure in which $g(x)$ is obtained from the measured kinetic-energy distribution. The most questionable of these first steps is the extrapolation of $X(E)$ used to circumvent inversion of Eq. (5).[1] But in addition our problem is "incorrectly posed" because we substitute $f = u * u$, Eq. (1), for $g = v * w$, Eq. (2), and because no inversion procedure is exactly accurate except for limiting functions in some cases.

It is reassuring at this point to note that we are by no means in a unique position. Many, perhaps most, problems in physics are incorrectly posed in one way or another. Most theoretical formulations either leave out or approximate troublesome terms. Experimental distortion of data is well known. But it is also true that this situation can

be more serious, in fact at times catastrophic, in an *inverse* as opposed to a *direct* problem in physics.[7]

The procedure by which a physically meaningful answer is to be obtained in INS has three indispensable components. It must (i) use some *a priori* knowledge of the result, (ii) provide a significant test which is capable of detecting error in the answer, and (iii) provide means for manipulating the given data so as to remove the principal experimental error. Our problem is made tractable by the fact that the physics restricts us in many cases to a specific class of fold and unfold functions which can be defined with sufficient precision for our purposes. These are functions which lie near the limiting functions $g(x)=x$, $v(x)=C$, $w(x)=1/C$ (C, a constant), $x>0$, or in the convolution square formulation, $f(x)=x$, $u(x)=1 (g=f=u=v=w=0$, $x<0)$. The departure from this functional limit which the physics allows is, in many cases, in the direction of a specific class of u, v, w functions we shall term "peaked" functions, not unlike $U(\zeta)$ in Fig. 2, consisting of a relatively smooth background which cuts off sharply at the origin as does $u(x)=1$ and upon which are superposed one or more peaks, each of full-base width 0.1–0.3 of the total extent of the argument $0<x<x_m$.

In the case of the convolution product the v and w functions can have smooth backgrounds and peaks of different magnitudes but each must exhibit peaks only at positions where peaks appear in both functions. The smooth background of the peaked function corresponds to the s, p bands of the bulk solid and the peaks are "resonances" due to tight-binding bulk bands—d bands, for example—or electrons in surface orbitals. We expect that the same peaks will appear in both the v and w functions since they have the same physical origin in each. The functional limitations which the physics imposes constitute the *a priori* information we need to solve the problem.

It is of significance to note that the sequential inversion procedure of Sec. III yields an absolutely correct answer for the limiting case $f(x)=x$, $u(x)=1$, an advantage not possessed by other inversion formulations. The $u(x)$ function of the limiting set $f(x)=x$, $u(x)=1$, $x>0$; $f=u=0$, $x<0$ corresponds to the Fermi function at zero absolute temperature and thus to a transition density at this temperature for constant density of states and constant transition probability.

We must recognize that for some solids or surface preparations either the local state density or transition-probability components of $u(x)$ will produce deviations from our limiting function $f(x)=x$ and $u(x)=1$ by amounts which are so large as to put $u(x)$ outside the class of peaked functions for which the procedures of this paper are satisfactory. A possible example is a semiconductor with filled low-density surface states. This results in $u(x)$ having a very low value at $x=0$ (the Fermi level). The methods of this paper require a "reasonable" step in $u(x)$ at the Fermi level. Functions not possessing this can undoubtedly be inverted by one means or another but then different *a priori* information concerning the answer and a different testing procedure will be needed to solve the problem. In any event the testing procedure of this paper are sufficient to indicate when the given function lies outside the limits within which a correct answer can be obtained by the methods of this paper.

At this point it would be well to emphasize the distinction which must be made between error introduced by the inversion and error introduced because the starting function $f(x)$ is in some way or other incorrect. If the data are smooth enough to avoid point-by-point runaway of the sequential inversion, that is, to avoid successive rapidly increasing positive and negative deviations of the calculated u_n from the correct mean, then the relatively smooth u_n sequence obtained is very close to the correct mathematical deconvolution of the given f_n. It is not necessarily the correct physical solution to the problem, however. To find and demonstrate the correct physical solution requires the tests and data manipulation described in this paper.

We have reduced the noise in the input f_n data by averaging 10–15 runs in a multichannel scaler. In addition it is customary for us to smooth over five points three times during the digital calculations which interpolate, normalize, and invert the input data. This smoothing is the equivalent of the smoothing inherent in a global method which uses approximately 16 harmonics in the expansion of the functions described by 250 digital data points.

V. CHARACTERISTICS OF FOLD OF A "PEAKED" UNFOLD FUNCTION

In this section we shall give a brief analysis of the structure of $f(x)$ and $f'(x)$ for a so-called peaked $u(x)$ function. It is most convenient to do this for $u(x)$ in which a single symmetrical peak $p(x)$, centered at $x=a$ of full-base width $2w$, is placed on a smoothly varying background $b(x)$. We write this as

$$u=b+p\ . \tag{15}$$

Such a function for *constant* background is shown in Fig. 4(a). Putting Eq. (15) into Eq. (2), we obtain

$$f=b*b+2b*p+p*p\ . \tag{16}$$

The three terms in this expression are plotted as curves 1, 2, and 3, respectively, in Fig. 4(b) for the u of Fig. 4(a). $b*b$ is the fold of the background over itself, $b*p$ is the fold of the peak over

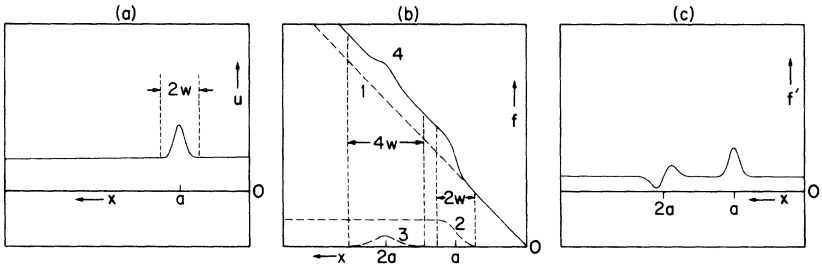

FIG. 4. Plots of unfold u at (a), fold f at (b), and derivative f' of the fold at (c) for a simple $u(x)$ consisting of a constant background with a Gaussian-like peak centered at a. The $b*b$, $2b*p$, and $p*p$ components, curves 1, 2, 3, respectively, add to curve 4 for f in part (b). The derivative of the $p*p$ feature at $2a$ in f' is seen. Note that u has no feature at $2a$.

the background, and $p*p$, the fold of the peak over itself. If p is symmetrical and of full-base width $2w$, $p*p$ is also a symmetrical peak of full-base width $4w$. If $b=1$, the asymptotic value of $2b*p$ is twice the area of the peak p, and the maximum value of $p*p$ is the area of p^2.

Next we consider f'. Corresponding to Eq. (2) we may write

$$f'(x)=u(0)u(x)+\int_0^x u'(x-t)\,u(t)\,dt \tag{17}$$

or

$$f'=u_0u+u'*u\ . \tag{18}$$

Inserting Eq. (15) into (18), we obtain

$$f'=u_0(b+p)+b'*u+p'*b+p'*p\ . \tag{19}$$

Since b is assumed smooth and featureless, $b'\sim 0$, and the $b'*u$ term will be dropped. The $p'*b$ term may be simplified by use of the equation

$$u'(x)*S(x)=u(x)\ , \tag{20}$$

in which $S(x)$ is a unit step function of the form $S(x)=1$, $x>0$; $S(x)=0$, $x<0$. Since $b(x)\cong b_0S(x)$, it is clear that $p'*b\cong p'*b_0S(x)=u_0p$, assuming $p_0=0$ and $u_0=b_0$. Thus Eq. (19) reduces to

$$f'=u_0b+2bp+p'*p\ . \tag{21}$$

In Eq. (21) the first two terms of f' reproduce the background b and the peak p of u, Eq. (15), but with altered relative magnitudes. The third term $p'*p$ can be seen to have the same general form as p' but total base width $4w$ instead of $2w$; i.e., it resembles $p'(\frac{1}{2}x)$. Functions which we take to have the same form or to resemble one another have the same number of zeros and positive and negative portions in the same relative positions. That this is true for $p'*p$ and $p'(\frac{1}{2}x)$ may be seen by substituting p' and p into Eq. (1) and observing what the integration of the convolution produces. f' is plotted in Fig. 4(c) for the f of Fig. 4(b).

We see that the peak at a in f' is proportional in magnitude to the peak at a in u, but that the $p'*p$ feature in f' at $2a$ is proportional to the square of the magnitude of the peak. For p sufficiently small with respect to b and/or sufficiently broad, it is possible for the $p'*p$ term in f' to be essentially invisible, in which case f' will show only features of the form $2bp$ above background and will show the same peaks as u, as is the case in Fig. 2.

In Fig. 3, on the other hand, the main peak in u is of such strength as to produce a clearly visible $p'*p$ term. Since in this case $a\sim w$, the $p'*p$ feature of width $4w$ placed at $2a$ will be half hidden under the peak p itself. We have indicated in Fig. 3 approximate resolutions of U and F' into their component parts.

It is clear that in each of the above cases we understand the structure of f' and can use it as a test of the validity of the peaks to be seen in u. The above characteristics result from the fact that our examples lie reasonably close to the limit $f(x)=x$, $u=1$ for which $f'=u$, and that the functional departures from this limit are in the form of the peaked functions described above. In the above discussion of f' and u we have limited ourselves to a $u=b+p$ with a single peak. If $u=b+p_1+p_2+p_3+\cdots$, it is clear that f will have p_i*p_j terms. f' will then include $p_i'*p_j$ terms. However, f' will have bp_i terms at the positions of the p_i peaks in u as before and the $p_i'*p_j$ terms will in general be small when the $p_i'*p_i$ terms are.

VI. DEPARTURE FROM $f=u*u$ TO $g=v*w$

We now face up to the possibility that the function we are given to deconvolve is actually a convolution product $g=v*w$ and not a true convolution square $f=u*u$. As we have seen, we expect the two transition densities v and w to have the same general form assumed for u, namely, $b+p_1+p_2+p_3$, etc. We expect both v and w to have relatively smooth backgrounds with peaks due to resonances in the same positions but of possibly differing relative magnitudes. This severe but reasonable restriction on the class of functions v and w makes it possible to extract from g a function of the form $b+p_1+p_2+p_3$ representing the convolution mean between v and w and revealing the structure of resonances in the electronic states of the surface of the solid which is the goal of INS.

First, we shall dispose of trivial cases. If $v=b_1$ and $w=b_2$, each convolution factor is a smooth

background function with a step at $x=0$ but without peak features. It is then easy to demonstrate that no peak features will be generated by deconvolving g as though it were a convolution square. $g^{(1/2)}$ will be what we have termed the convolution mean between b_1 and b_2. An example of this shown in Fig. 18 of Ref. 1. It can also be shown that when a peak is placed on different smooth backgrounds, as in $v=b_1+p$ and $w=b_2+p$, $(v*w)^{(1/2)}$ will exhibit a peak p at the proper position on a background equal to the mean $(b_1*b_2)^{(1/2)}$. No spurious peaked features are introduced into $(v*w)^{(1/2)}$ by this difference in the smooth backgrounds of v and w. Another trivial case is that in which $w=\alpha v$. Then $g=\alpha(v*v)$ is a convolution square of the function $g^{(1/2)}=\alpha^{1/2}v$.

We shall carry on our discussion for single-peaked functions v and w like the u of Fig. 4(a). The most general difference between v and w we shall admit may be written

$$v=b+p\ ,\quad w=b+\alpha p\ . \tag{22}$$

Then,

$$g=v*w=b*b+(1+\alpha)b*p+\alpha p*p\ . \tag{23}$$

Rewriting (23) in a form which approximates that of $f=u*u$ [Eq. (16)] as closely as possible, we obtain

$$g=b*b+2[b*\tfrac{1}{2}(1+\alpha)p\,]+[4\alpha/(1+\alpha)^2\,]\times[\,\tfrac{1}{2}(1+\alpha)p*\tfrac{1}{2}(1+\alpha)p\,]\ . \tag{24}$$

This would have the form of $u*u$ with $u=b+\frac{1}{2}(1+\alpha)p$ if the coefficient $4\alpha/(1+\alpha)^2$ were unity. This, of course, would require $\alpha=1$. We see from (24) that g differs from a true convolution square only in the coefficient of the $p*p$ term. We shall term this an amplitude inconsistency of the $p*p$ and $b*p$ terms in the fold function $v*w$.

The most general form of $g=v*w$ which the physics of ion neutralization requires us to admit may thus be written

$$g=v*w=b*b+2b*p+kp*p\ , \tag{25}$$

in which k may be greater or less than unity but is always positive. g may also be written in the form

$$g=v*v+(k-1)p*p\ . \tag{26}$$

For functions with these limitations let us now consider g' and $g^{(1/2)}$. Taking the derivative of both sides of Eq. (26), we obtain

$$g'=v_0v+v'*v+(k-1)(p_0p+p'*p)\ . \tag{27}$$

Using $p_0=0$, $b'\cong 0$, $p'*b\cong b_0p$ as was done in deriving Eq. (21) we obtain

$$g'=v_0(b+2p)+kp'*p\ . \tag{28}$$

Thus g' resembles the derivative of a true convolution square, f' of Eq. (21), differing only in the magnitude but not the sign of the $p'*p$ term. We are interested in the form of g' for its use in testing for spurious features in $g^{(1/2)}$.

The really important question concerns the form of $g^{(1/2)}=(v*w)^{(1/2)}$, the convolution square root of what is really a convolution product. Our discussion continues for a single peak of full-base width $2w$ placed at a where $a\gg w$. The condition $a\gg w$ guarantees that the features of higher order produced at na will not overlap for at least the first several orders, making our discussion in terms of peaked functions easier. Since g of Eq. (25) has the form of a true convolution square in the $b*b$ and $2b*p$ terms, sequential unfolding gives $b+p$ out to the point $2(a-w)$ where the $p*p$ feature in g is encountered. Clearly the folding of p over p will produce $p*p$, not the required $kp*p$, so a new feature q_2 must appear in $g^{(1/2)}$ near $x=2a$ of such form and magnitude as to correct $p*p$ in the convolution product g to $kp*p$. Thus for $x<2(a+w)$,

$$g=g^{(1/2)}*g^{(1/2)}=b*b+2b*p+p*p+2b*q_2\ . \tag{29}$$

Comparison with Eq. (25) yields

$$b*q_2=b_0S(x)*q_2=\tfrac{1}{2}(k-1)p*p\ , \tag{30}$$

in which we have again used the fact that, near $x=0$, $b\cong b_0S(x)$. Using Eq. (20), we obtain

$$q_2=\tfrac{1}{2}[(k-1)/b_0](p*p)' \tag{31}$$

as the form of the feature near $x=2a$ developed in $g^{(1/2)}$ because g is not a true convolution square. q_2 has breadth $4w$.

Because $g^{(1/2)}$ differs from b by q_2 near $x=2a$ it can readily be shown that a sequence of features q_n of breadth $2nw$ at $x=na$ will be produced in $g^{(1/2)}$. We derive the relations

$$q_3=-(p*q_2)'=-[\tfrac{1}{2}(k-1)/b_0\,][p*(p*p)'\,]' \tag{32}$$

and

$$q_n(x)=-[(g^{(1/2)}-b)*(g^{(1/2)}-b)\,]'/b_0\ . \tag{33}$$

$q_n(x)$ may be developed sequentially using

$$g^{(1/2)}=b+p+\sum_{n\geq 2}q_n\ . \tag{34}$$

This procedure will be an adequate representation to the x value at which the features q_n begin to overlap each other.

It is possible to discuss the general form of the feature $q_n(x)$. As was shown for q_2 and q_3, the general expression (33) can be reduced to expressions involving the derivatives of the convolutions of p, i.e., $(p*p)'$ and derivatives of convolutions of these with p, $[p*(p*p)'\,]'$, etc. If we use the fact

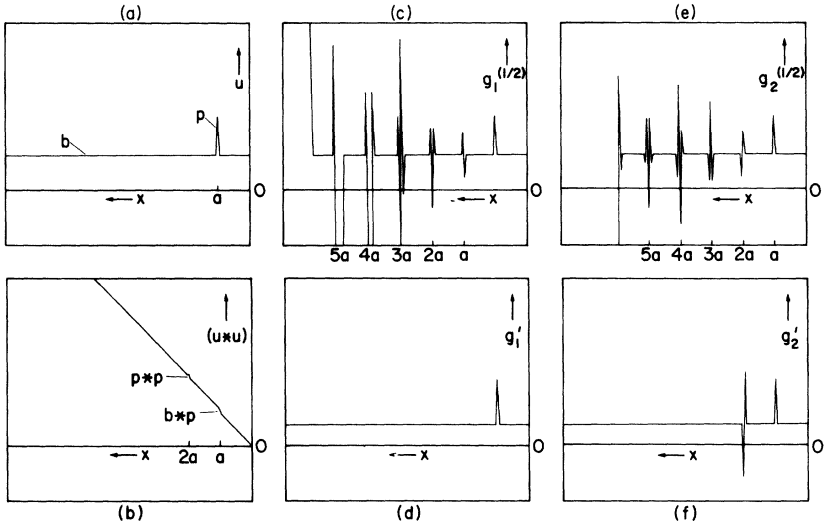

FIG. 5. Plots of a $u(x)$ function consisting of a uniform background b, and a "one-point" peak p, at $x=a$ and its fold $f(x)$ at parts (a) and (b), respectively. When $f(x)$ is distorted by removal of the $p*p$ feature, $k=0$, one obtains the inversion $g_1^{(1/2)}$ at (c) and the derivative of the mutilated fold g_1' at (d). When the $p*p$ feature in f is doubled, one obtains the unfold $g_2^{(1/2)}$ at (e) and the derivative g_2' at (f).

that for our restricted functions $p_0=0$, we can show by Eq. (17) that $(p*p)'=p'*p$. As we have seen earlier $p'*p$ has the same form as $p'(\frac{1}{2}x)$. By an extension of this line of argument we see that $-[p*(p*p)']'=-[p*(p'*p)]'$ which will have the form of $-[p*p'(\frac{1}{2}x)]'$ which in turn has the form of $-p''(\frac{1}{3}x)$. In general, then, the form of the nth spurious feature $q_n(x)$ will be

$$\frac{(1-k)(-1)^n d^{n-1}p(x/n)}{dx^{n-1}}$$

for $n\geq 2$.

Examples of $g^{(1/2)}$ functions obtained from the fold of a u function which has a "one-point" peak at $x=a$ are shown in Fig. 5. We note that the structures $q_n(x)$, $n\geq 2$, are those predicted above. Examples for peaked functions with peaks of magnitudes differing by a factor 5 are shown in Fig. 6 for two peak-to-background ratios. In Fig. 7, we have constructed v and w functions from a sloping background and three Gaussian peaks of 5 to 1 ratio. Note that $(v*w)^{(1/2)}$ reproduces the peaks very well and that only very minor spurious undulations occur in the range $x>4$.

Let us now compare terms above background in g' and in $g^{(1/2)}$ near $x=2a$. From Eq. (28) we obtain in g' the term $kp'*p$ which has the form of $p'(\frac{1}{2}x)$ as we have seen. The q_1 term in $g^{(1/2)}$ is of the form $(1-k)(-1)^1 p'(\frac{1}{2}x)$ as discussed above. Thus the corresponding terms in g' and $g^{(1/2)}$ will have the same form if $k>1$ but one will be the negative of the other if $k<1$. This also is illustrated in Fig. 5.

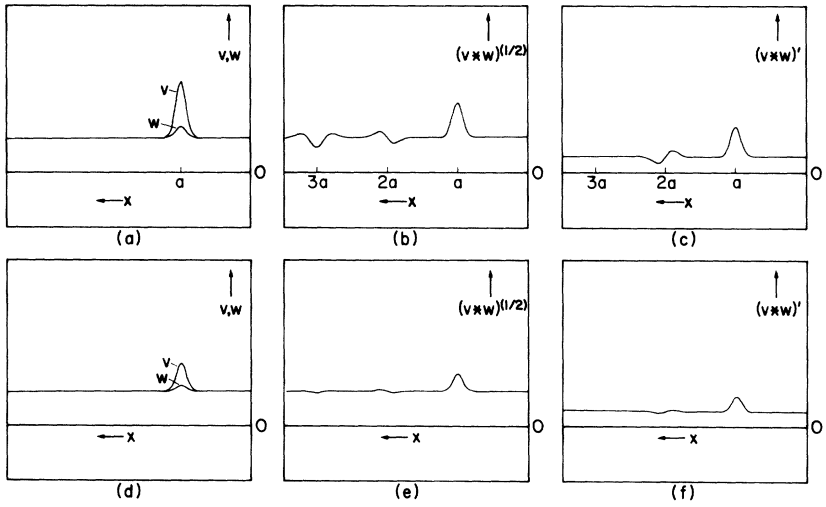

FIG. 6. Plots illustrating spurious features produced by magnitude inconsistency of the $2b*p$ and $p*p$ features for a $u(x)$ having a Gaussian-like peak centered at $n=46$ sitting on a constant background. (a) v and w functions with peaks differing in magnitude by a factor of 5; (b) $(v*w)^{(1/2)}$, the convolution square root of the convolution product of v and w shown at (a); (c) $(v*w)'$, the derivative of the convolution product of v and w shown at (a); (d) v and w having smaller peaks again in the ratio 5:1; (e) $(v*w)^{(1/2)}$, the convolution square root of the convolution product of v and w shown at (d); (f) $(v*w)'$, the derivative of the convolution product of v and w shown at (d).

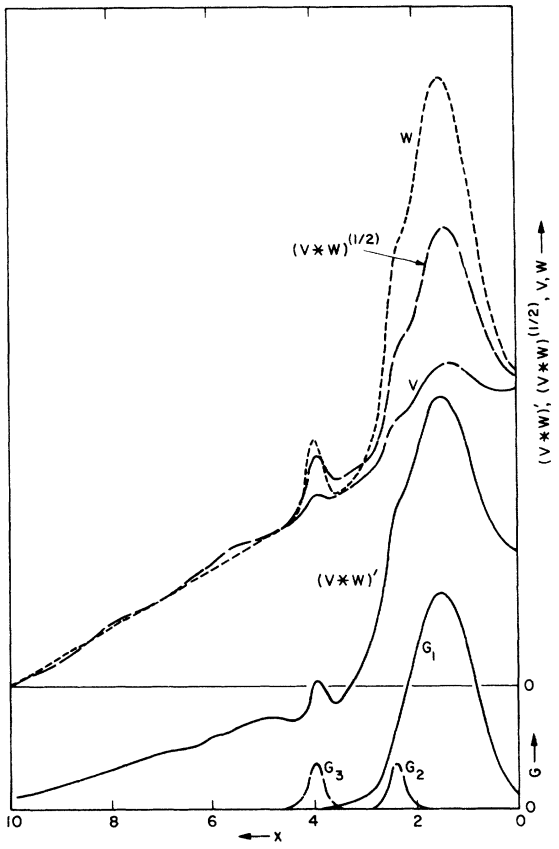

FIG. 7. Plots showing the effect of amplitude inconsistency for an example which resembles what one could get as a result from INS. The dashed lines at the top of the figure indicate the $v(x)$ and $w(x)$ functions. $v(x)$ and $w(x)$ are compounded of the straight line $0.2-0.02x$ to which have been added $\frac{1}{3}$ and $\frac{5}{3}$ times the Gaussian-like functions shown at the bottom of the figure, respectively. $(v*w)^{(1/2)}$ is the convolution square root of the convolution product of v and w. Also shown in the middle of the figure is $(v*w)'$, the derivative of the convolution product. Note the small spurious structure developed at $x > 4$ in both $(v*w)^{(1/2)}$ and $(v*w)'$.

VII. TESTS FOR GENUINENESS OF PEAKED FEATURES IN $(v*w)^{(1/2)}$

We now return to our main theme. It is to determine what spurious features appear when we perform the operation $g^{(1/2)}$ where $g = v*w$ with $v = b + p_1 + p_2 + p_3 + \cdots$ and $w = b + \alpha p_1 + \beta p_2 + \gamma p_3 + \cdots$. For each peak we know that $g^{(1/2)}$ will reproduce the background and the peak itself at a but a sequence of features resembling higher derivatives of the peak will appear at $2a$, $3a$, $4a$, etc. Now it is clear that it is prudent to restrict ourselves to functions having only a few peaks as in Fig. 2. Then there is definite promise that one can verify whether a peak in $g^{(1/2)}$ is a reproduction of a peak in v and w or a spurious feature produced by the fact that g is really a convolution product and not a convolution square.

It is now possible to state a procedure whereby one can decide as to the genuineness of peaked structures in $g^{(1/2)} = (v*w)^{(1/2)}$.

(i) As we proceed from the origin in the function $g^{(1/2)} = (v*w)^{(1/2)}$ it is evident that the first peak cannot be spurious. For peaked functions a spurious feature is always the offspring of a bona fide feature lying $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, etc., of its distance from the origin.

(ii) From the form of the first bona fide peak $p_1(x)$, we can predict the form of the first possible spurious feature in $g^{(1/2)}$. It must look like $p_1'(\frac{1}{2}x)$ and appear either directly or inverted in g'. Thus if the second peak encountered in $g^{(1/2)}$ appears in g' but does *not* meet the form requirement we can consider it as bona fide.

(iii) If all peaks in g' also appear in $g^{(1/2)}$ we can conclude that the $kp'*p$ terms in g' [Eq. (28)] are too small to be observed and that, therefore, all peaks observed in $g^{(1/2)}$ are bona fide representations of a mean function between v and w representing a mean transition density function for the up and down electrons in the Auger process. We consider this procedure to be reliable for a limited number of peaks, possibly four or five as one proceeds away from the origin.

(iv) As indicated in the discussion of Fig. 3, a feature can appear in g' which does not appear in the direct or inverted form in $g^{(1/2)}$. This in general will require, as the above analysis indicates, that $g \sim f$ with $k \sim 1$. Thus the feature in g' will be the required $p'*p$ term of Eq. (21) which is present to prevent the appearance of a feature at $2a$ in u.

VIII. EFFECTS OF AMPLITUDE DISTORTION OF $f(x)$

In this section we shall discuss the effects of amplitude distortion of the parts of $f(x)$ which must be amplitude consistent with one another if $f(x)$ is to be a true convolution square. In terms of our example $u = b + p$, $f = b*b + 2b*p + p*p$, we see that such amplitude distortion will make the $p*p$ term inconsistent in amplitude with the $2b*p$ term. Thus f really has become a g function like that of Eqs. (25) and (26) and the discussion of errors given above for this case apply here as well. The tests for spurious features produced by amplitude error are thus the same as those listed at the end of Sec. VII.

In Fig. 8, we give another example of the effect of amplitude inconsistency. Here we achieve amplitude inconsistency of the fold function by adding a smoothly varying function directly to a true convolution square before deconvolution. Curve 1 in Fig. 8 is an unfold function constructed of the three Gaussian-like peaks of Fig. 7 plus a linearly decreasing background. Curve 2 is the fold of curve

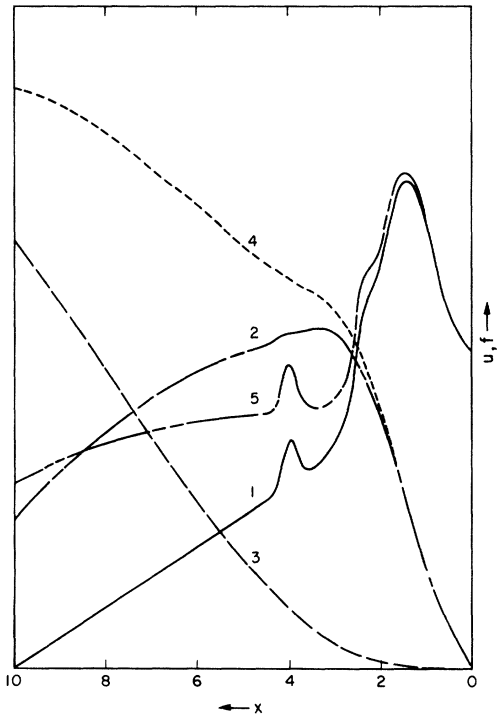

FIG. 8. Graph showing the effect on a convolution square root of adding a smoothly varying function, curve 3, which is zero at the origin, to a true convolution square, curve 2. This yields the combination function, curve 4, whose convolution square root is curve 5. Curve 1 is the convolution square root of curve 2. The construction of curves 1 and 3 is discussed in the text.

1. Curve 3 is a smoothly rising function which starts from zero at the origin. It has specific physical interest and is derived as described below. When curve 3 is added to curve 2, we obtain curve 4 which is then deconvolved as a convolution square to obtain curve 5. We see illustrated the important result that the peak structure in curve 5 is the same as that in curve 1; only the level of the smooth background of the function at larger x has been affected.

The result of Fig. 8 is an important one for INS because it demonstrates that the process of deconvolution "picks out" the true self-convolution features of the function to be deconvolved and ignores a smoothly varying modification of the convolution as far as peak structure is concerned. A modification of the data similar to this could arise in INS by virture of the use of an incorrect escape probability curve $P(E)$, by which the data are divided at one point. A second way that the convolution could be modified in INS in the manner of Fig. 8 is by the addition of either a three-electron Auger component or the energy degradation of some of the excited electrons by collision with other band electrons.

The problem can be set up and solved on the basis of the following assumptions. Suppose the initial states of the electrons to lie at x_1 and x_2 with $x_2 > x_1$, x being quantized to the values $x = nh$. We then assume that the energy released on ion neutralization is shared by two electrons such that their energies *above the Fermi level* are y_n and y_{m-n}, where n varies from 0 to m. We further assume that all pairs of such final excited states are equally probable. Clearly, we can achieve this double excitation either by direct double excitation in a three-electron Auger process or by energy sharing between the one excited electron and a third electron subsequent to a two-electron Auger process. The fold function for such processes can be written down. It is

$$f_n = 2h \sum_{m=1,n-1} \left(\sum_{p=0,m-1} u_{m-p-1} u_p \right) \left(\sum_{k=0,n-m-1} u_k \right) . \tag{35}$$

This product series may be summed to give the result

$$f_n = \tfrac{1}{3}(n^3 - n) = u * u * u , \tag{36}$$

plotted as curve 3 in Fig. 8. Curve 3 is normalized by the factor 10^{-3} before addition to curve 2. Thus curve 4 is $(u * u) + 10^{-3}(u * u * u)$, u being curve 1.

IX. EFFECTS OF ORIGIN ERROR AND MEANS FOR ITS ELIMINATION

We discuss now a second form of degradation of the experimental data from a true convolution square, namely, that due to origin error. The origin of the $f(x)$ or $F(\zeta)$ function is obscured in the physical data by energy broadenings of the kinetic-energy distributions of the ejected electrons, and the procedure of INS may yield an origin which is somewhat in error. Origin error results in a position inconsistency between the $2b * p$ and the $p * p$ terms in f.

Although a peaked function with a "one-point" peak like that in Figs. 5(a) and 9(a) gives a distorted view of what is seen in practice, such functions are useful pedagogically because for them spurious features are large and remain separate and distinct to reasonably high order. In Fig. 9, we see what happens when the $2b * p$ and $p * p$ features of $u * u$ are shifted to the left or right along $b * b$. These mutilations of g produce functions whose convolution square roots are shown at Figs. 9(b) and 9(c), respectively. If, for example, the $2b * p$ feature is moved from a to $a + 10$ and its $p * p$ feature from $2a$ to $2a + 10$, these terms are position inconsistent because the $p * p$ term should lie at $2a + 20$. When sequential deconvolution of such a function is attempted, a spurious feature will appear at $2a + 10$ in $g_{+10}^{(1/2)}$ because the misplaced $p * p$ feature is to be found there, and another spurious feature will appear at $2a + 20$ because no $p * p$ feature is present

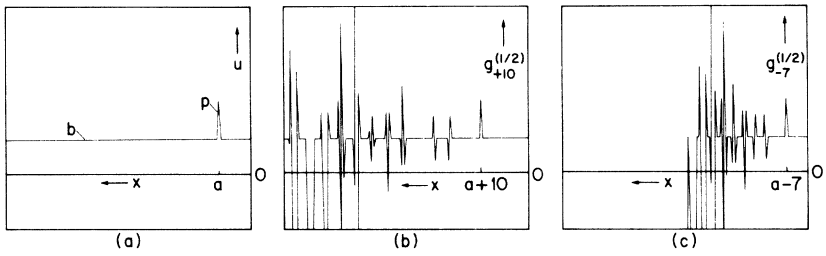

FIG. 9. Plots at (a) of a $u(x)$ function consisting of a uniform background b, with a one-point peak p, at $x=a$; at (b) of the unfold $g^{(1/2)}_{+10}$ of a fold g_{+10} obtained from $f=u*u$ by sliding the $2b*p$ and $p*p$ features of f a distance 10 points to larger x; at (c) of the unfold $g^{(1/2)}_{-7}$ of a fold g_{-7} obtained from $f=u*u$ by sliding the $2b*p$ and $p*p$ features of f a distance 7 points to smaller x.

there ($k=0$ in the analysis of Sec. IV). It can be seen that the spurious features at $2a+10$ and $2a+20$ have the $p'(\frac{1}{2}x)$ form and are the negative of each other as Fig. 9(b) shows. Furthermore, each of these spurious features will give rise to an infinite sequence of spurious features of higher-derivative form. Thus position inconsistency as a result of origin error leads to a more complicated sequence of spurious features than does amplitude inconsistency.

Examples of errors introduced by origin error for a broader peak are shown in Fig. 10 where the spurious features are resolved to at least second order. When the principal peak in u lies closer to the origin, the increasing orders of spurious features can coalesce into a remarkably sinusoidallike wave as is seen in Fig. 11. Figure 11 also illustrates how these errors decrease in magnitude and converge to zero when the principal peak height above background is reduced.

It is apparent that our data must be tested by origin shift for the presence of origin error. The correct origin of $f(x)$ will be that which removes the spurious structure present when the origin is in error and yields an unfold $u(x)$ whose structure at larger x is minimized and agrees with that of $f'(x)$. This origin sensitivity emphasizes the nonlocal character of $u(x)$ and the power and significance of the test comparison with the local function $f'(x)$.

We now define what we shall mean by an origin shift and present the method we have developed for achieving it. We are given a function $f(x)$, expressed digitally as f_n, which starts at $f_0=0$ at $n=0$. The corresponding x values are $x=nh$. We wish to shift the origin to the point $x_1=n_1h$, where n_1 may be positive or negative but is always an integer. We intend to accomplish this by replacing the values f_n from $n=0$ to $n=n_2$ by a new set of values f^r_n, $n=n_1, n_2$. The point $n=n_2$ is usually defined such that $f(x)$ passes through its first point of inflection between the points $n=n_2$ and n_2+1, i.e., the first difference $\Delta_n=f_{n+1}-f_n$ is maximum at $n=n_2$. We require (a) that the new set of data must leave the point at $n=n_2+1$ unchanged, i.e., $f^r_{n_2+1}=f_{n_2+1}$, and (b) that the differences of the first differences of the old and new data shall decrease linearly as n increases. By definition,

$$\Delta\Delta_n=(\Delta^r_n-\Delta_n)=(f^r_{n+1}-f^r_n)-(f_{n+1}-f_n)\,. \tag{37}$$

These conditions, together with the obvious requirement that $f^r_{n_1}=0$, lead to the expression for $\Delta\Delta_n$,

$$\Delta\Delta_n=(n_2-n+1)(\Delta\Delta_{n_1})/(n_2-n_1+1)\,, \tag{38}$$

and then to the new number sequence for $f(x)$,

$$f^r_n=f_n-f_{n_1}(n_2-n+1)(n_2-n+2)/(n_2-n_1+1)(n_2-n_1+2). \tag{39}$$

The above procedure of origin shift limits the distortion of the given data caused by origin shift to the initial portions near the origin and guarantees that the new function will join smoothly onto the old at its first point of inflection. We emphasize that we have used this procedure in INS only for very small origin shifts which amount at most to 0.3 eV out of a total function extent of about 12.0 eV and an interval to the first point of inflection of the order of 1.0 eV. When applied to a smooth background function by picking a point $n=n_2$ (not an inflection)

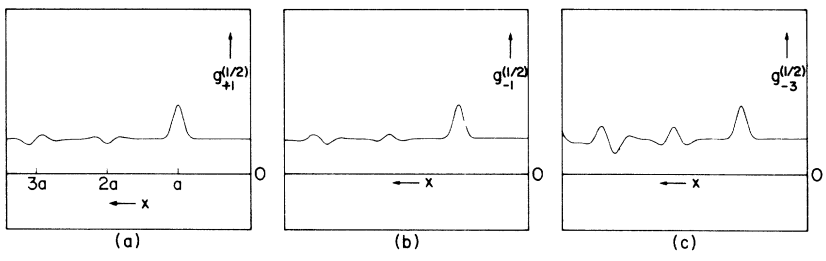

FIG. 10. Plots illustrating spurious features generated by position inconsistency of the $2b*p$ and $p*p$ features for a u_n sequence having a Gaussian-like peak centered at $n=46$ sitting on a constant background. In the fold of this u_n, the $2b*p$ and $p*p$ features are shifted by $\Delta n=+1, -3$ points to produce g functions whose convolution square roots are shown at (a), (b), and (c), respectively.

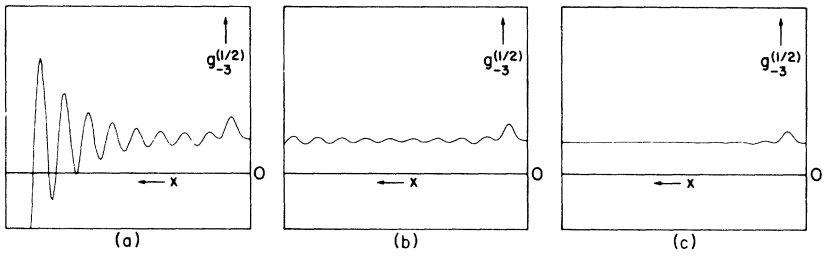

FIG. 11. Plots of unfolds $g_{-3}^{(1/2)}$ obtained from a fold in which the origin has been shifted by three points from that of a true convolution square in a similar manner to that described in connection with Fig. 9. Parts (a), (b), and (c) differ in the p/b magnitude ratio and illustrate how spurious features disappear as this ratio is reduced.

near the origin, it can be shown that the effect of origin shift is limited strictly to the region near the origin, which for small shifts is limited to two to three times n_2. Thus the production of spurious features in $u(x)$ at large x values by an incorrect origin is a direct result of the peaked structure of $u(x)$.

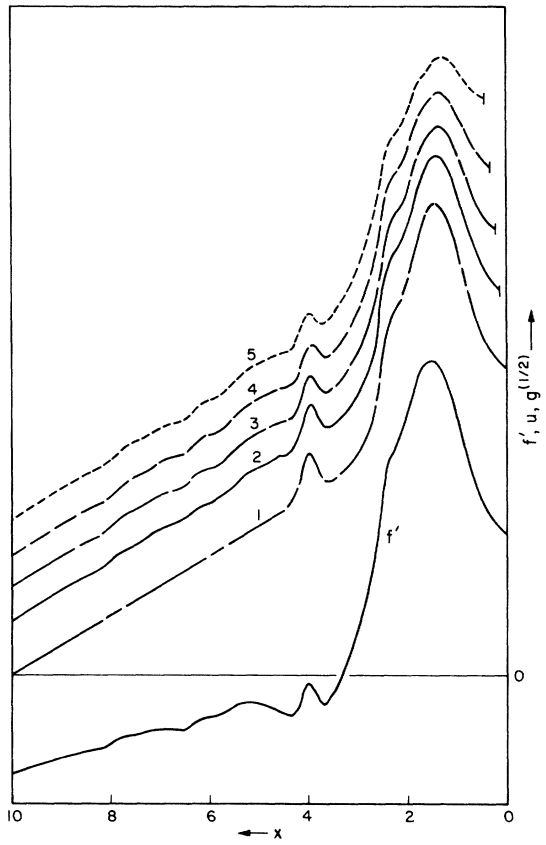

FIG. 12. Plots showing the effect of position inconsistency for the example of Fig. 7 which resembles a possible physical case. The original $u(x)$, curve 1, is compounded of a straight line and the three Gaussian-like peaks shown in Fig. 7. After folding u to produce the fold f a series of g functions with increasing position inconsistency is derived from the original f by the method of origin shift described in the text. The convolution square roots of these g functions, $g^{(1/2)}$, are shown shifted vertically for clarity as curves 2–5 corresponding to origin shifts of $\Delta x = 0.1$, 0.2, 0.3, and 0.4, respectively. Shown below is the derivative of f, f'.

Illustrations of the effects of origin shift on functions like those encountered in INS are given in Figs. 12–14. In Fig. 12, the origin of a constructed function like that of Fig. 7 is shifted to positive x with remarkably little effect. In Fig. 13, we also see that shift to positive x for a function with a large peak near the origin has little effect, but that shift to negative x produces a large effect. In Fig. 14, we see that (when the peak lies some distance from the origin) positive and negative origin shifts produce first-order spurious features which are the negative of each other as expected. We conclude that an experimental $u(x)$ having a peak at $x = a$ and *no* features at $x = 2a$ demonstrates *both* the possession of the correct origin and the fact that $g = v * w$ is very close to a true convolution square $f = u * u$.

X. SUMMARY AND CONCLUSIONS

It is now possible to summarize in a series of statements the nature of the interrelation of physics and mathematics involved in the two-electron Auger neutralization process.

(i) The general case of a two-electron Auger ejection process involving different transition probabilities for the up and down electrons requires that $X(E)$, the ejected electron's kinetic-energy distribution, have the form of the convolution product $g = v * w$ which cannot be inverted knowing g alone.

(ii) The physical limitation of v and w to so-called peaked functions having peaks at the same energy position (x) but of possibly differing magnitude, placed on relatively smooth backgrounds which cut off rather sharply at the Fermi level, makes possible the inversion of $v * w$ as a convolution square root. Peaks in $(v * w)^{(1/2)}$ will lie where those in v and w are and will be averages of them.

(iii) An analysis of the fold functions g of peaked v and w functions has led to an understanding of possible errors in $(v * w)^{(1/2)}$ and to the devising of tests for the genuineness of its features.

(iv) Unknown but relatively smooth variations in electron escape probability or density of final states are shown to have negligible effect on the ability of the deconvolution procedure to "pick out" the peaked structure present in the unfold func-

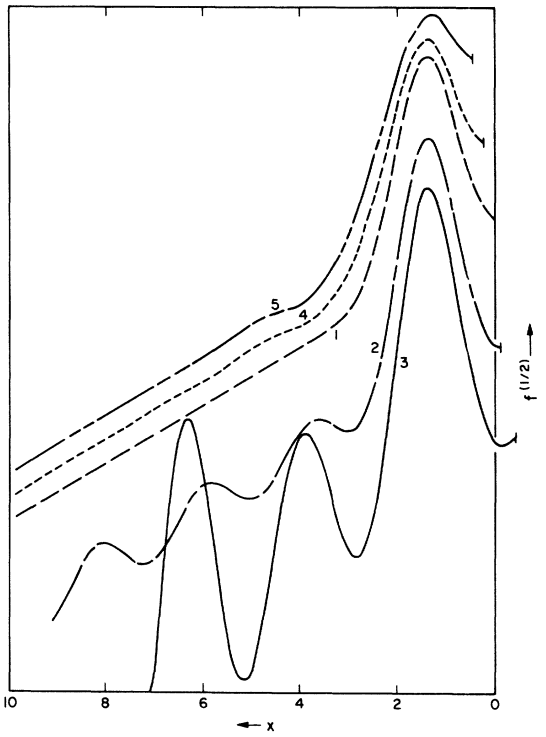

FIG. 13. Illustration of origin shift for a function in which the principal peak lies close to the origin. The curves have been shifted vertically relative to each other for clarity.

tion. Energy degradation behaves similarly. Sharp peaklike structure in the escape probability or final-state density will produce sharper structure in $u(x)$ which can be identified as discussed in Ref. 4 by comparing $u(x)$ curves for two ions of different neutralization energy.

(v) Energy broadening inherent in the electron ejection process obscures the origin of the $g(x)$ function to be deconvolved but the correct origin can be found using an origin-shifting procedure which thus eliminates any possible error due to origin position. This also locates the Fermi level on the electron energy scale which, together with work function and contact potential information, leads to a value for the effective neutralization energy of the ion at the atom-solid separation at which the Auger process occurs.

Similarly we can state a series of conclusions concerning the viability and accuracy of the INS method.

(a) We have satisfied ourselves that the so-called step-midpoint sequential inversion procedure is by far the most satisfactory among either sequential or global techniques for the general class of peaked functions we encounter in this work.

(b) Comparison of the nonlocal $u(x)$ with the local $f'(x)$ and an investigation of the form and position of deeper-lying structure with respect to structure lying closer to the origin provide powerful and adequate tests of the genuineness of $u(x)$.

(c) Possession of an origin-shifting procedure is essential to the proper inversion of a convolution square by whatever method used. Origin error is the dominant problem encountered in this work but it is completely handleable, with the correct origin comparatively easy to determine.

(d) All data thus far obtained and published by the method of INS have produced $u(x)$ functions which either agree in peak structure with $f'(x)$ or differ from $f'(x)$ in a completely understood manner.

(e) Although it must be admitted that the requirement to invert the basic data in INS is a complication not enountered in the implementation of a one-electron emission spectroscopy, we claim that it is not serious or cumbersome once its characteristics are understood and the means of its execution are at hand. In many instances unfolding could be replaced by differentiation where $u(x) \sim f'(x)$, but this is not universally valid. It should not be too difficult to devise an on-line data-taking and computing procedure which would perform all the data taking, data processing, and testing involved in INS.

ACKNOWLEDGMENTS

The authors wish to acknowledge with thanks the

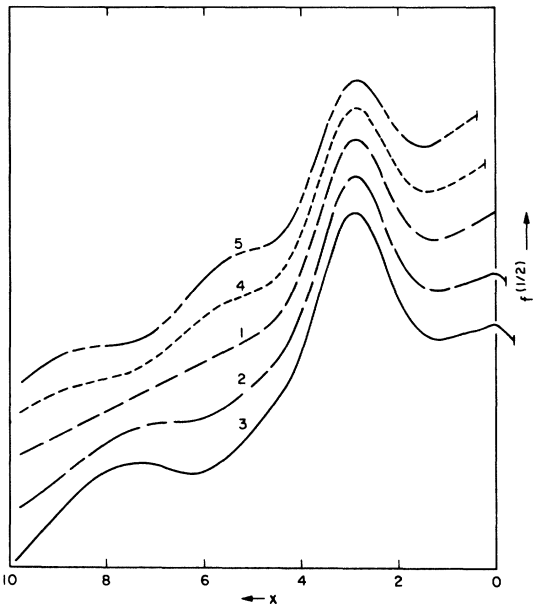

FIG. 14. A sequence of unfold functions in which a single peak lies at larger x than in Fig. 13 derived from g functions differing in origin position. For clarity the curves in this figure are shifted vertically relative to each other by arbitrary amounts.

help and clarification obtained in numerous discussions with their colleagues at Bell Laboratories, particularly L. R. Walker, R. B. Blackman, and M. C. Gray. Others who have kindly contributed to this work and to whom we are also grateful are mentioned in the text. Computer programming has been done for us principally by I. Bogdanski, R. C. Fulton, and B. C. Chambers, and technical assistance has been provided by P. Petrovich. To all of these persons it is a pleasure to express our thanks.

APPENDIX A: OTHER SEQUENTIAL-DIGITAL-UNFOLDING FORMULATIONS

At one point in our studies of sequential digital inversion, before we had understood the full significance of origin error, we were led to believe that our situation would be improved by a more accurate sequential-digital-unfolding formulation. This turned out not to be the case but in the course of these investigations we derived a total of four formulations and studied, or, more properly, experimented with, their stability properties. It appears that a brief statement of this work and its results might be of interest to the reader and could correct possible misconceptions.

Sequential-digital-inversion formulations can be derived in a straightforward and systematic way by inverting so-called closed digital-quadrature rules of increasing accuracy and complexity. When this is attempted, a series of interesting facts emerge, which we state briefly.

(i) Only one formulation, that of Table I, inverts without the independent calculation of the first u point u_0. This is becuase f_1 is expressed as a function of u_0 only. Further, it is true only for this formulation that

$$\lim_{h\to 0} u_0 = \lim_{h\to 0} \frac{(f_1 - f_0)}{h} = [f'(0)]^{1/2} = u(0) . \quad \text{(A1)}$$

(ii) The next group of three more complex formulations require the independent calculation of u_0 because f_1 depends on both u_0 and u_1. This makes these formulations unwieldy. Even with an accurate determination of u_0, obtained by passing a high-order curve through several f_n to get an accurate $f'(0)$ for use in Eq. (A1), we have found these formulations to be much more unstable than that of Table I.

(iii) Higher-order quadrature rules of the closed Newton-Coates series require the independent calculation of more than one initial u value and are completely useless.

(iv) We have examined how each of the first four formulations derived handles a one-point noise deviation in f_n. Only the formulation of Table I has satisfactory features. It also has excellent characteristics with respect to round-off error.

APPENDIX B: COMPARISON OF SEQUENTIAL AND GLOBAL METHODS OF INVERSION

We have done some experimenting with so-called global methods for inverting the self-convolution integral. It is illuminating to discuss them briefly, even though none has proved to be competitive with the step-midpoint unfold for the class of functions to which we have limited ourselves. Any global method we have tried to use has proved to be cumbersome, to require considerable amounts of computer time, and, in some cases, to have difficulty with round-off errors. All give relatively crude answers in that they specify the answer by means of a number of parameters which is relatively much smaller than the number of given data points. In general, this accounts for the stability of global methods. We have found, however, that an amount of smoothing of the given data which is sufficient to stabilize the sequential inversion does not degrade the answer as much as does the limitation to a practicable number of independent parameters in the case of the global methods we have tried. We should distinguish clearly between a true point-by-point instability in sequential inversion and the periodic variations which can be introduced into $u(x)$ if $f(x)$ is not a true convolution square.

We discuss four types of global inversion: methods based on the Laplace or Fourier transform, a method based on a network analogy, and two iterative methods.

It is well known that $u(x)$ can be represented as the inverse Laplace or Fourier transform of the square root of the transform of $f(x)$. It is not surprising that digital schemes based on this equation are possible and suggestions have been made by Blackman,[8] Gentleman,[9] and Amelio and Scheibner.[10,11]

Gentleman[8] has tried a method based on the so-called fast Fourier transform. Two difficulties were encountered. One relates to the way in which the given f_n data are extended to complete one cycle of the cyclic form in which f_n must be used. Thus, if the given f_n data are put on a circle from 0 to π, the requirement that $f(x) = 0$ for $x < 0$ would put zeros in the range 0 to $-\pi$. This amounts to an incompatible extension of f_n in the range π–2π. Gentleman found that his method would work if he would guess an answer, u_n, and fold it to give an initial set of f_n values for the range π–2π which were then used with the given f_n in the range 0–π. A second problem encountered in Gentleman's attempt relates to the ambiguity in phase angle when taking the square root. If the data are sufficiently close together, the choice of phase angle at a given point to be that closest to the phase angle of the preceding point was satisfactory. The data interpolation and the several steps required to produce the correct unfold mean that the method is not competitive

in computer time or over-all simplicity with the sequential method. The method of folding using the fast Fourier transform, however, did provide the excellent check of round-off error in the sequential inversion. Gentleman found that the match error [Eq. (B2) below] between his fold of the sequential unfold and the given f_n data was approximately 10^{-7}.

Amelio and Scheibner[10] and Amelio[11] have also devised digital techniques based on the convolution theorem. In one method the given data are expanded in a polynomial for which orders as high as 15 have been used. The second method replaces f_n by straight line segments, which is equivalent to replacing u_n by the step function on which the step-midpoint fold is based. As many as 30 segments have been used. In each method the basic integral equations are transformed to a linear algebraic system and are then solved. The method will be tedious when pushed to the high order required if the experimental resolution in the data is not to be degraded by the inversion. In his latest work Amelio[11] has used the global technique as a test for a sequentially inverted result. On the basis of this test the given data are varied until they meet a specific requirement. Clearly this data manipulation is required because of noisy initial data due to a weak signal. Otherwise use of the sequential method alone would have been satisfactory.

It is also possible to make use of the electrical-network interpretation of Eq. (2) as the basis of a global inversion scheme. If $f(x)$ is the unit-impluse response of two identical four-terminal networks in tandem, then $u(x)$ is the unit-impulse response function of the network itself. Another way to state it is that we wish to find a network whose response is $f(x)$ when its excitation function is equal to its unit-impulse response function $u(x)$. Simone arranged to integrate the differential equations of a simple network repetitively on an analogue computer and to display $u(x)$ and $f(x)$ on alternate cycles. The circuit parameters could then be varied until the observed $f(x)$ approximated the given data. To achieve accuracy, one would be required to employ a rather complicated network with many adjustable parameters, and it is then not readily apparent how to vary these in order to achieve the desired modification of $f(x)$. Thus the method is instructive, but is not practical for day-to-day use with experimental data.

Iterative methods which avoid direct inversion are also possible. In this general class of method a first guess at the unfold function $u(x)$ is characterized by a limited series of parameters, say, a_i, $i = 0, k$. These could be the coefficients of a polynomial expansion

$$u(x) = \sum_{i=0,k} a_i P_i(x) \tag{B1}$$

in terms of the Chebyshev polynomials, for example. $u(x)$ is then folded with itself to obtain a so-called test set f_n^t. This can be done either by folding the polynomials once and for all and digitalizing the resulting series expansion for $f(x)$, or by digitalizing $u(x)$ and folding these values digitally, say, by the step-midpoint fold. The first of these methods involved the tedious folding of polynomials to high order and also turned out to have very large round-off errors. The second method involving digital folding provided a satisfactory subroutine for generating f_n^t from the given a_i, $i = 0, k$. The test values f_n^t thus obtained were compared with the given data f_n and the match error

$$\Phi = \sum_{i=0,k} (f_n - f_n^t)^2 \Big/ \sum_{i=0,k} f_n^2 \tag{B2}$$

was calculated. A sucessive approximation program was then used to vary the k parameters a_i, $i = 0, k$, so as to minimize the match error. A method using a least-squares steering program of Semmelman was made to work. In many ways it proved to be cumbersome. If k is large enough to give an accurate representation of $u(x)$, the successive-approximation program becomes very time consuming and unwieldy.

We draw the following conclusion from a comparison of global and sequential techniques. It is certainly possible that for some types of functions a global method would work where the sequential method would not. However, for the general class of functions to which we limit ourselves, we contend that the sequential method is far superior. We observe that the sequential method will unfold $f(x) = kx$ with absolute accuracy whereas a global method will have difficulties in producing the finite step in $u(x)$ at $x = 0$. For the same accuracy and resolving power, the sequential method is very much faster and less cumbersome than global methods, and it is possible to push the resolving power further using the sequential method. Resolving power is limited in global methods by the limitation on the maximum number of parameters which can be varied in iterative methods, or solved for in methods which reduce to simultaneous linear equations, before the method becomes too tedious or, in fact, inoperable. Resolving power is limited in the sequential method only by the amount of data smoothing required to keep the method stable enough to produce a relatively smooth curve. Resolving power using the sequential method can thus be improved readily by arranging to obtain "better" experimental data using averaging of many digital runs in a multichannel scaler. In order to take advantage of such improvement of data quality when using a global method, it would be necessary to increase the number of parameters employed to the point where the method could become impractical.

[1]H. D. Hagstrum, Phys. Rev. 150, 495 (1966).
[2]H. D. Hagstrum and G. E. Becker, Phys. Rev. 159, 572 (1967).
[3]H. D. Hagstrum, J. Res. Natl. Bur. Std. 74A, 433 (1970).
[4]H. D. Hagstrum and G. E. Becker, J. Chem. Phys. 54, 1015 (1971).
[5]J. J. Lander, Phys. Rev. 91, 1382 (1953).
[6]H. D. Hagstrum, Y. Takeishi, and D. D. Pretzer, Phys. Rev. 139, A526 (1965).
[7]V. F. Turchin, V. P. Kozlov, and M. S. Malkevich, Usp. Fiz. Nauk. 102, 345 (1970)[Sov. Phys. Usp. 13, 681 (1971)].
[8]R. B. Blackman (private communication).
[9]W. M. Gentleman (private communication). See W. M. Gentleman and G. Sande, in *Proceedings of the Fall Joint Computer Conference, November*, 1966 (Spartan Books, Washington, D. C., 1966), AFIPS Conf. Proc. 29, 563 (1966).
[10]G. F. Amelio and E. J. Scheibner, Surface Sci. 11, 242 (1968).
[11]G. F. Amelio, Surface Sci. 22, 301 (1970).

PHYSICAL REVIEW B VOLUME 4, NUMBER 12 15 DECEMBER 1971

Azbel'-Kaner Cyclotron Resonance in Mercury†

R. G. Poulsen* and W. R. Datars
Department of Physics, McMaster University, Hamilton, Ontario, Canada
(Received 1 July 1971)

Azbel'-Kaner cyclotron-resonance experiments have been carried out with two very flat mercury crystals at a microwave frequency of 34.28 GHz and a temperature of 1.2 °K. Cyclotron effective masses of ten orbits were measured with an error of less than 2%. Five of the orbits (labeled μ_γ, γ_2, κ, μ_2, and ϵ_1) were observed for the first time by Azbel'-Kaner cyclotron resonance. The cyclotron masses of α orbits in the electron lenses were represented by an interpolation scheme which gives the mass for any field direction. This interpolation scheme showed that the second-zone electron lens is tipped 3° out of a (100) plane of the reciprocal lattice toward the [111] direction and that there is a 9% anisotropy of the mass in the (100) plane. A similar interpolation scheme describing the frequencies of de Haas–van Alphen (dHvA) oscillations, which correspond to the β arms of the first-zone hole surface, is also presented. The oscillations were caused by two effects which could not be separated: quantum oscillations of the microwave surface impedance and dHvA torque. Methods for accurately determining the crystal orientation within the experimental apparatus, using the symmetry of the electron-lens masses and of signal peaks arising from open-orbit induced-torque effects, are presented. Cyclotron resonance with the magnetic field inclined to the sample surface is discussed. Several effects indicate anomalous penetration of the electromagnetic field into the metal.

I. INTRODUCTION

The cyclotron effective mass in a metal is proportional to the derivative with respect to energy of the area of a cyclotron orbit and is greater than that determined from the band structure because of mass enhancements arising from electron-electron and electron-phonon interactions.[1] However, contributions of electron-electron interactions to the cyclotron mass are often partially folded into band-structure calculations,[2] and are small and independent of energy and temperature.[3] Thus, in addition to providing information concerning Fermi-surface topology, determinations of the cyclotron mass also yield information about electron-phonon mass enhancement. This is particularly important for mercury in which there is a large electron-phonon interaction. The cyclotron mass, the electron-phonon mass enhancement factor $1+\lambda$, and their variations with temperature and energy are conveniently measured by Azbel'-Kaner cyclotron resonance.

Azbel'-Kaner cyclotron resonance (AKCR) is observed by shining microwave radiation on the surface of a flat metal sample in the presence of an applied magnetic field aligned parallel to the sample surface.[4] Electrons spiraling along the magnetic field direction return an average of $\omega_c\tau/2\pi$ times to the skin depth region at the sample surface. $\omega_c = eH/m_c c$ is the cyclotron frequency, H is the magnetic field, and m_c and τ are the cyclotron mass and relaxation time, respectively. Resonance occurs when electrons and the rf microwave field are in phase for successive cyclotron revolutions. This occurs at fields $H_N = c\omega m_c/Ne$ (N is an integer) for which the microwave frequency ω is equal to, or an integral multiple of, ω_c. The reciprocal fields $1/H_N$ are periodic.

Azbel' and Kaner have shown that, with appropriate orientation of the x and y axes in the sample surface, the surface impedance $Z(H)$ is

$$Z_{xx,yy}(H) = Z_{xx,yy}(0)\left[1 - \exp\left(-2\pi i\omega/\omega_c - 2\pi/\omega_c\tau\right)\right]^{1/3} \quad (1)$$

for electrons of common mass, where $Z(0)_{xx,yy}$ is