Band gaps in a generalized density-functional theory

L. Fritsche and Y. M. Gu*

Institut für Theoretische Physik der Technischen Universität Clausthal, W-3392 Clausthal-Zellerfeld, Germany

(Received 20 November 1992)

As has recently been shown by one of the authors (L.F.), the density-functional scheme of Hohenberg, Kohn, and Sham can consistently be extended to excited states [Physica B **172**, 7 (1991)]. Within this generalized density-functional theory it turns out that the energy for electronic excitations across the gap of insulators and semiconductors can be expressed as the sum of the so-called Kohn-Sham gap and a correction that is usually of the same order of magnitude. (This correction proves to agree up to first perturbational order with that obtained by Godby *et al.* [Phys. Rev. B **37**, 10 159 (1988)] within the so-called *GW* approximation.) The present article reports refined calculations on the band gaps of solid rare gases, alkali halides, diamond, and silicon. The results, are to some extent, still affected by the atomic-sphere approximations which we have been employing, but show relatively fair agreement with the experimental data. We also discuss Janak's theorem, the insulator-metal transition under hydrostatic pressure, and the problem of the Fermi surface in metals.

I. INTRODUCTION

Notwithstanding the growing interest in applications of conventional density-functional (DF) theory based on the paper by Kohn and Sham,[1] there is little awareness of certain weaknesses in the foundation of this scheme. The conceptual shortcomings have repeatedly been addressed by one of the present authors (L.F.) and associates.[2–5] In eliminating these deficiencies, it turns out that an alternative foundation automatically yields an extension of the DF framework to excited states, and the resulting scheme may be viewed as a generalized density-functional (GDF) theory. The present paper rests on this extension and draws, in particular, on the interconnection between the true excitation energy, ΔE, referring to a total energy difference of the N-electron system under study, and the difference in the band energies of the one-particle states that are, respectively, depleted and filled in the process of an interband transition. This band-energy difference has to be added onto a correction which is commonly (but quite inappropriately) referred to as a "many-body correction" and may be viewed as a justification of the so-called "scissors operator."

It is evident from the derivation of Kohn-Sham- (KS) type one-particle equations that the effective potential in those equations cannot depend on the one-particle state that the associated equation refers to. The self-interaction-corrected (SIC) scheme suggested by Perdew and Zunger[6] is at variance with this fundamental property in that it explicitly requires an orbital-by-orbital cancellation of the electronic self-energy. In a consistent N-electron theory of indistinguishable particles, each electron appears with the same probability in each of the orbitals, and hence a cancellation of self-energy can only occur within the electron-electron interaction integral containing the pair density $\rho(\mathbf{r}',\mathbf{r})$ which does not depend on individual orbitals. Exchange-correlation potentials based on suitably chosen pair-correlation factors that obey the so-called sum rule, guarantee the nonoccurrence of self-interaction, and are nevertheless orbital independent. (See, e.g., Fritsche and Gollisch.[7])

In Sec. II we briefly discuss the origin of the gap correction. We also address Janak's theorem[8] and point out the limitations of its validity. Section III is concerned with the predictability of the sign of the gap correction and its connection to the definition of the Fermi surface in DF theory. Results of gap corrections, which will be presented in Sec. IV, contain (except for the alkali halides) two sets of gap values and their respective corrections based on self-interaction-free exchange-correlation potentials. These potentials are connected to certain pair-correlation factors. We shall refer to the associated gap values as resulting from a "nonlocal potential," although this is a rather misleading (but widely used) name for a potential that is actually local, such as the Hartree potential. To demonstrate that the finiteness of the gap correction is not tied to nonlocal potentials, we have recalculated the various band structures, in particular, the gaps and gap corrections, by using the local exchange-correlation potential derived by von Barth and Hedin.[9] Where corresponding data were available, we have also compared our results to those obtained within the SIC scheme and the so-called "*GW* approximation" to the self-energy in a quasiparticle description of electrons in solids. The idea of the latter approach was already put forward by Hedin[10] in 1965, but only twenty years later applied to the band problem of insulators and semiconductors by Hybertsen and Louie.[11] The method has later been taken up and technically been improved by Godby, Schlüter, and Sham[12] and Godby and Schlüter.[13] In Sec. V we discuss the possible origin of smaller discrepancies that still exist with our results and also address the prospects of practicable improvements.

II. THE GDF CONCEPT

Conventional DF theory rests on the concept of adiabatically switching off the electron-electron interaction and simultaneously turning on an additional external potential $\tilde{V}_{\rm ext}(\lambda,\mathbf{r})$ so that the original one-particle density $\rho(\mathbf{r})$ is conserved. It can be shown that this potential exists for any N-electron eigenstate $\Psi_n(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)$ and has the form[4]

$$\tilde{V}^{(n)}_{\rm ext}(\lambda,\mathbf{r})=(1-\lambda)V^{(n)}_{H}(\mathbf{r})+V^{(n)}_{\rm XC}(\mathbf{r})-\lambda V^{(n)}_{\rm XC}(\lambda,\mathbf{r}) , \tag{1}$$

where $V^{(n)}_H(\mathbf{r})$ is the Hartree potential associated with the density $\rho_n(\mathbf{r})$ that derives from Ψ_n, and $V^{(n)}_{\rm XC}(\mathbf{r})$ denotes the exchange-correlation potential associated with the variation of the exchange-correlation energy $E^{(n)}_{\rm XC}$ by

$$\delta E^{(n)}_{\rm XC}=\int V^{(n)}_{\rm XC}(\mathbf{r})\delta\rho_n(\mathbf{r})d^3r . \tag{2}$$

The extra potential $\tilde{V}^{(n)}_{\rm ext}(\lambda,\mathbf{r})$ exists even for spin-ordered systems in a spin-dependent form $\tilde{V}^{(n)}_{\rm ext}(\lambda,\mathbf{r},s)$ that guarantees the conservation of spin-classified densities $\rho_{ns}(\mathbf{r})$, where $s=\pm1$ denotes spin-up and spin-down, respectively. The existence proof of $\tilde{V}^{(n)}_{\rm ext}(\lambda,\mathbf{r})$ invalidates an earlier study by Harris[14] that seemed to indicate that an adiabatic connection of Ψ_n to a noninteracting state cannot rigorously be ensured.

If $\lambda=0$, one is dealing with a noninteracting N-electron system. The associated Schrödinger equation can be separated, then, and one is led to N one-particle equations (in Hartree units)

$$[-\tfrac{1}{2}\nabla^2+V^{(n)}_{\rm eff}(\mathbf{r},s)]\psi_{is}(\mathbf{r})=\epsilon_{is}\psi_{is}(\mathbf{r}) , \tag{3}$$

where

$$V^{(n)}_{\rm eff}(\mathbf{r},s)=V_{\rm ext}(\mathbf{r})+V^{(n)}_{H}(\mathbf{r})+V^{(n)}_{\rm XC}(\mathbf{r},s) . \tag{4}$$

The latter relation follows from Eq. (1) for $\lambda=0$. The associated N-electron wave function has the form of a $N\times N$ Slater determinant $\Phi_n(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)$ containing N orbitals $\psi_{is}(\mathbf{r})$ that solve Eq. (3). The total set of solutions to Eq. (3) constitutes a complete orthonormal set of function in terms of which we can construct an infinite orthonormal set of determinants Φ_k by systematically selecting different subsets of N orbitals. The true wave function may then be expanded in a configuration-interaction series

$$\Psi_n(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)=\sum_k c_{nk}\Phi_k(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N) ,$$

which may be rewritten

$$\Psi_n=\Phi_n+\tilde{\Psi}_n , \tag{5}$$

where

$$\tilde{\Psi}_n=\sum_k c'_{nk}\Phi_k \tag{6}$$

and

$$c'_{nk}=\begin{cases} c_{nk}-1 & \text{for } k=n \\ c_{nk} & \text{otherwise} .\end{cases}$$

Since Ψ_n and Φ_n yield the same densities $\rho_{ns}(\mathbf{r})$, it follows from Eq. (5) that $\Phi^*_n\tilde{\Psi}_n+\Phi_n\tilde{\Psi}^*_n+\tilde{\Psi}^*_n\tilde{\Psi}_n$ integrates exactly to a zero-density contribution if one integrates this expression with respect to $N-1$ electron coordinates. It is, furthermore, evident from Eq. (5) that all coefficients c'_k in the sum of the right-hand side of Eq. (6) tend to zero as the coupling strength λ is reduced to zero. Hence, in performing the adiabatic switching one maps the true wave function Ψ_n uniquely onto a determinant Φ_n. [In case where $V_{\rm ext}(\mathbf{r})$ has central or axial symmetry Φ_n may consist of a linear combination of a few determinants that differ in their highest-lying degenerate orbitals. (See Cordes and Fritsche.[3])] As has been shown by Harriman,[15] the requirement that an $N\times N$ determinant yield the same one-particle density as Ψ_n does not uniquely define the orbitals from which that determinant is formed. Hence, in order to make the mapping $\Psi_n\rightarrow\Phi_n$ unique, it is absolutely crucial that the extra-potential $\tilde{V}^{(n)}_{\rm ext}(\mathbf{r},s)$ defining $V^{(n)}_{\rm eff}(\mathbf{r},s)$ in Eq. (3) can be uniquely constructed. This one-to-one correspondence between Ψ_n and Φ_n allows one to classify the ground state and the excited states as in Hartree-Fock (HF) theory. It should be noted, however, that the orbitals $\psi_{is}(\mathbf{r})$ solve Eq. (3) where $V^{(n)}_{\rm XC}(\mathbf{r},s)$ is a local potential and, hence, they are definitively different from HF orbitals. This applies as well to the total energy which—by construction—is given by the eigenvalue E_n associated with Ψ_n and can be shown to have the form[4]

$$E_n=\sum_s\sum_i^{(N_s)}\epsilon^{(n)}_{is}-\tfrac{1}{2}\int\rho_n(\mathbf{r})V^{(n)}_{H}(\mathbf{r})d^3r+\sum_s\int\rho_{ns}[\bar{\epsilon}^{(n)}_{\rm XC}(\mathbf{r},s)-V^{(n)}_{\rm XC}(\mathbf{r},s)]d^3r . \tag{7}$$

The function

$$\epsilon^{(n)}_{\rm XC}(\lambda,\mathbf{r},s)=-\frac{1}{2}\sum_{s'}\int\frac{\rho_{ns'}(\mathbf{r}')f^{(n)}_{s's}(\lambda,\mathbf{r}',\mathbf{r})}{|\mathbf{r}'-\mathbf{r}|}d^3r' , \tag{8}$$

constitutes the exchange-correlation energy per particle at coupling strength λ, and $f^{(n)}_{s's}(\lambda,\mathbf{r}',\mathbf{r})$ denotes the so-called correlation factors (originally introduced by McWeeny[16] with a different sign) which are defined as one minus the respective pair-correlation functions. The averaged exchange-correlation energy per particle that appears in Eq. (7) and is characterized by an overbar is defined in analogy to $\epsilon^{(n)}_{\rm XC}(\lambda,\mathbf{r},s)$ except that

$$\bar{f}^{(n)}_{s's}(\mathbf{r}',\mathbf{r})=\int_0^1 f^{(n)}_{s's}(\lambda,\mathbf{r}',\mathbf{r})d\lambda$$

stands in place of the λ-dependent correlation factors.

An interband transition bridging the gap of an insulator or semiconductor may, within the GDF framework, be described as a transition from the ground state Ψ_i (mapping onto Φ_i) to the lowest-lying excited state Ψ_f that maps onto Φ_f. The respective total energies are given by Eq. (7) and their difference may be cast into the form[4]

$$\Delta E=\epsilon_f-\epsilon_i+\Delta_{fi} , \tag{9}$$

where

$$\Delta_{fi}=\Delta E_{\rm XC}-\int V^{(0)}_{\rm XC}(\mathbf{r})\Delta\rho(\mathbf{r})d^3r \tag{10}$$

is associated with the change of the charge density $\Delta\rho(\mathbf{r})$ that is connected with the transition $\Psi_i\rightarrow\Psi_f$. This quantity is finite for the following reason: For an infinitesimal change $\delta\rho(\mathbf{r})$ of the ground-state density $\rho_0(\mathbf{r})$, which occurs when applying a small perturbative potential $\delta V(\mathbf{r})$, we have, according to Eq. (2),

$$\delta E_{\mathrm{XC}} - \int V_{\mathrm{XC}}^{(0)}(\mathbf{r})\delta\rho(\mathbf{r})d^3r = 0 \; . \tag{11}$$

By contrast, $\rho_f(\mathbf{r})$ and $\rho_i(\mathbf{r})$ derive from two eigenfunctions Ψ_f and Ψ_i, neither of which can be considered an infinitesimal distortion of the other, and both belong to the same external potential $V_{\mathrm{ext}}(\mathbf{r})$. Consequently, Eq. (11) is invalidated if $\delta\rho(\mathbf{r})$ is replaced by $\Delta\rho(\mathbf{r})$. It can be shown that Eq. (10) may be rewritten[4]

$$\Delta_{fi} = \int [2\bar{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r}) - V_{\mathrm{XC}}^{(0)}(\mathbf{r})][|\psi_f(\mathbf{r})|^2 - |\psi_i(\mathbf{r})|^2]d^3r \; , \tag{12}$$

which constitutes the basis of our calculations discussed in Sec. III. Clearly, both $\bar{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r})$ and $V_{\mathrm{XC}}^{(0)}(\mathbf{r})$ contain many-body information via the pair correlation to which they are connected. On the other hand, ϵ_f and ϵ_i, being eigenvalues of the one-particle equation (3), contain information of this kind as well, since $V_{\mathrm{XC}}^{(0)}(\mathbf{r})$ explicitly contributes to the effective potential in this equation. It is, therefore, hardly illuminating to refer to Δ_{fi} as being a many-body correction to $\epsilon_f - \epsilon_i$.

In concluding this section we want to address a popular misconception concerning Janak's theorem.[8] It appears to be a commonly accepted view that this theorem proves the vanishing of Δ_{fi} in Eq. (9). As will become evident from the following consideration, this fallacy is connected to the fact that Eq. (11) does not hold for discontinuous changes $\Delta\rho(\mathbf{r})$. In deriving Janak's theorem one first rewrites Eq. (7) in the form

$$E_n = \sum_s \sum_i n_{is}\epsilon_{is}^{(n)} - \tfrac{1}{2}\int \rho_n(\mathbf{r})V_H^{(n)}(\mathbf{r})d^3r + \sum_s \int \rho_{ns}[\bar{\epsilon}_{\mathrm{XC}}^{(n)}(\mathbf{r},s) - V_{\mathrm{XC}}^{(n)}(\mathbf{r},s)]d^3r \; , \tag{13}$$

where

$$\rho_{ns}(\mathbf{r}) = \sum_i n_{is}|\psi_{is}(\mathbf{r}, n_{is1}, n_{is2}, \ldots)|^2$$

and

$$n_{is} = \begin{cases} 1 & \text{for } N_s \text{ occupied orbitals} \\ 0 & \text{otherwise} \; , \end{cases}$$

with the latter quantities denoting orbital occupation numbers. If one analytically continues E_n as a function of these occupation numbers to noninteger values, one can determine the change δE_n that results from varying n_{is}. Most of the terms that primarily occur on varying Eq. (13) cancel each other if one observes that the orbitals $\psi_{is}(\mathbf{r})$ obey Eq. (3) and that Eq. (11) holds in this particular case. The result may be written

$$\delta E_n = \sum_s \sum_i \epsilon_{is}^{(n)}\delta n_{is} \tag{14}$$

and represents Janak's theorem. Unfortunately, Janak gave Eq. (14) the form

$$\frac{\partial E_n}{\partial n_{is}} = \epsilon_{is}^{(n)} \; ,$$

which is actually incorrect because the partial derivative with respect to n_{is} implies that the other occupation numbers are kept constant. On the other hand, any admissible change in the occupation numbers must conform to the requirement of particle conservation

$$N = \sum_s \sum_i n_{is} \; , \tag{15}$$

since exchange is not defined for a system of a noninteger number of particles. Clearly, Eq. (15) is irreconcilable with varying only a single occupation number. If one allows two occupation numbers, say n_i and n_f, to vary, Eq. (14) attains the form

$$\delta E_{fi} = (\epsilon_f - \epsilon_i)\delta n_f \; , \tag{16}$$

where we have used Eq. (15) and dropped irrelevant indices. If one would, furthermore, naively integrate this equation with respect to n_f over its entire range from zero to one and justifiably assume that the one-particle energies do not change for an extended solid, one would arrive at

$$\Delta E_{fi} = \epsilon_f - \epsilon_i \; . \tag{17}$$

The error made in going from Eq. (16) to Eq. (17) consists in the fact that the validity of the former rests on Eq. (11), whereas the latter equation involves the finite change $\Delta\rho(\mathbf{r})$ which gives rise to the finiteness of Δ_{fi} according to Eqs. (10) and (11).

III. CONSEQUENCES OF THE GAP CORRECTION

To simplify the ensuing considerations we adopt the viewpoint of the atomic-sphere approximation (ASA), that is, we subdivide the lattice into atomic spheres centered at the atomic nuclei and allow for additional suitably chosen empty spheres in cases where the lattice under study is not close packed. Within each atomic sphere, the first bracketed expression under the integral in Eq. (12), which we shall denote by $-\tilde{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r})$, is essentially a monotonic function of the distance from the nucleus and proves to be generally negative. Direct band gaps in semiconductors and insulators are characteristic of p-type Bloch states at the top of the valence band and s-type states at the bottom of the conduction band. If we introduce the spherical average of $|\psi_{f/i}(\mathbf{r})|^2$ (which we shall indicate by an overbar), the sphere-integrated charge density $4\pi r^2\overline{|\psi_i(\mathbf{r})|^2}$ proves to be localized closer to the nucleus than the respective expression for the final state $\psi_f(\mathbf{r})$. As follows from inspection of Eq. (12), the sphere-integrated charge density of $\psi_i(\mathbf{r})$ appears with a negative sign, and since $-\tilde{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r})$ is negative and monotonic the gap correction Δ_{fi} must be positive in these cases, in agreement with the experiment. Similar arguments hold for the fundamental optical gap in transition-metal oxides where the initial state is d type rather than p type. We have so far studied only NiO and CoO and found positive correction of about 5 eV. A detailed discussion will be published elsewhere. The situation is different with transitions that interconnect the bottom of a d band in Ni metal, for example, with the topmost occupied d bands. The latter are associated with antibonding d-type Bloch states that are slightly stronger localized within the atomic sphere than the bonding d-type states at the bottom. As a result, the corresponding correction Δ_{fi} is negative and amounts to approximately -1 eV. In photoemission experiments the effective d-band width of Ni metal appears, therefore, narrower by ~ 1 eV than the calculation yields for the ground-state band structure. As regards the angular-momentum decomposition of the band states, their change around a band gap as a function of momentum $\mathbf{k}$ is small in a relatively large portion of the Brillouin zone. It turns out, as a result of this behavior, that $\Delta_{fi}(\mathbf{k})$ for optical transitions beyond the gap energy is almost constant within this energy regime. This may be interpreted in the spirit of the so-called scissors operator: in analyzing the experimentally observed interband transitions, the relevant portions of the conduction band seem to be rigidly shifted by the amount of the gap correction.

Another not exactly obvious consequence of our result on the total-energy difference given by Eq. (9) concerns the Fermi surface in DF theory. It is commonly assumed without justification that the ground-state image of Ψ_i, i.e., the determinant Φ_i, contains the N lowest-lying one-particle states $\psi_{\hat{n}s}(\mathbf{k},\mathbf{r})$ where $\hat{n}$ is the band index and $\mathbf{k}$ the momentum vector. (The fact that this assumption may be questionable has first been addressed by Harris.[14]) If the unoccupied and occupied states are not separated from each other by a gap, one has a Fermi surface in $\mathbf{k}$ space defined by

$$\epsilon_{\hat{n}s}(\mathbf{k})=\epsilon_F \ ,$$

where $\hat{n}$ denotes the uppermost occupied band(s), and ϵ_F is the Fermi energy. If δE_n, given by Eq. (14), were the most general expression for the variation of the total energy E_n, the above standard assumption on the filling of states in the ground state ($n=0$) would, in fact, lead to

$$\delta E_0>0 \ ,$$

as required. This result is immediate on using Eq. (15) in the form

$$\epsilon_F \sum_{is} \delta n_{is}^{(0)}=0$$

and subtracting this equation from Eq. (14) so that

$$\delta E_0=\sum_{i,s}(\epsilon_{is}^{(0)}-\epsilon_F)\delta n_{is}^{(0)} \ .$$

Because of Eq. (14), both $\delta n_{is}^{(0)}$ and its parenthesized factor in front are negative for states that are occupied in the ground state and positive for the remaining states. Hence, all terms under the sum are positive. Clearly, Eq. (14) is a consequence of rewriting the total energy, originally given by Eq. (7), in the fictitious generalized form defined by Eq. (13). However, in going through a consistent first-principles derivation of δE_0 one arrives, instead of Eq. (14), at

$$\delta E_0=\sum_s \sum_i \epsilon_{is}^{(0)}\delta n_{is}^{(0)}+\delta\tilde{T}_0 \ ,$$

where

$$\tilde{T}_0=\langle T_{e\text{-}e}\rangle_0-\langle T_0\rangle_0 \ ,$$

with $\langle T_{e\text{-}e}\rangle_0$ and $\langle T_0\rangle_0$ denoting the kinetic energy of the true interacting system and the noninteracting image system, respectively. (For details, see Fritsche.[4]) Although $\tilde{T}_0$ is a positive quantity, $\delta\tilde{T}_0$ can have either sign and, hence, one cannot exclude the possibility that δE_0 becomes negative for certain distortions of the wave function Ψ_i. Whether or not Φ_i really corresponds to the ground state can be checked by again using Eq. (9). To this end one considers a transition from the state Ψ_i ($\rightarrow\Phi_i$) to another state Ψ_f ($\rightarrow\Phi_f$), where Φ_f differs from Φ_i only in that one orbital (with energy ϵ_i) at the Fermi surface has been replaced by another one whose energy ϵ_f is above the Fermi surface by an infinitesimally small amount. Equation (9) yields, for the change of the total energy in that case,

$$\Delta E_{fi}=\Delta_{fi} \ . \tag{18}$$

If the Fermi surface is very aspherical, the orbitals at distinctly differently curved portions of the Fermi surface can differ sizably in their angular-momentum decomposition and, hence, Δ_{fi} can easily be of the order of 0.1 eV or larger. In those cases $\epsilon_{\hat{n}}(\mathbf{k})=\epsilon_F$ does not define an equi-energy surface as required for a true Fermi surface. In general, the ground state is defined by a surface in $\mathbf{k}$ space, which separates occupied from unoccupied states such that a transition $\Psi_i\rightarrow\Psi_f$ leads to changes

$$\Delta E_{fi}=\epsilon_f-\epsilon_i+\Delta_{fi}$$

that vanish for any possible pair (i,f) at that surface. It appears to be likely that some failures of standard DF theory in consistently describing ground-state properties of solids are associated with an incorrect choice of occupied orbitals close to ϵ_F (defined by the standard filling). The transition-metal oxides, for instance, CoO and NiO, are possibly an example of this type of failure. The calculations yield incompletely filled, relatively flat d bands which are Ni and Co derived, respectively. The incomplete filling gives rise to a Fermi surface that crosses the uppermost band(s). The strongly curved oxygen bands are lower by 3 eV and, hence, fully occupied according to the conventional understanding. However, on transferring an electron from the top of the oxygen bands to the Fermi level, one gains a small amount of energy if one employs Eq. (18) to calculate the associated energy change. In order to establish the true ground state, one therefore has to deplete the topmost oxygen state until ΔE_{fi} equals zero for all possible transfers between Ni (or Co) and oxygen states, both of which then form a Fermi surface in $\mathbf{k}$ space. If there were a very effective process that immediately annihilates excessive O holes and d electrons formed on applying an external electric field, the occupation in $\mathbf{k}$ space would stay centrosymmetric, which would be tantamount to zero conductivity. Although this picture may be very speculative, it at least demonstrates the existence of new territory within DF theory that is bound to go unnoticed as long as one follows the standard "filling rules" that are not proven.

The Fermi-surface problem is closely related to the insulator-metal transition of semiconductors, of which we only consider here Si as an example. If one subjects Si to hydrostatic pressure, the conduction-band minimum drops and eventually lines up with the maximum of the valence band. On further increasing the pressure, the minimum keeps shifting downward and starts introducing empty conduction-band levels below the occupied valence levels. Again using DF standard rules, one would argue that such unoccupied levels cannot occur in the Si ground state associated with that pressure and, hence, the topmost valence levels ought to be depleted in favor of the lower-lying conduction-band states until the respective Fermi levels agree. An electronic structure of this kind would be associated with metallic conductivity, which, however, is not observed at that pressure. In fact, even at considerably higher pressure Si remains insulating. This becomes understandable if one assumes that the conduction band remains unoccupied even when its minimum has already dropped substantially below the valence-band maximum because the downward transfer of an electron from the topmost valence level would require a positive energy ΔE_{fi} as follows again from using Eq. (18). Within a very large range of pressure Δ_{fi} is positive and amounts to 0.5–1.0 eV. Hence, the difference $\epsilon_f - \epsilon_i$, which decreases monotonically as one raises the pressure, must attain a negative value of approximately this magnitude before ΔE_{fi} turns negative and conduction-band states start getting filled. In other words, the insulator-metal transition occurs when ΔE_{fi} changes sign, rather than $\epsilon_f - \epsilon_i$.

It is evident from the above considerations that Eq. (9) lends itself to discussing many solid-state properties that are not satisfactorily accessible within conventional DF theory or lie definitely outside its framework. A further virtue of Eq. (9) consists in the simple form of the gap correction as given by Eq. (12) which is considerably less involved than its equivalent definition within the *GW* approximation. The latter hardly allows one to predict the sign of Δ_{fi} in a simple way and requires substantially more computational effort than our expression.

IV. RESULTS

A. Solid rare gases

The simple fcc structure of the solid rare gases makes them particularly suitable for a first test of our GDF approach. Moreover, the optical absorption—in particular, the fundamental gap of these archetypal insulators—is experimentally well studied. Preliminary results on solid Ne, Ar, and Kr have already been published elsewhere.[4] The calculations have been refined using an elaborate computer code to calculate the nonlocal exchange-correlation potentials, and we also included solid Xe. We use Andersen's linear-muffin-tin-orbital (LMTO) method[17] in the ASA version. The results are listed in Table I. The data marked "nonlocal" refer to a band-structure calculation based on a nonlocal potential where the correlation factor was chosen to have the form of either a Gaussian or a Lorentzian to the power $\frac{5}{2}$. The differences with respect to experimental data neither indicate a clear trend as a function of the atomic number nor favor one of the two correlation factors. On the other hand, one clearly recognizes that the SIC scheme yields clearly less satisfactory results. Even the gap corrections obtained using a local approximation to

TABLE I. Interband transition energies for solid rare gases. The shorthand notations "Gauss" and "Lorentz" refers to the respective model forms of the correlation factor $f_{s's}(\mathbf{r}',\mathbf{r})$ used in constructing $V_{\mathrm{XC}}(\mathbf{r},s)$. The results listed as "local *B-H*" refer to calculations where $V_{\mathrm{XC}}(\mathbf{r},s)$ was assumed to have the form derived by von Barth and Hedin (Ref. 9). Recent results of Bacalis, Papaconstantopoulos, and Pickett (Ref. 18) that have been calculated within the self-interaction correction scheme are also listed for comparison. The experimental reference data are the same as those quoted in that paper. All energies are given in units of eV.

	Exchange-correlation potential V_{XC}		Kohn-Sham band gap ϵ_g	Band-gap correction Δ_{fi}	Excitation energy	
					Theory $\epsilon_g + \Delta_{fi}$	Experiment
Ne	Nonlocal	Gauss	11.99	10.97	22.96	
		Lorentz	11.52	11.08	22.60	20.82
	Local	*B-H*	11.63	10.27	21.90	
		SIC	11.40	5.16	16.56	
Ar	Nonlocal	Gauss	8.88	4.79	13.67	
		Lorentz	8.93	4.85	13.78	13.88
	Local	*B-H*	8.45	5.90	14.35	
		SIC	3.86	5.90	11.96	
Kr	Nonlocal	Gauss	7.88	3.39	11.27	
		Lorentz	7.84	3.27	11.11	11.43
	Local	*B-H*	7.13	5.45	12.56	
		SIC	6.76	3.35	10.11	
Xe	Nonlocal	Gauss	6.85	2.53	9.38	
		Lorentz	6.51	2.23	8.74	9.12
	Local	*B-H*	6.29	4.70	10.99	
		SIC	5.56	2.67	8.23	

$[2\bar{\epsilon}_{XC}^{(0)}(\mathbf{r})-V_{XC}^{(0)}(\mathbf{r})]$ in Eq. (12) and combining these corrections with the associated gaps from a "local" band structure give a better overall agreement than the SIC data. One also recognizes that Δ_{fi} becomes systematically smaller as the atomic number Z increases. This correlates with the shrinking gap size when Z is raised. As discussed in Sec. III, the magnitude of Δ_{fi} is largely determined by the effect that a lowest unoccupied atomic-type state is more or less inflated compared to highest atomic-type state. The lower the excitation (gap) energy ϵ_g, the smaller this inflation effect becomes and, hence, Δ_{fi} drops with decreasing ϵ_g.

B. Diamond and silicon

Diamond has for many years been the example used most to demonstrate the incapability of local-density-approximation (LDA) -DF theory to account properly for many-body effects determining the true magnitude of the fundamental gap. The latter turns out to be 1.7 eV larger than that obtained from a LDA calculation. It is interesting to note that—as with the solid rare gases—the gap in the band structure is hardly affected by replacing the LDA exchange-correlation potential with one of our nonlocal versions. The pertinent data are listed in Table II. (The calculated results are based on a lattice parameter of 3.57 Å.) As can be seen from the quantities Δ_{fi}, they achieve the necessary correction quite satisfactorily, although there is a striking difference in the angular-momentum character of the final state which is filled in crossing the gap: that conduction-band state is dominantly p type rather than s type as with the solid rare gases. Since the initial state at the top of the valence band is p type as well, one would not expect a large correction Δ_{fi} to occur because it depends directly on the difference of the square moduli of these two states. The conduction state, however, inflates to some extent into the interstitial region, thereby lowering the contribution of $|\psi_f(\mathbf{r})|^2$ within the atomic spheres, which gives rise to a sizable difference $|\psi_f(\mathbf{r})|^2-|\psi_i(\mathbf{r})|^2$. Similar considerations apply to the data listed in Table III which refer to silicon in the diamond structure. (The lattice parameter has been chosen to be 5.43 Å in agreement with the pertinent literature.) Except for the L point, where one still has a discrepancy of about 25%, our results again compare reasonably well with experiment. In Figs. 1 and 2 we have plotted the dependence of the effective gap ΔE_{fi} of Si as a function of its relative volume. As explained in Sec. III, the material should become metallic when ΔE_{fi} drops below zero. The lower two curves refer to the band-energy difference $\epsilon_f-\epsilon_i$ for the direct gap (Fig. 1) and the indirect gap (Fig. 2) where the results marked "LDA" were obtained by using the LDA potential and those marked "KST" are based on the true exchange-correlation potential defined by Godby, Schlüter, and Sham.[12] The notation for the upper two curves is self-explanatory. One clearly recognizes that a naive filling scheme to which the LDA and KST curves refer would predict a metal-insulator transition around $V/V_0=0.78$, which is not observed. This is in keeping with the result that the effective gap at this relative volume is still far from being zero, in both the GW and the GDF approximations. On the other hand, it is obvious from comparing Figs. 1 and 2 that the transition will eventually take place by filling the lowered conduction states above the indirect gap, whereas the states above the direct gap are shifted upwards as the volume shrinks.

TABLE II. Interband transition energies at main symmetry points of the diamond band structure and for the indirect gap ("Min."). The GW results with and without parentheses were obtained by Louie (Ref. 19) and by Godby and Schlüter (Ref. 13), respectively. The latter authors calculate the Kohn-Sham band gap by using the "true exchange-correlation potential." The experimental result for the Γ point is taken from a recent paper of Armon and Sellschop (Ref. 20). The other experimental data are identical to those quoted by Godby and Schlüter (Ref. 13). Energies are given in units of eV.

Gap	Exchange-correlation potential V_{XC}		Kohn-Sham band gap ϵ_g	Band-gap correction Δ_{fi}	Excitation energy Theory $\epsilon_g+\Delta_{fi}$	Experiment
Γ	Nonlocal	Gauss	5.56	0.94	6.50	
		Lorentz	5.40	0.97	6.37	6.5
	Local	*B-H*	5.54	0.56	6.10	
		GW	5.72	1.54	7.26(7.5)	
X	Nonlocal	Gauss	10.82	1.94	12.76	
		Lorentz	10.65	1.84	12.49	12.5
	Local	*B-H*	10.56	1.00	11.56	
		GW	11.07	1.48	12.55(12.5)	
L	Nonlocal	Gauss	10.98	0.84	11.82	
		Lorentz	10.81	0.89	11.70	
	Local	*B-H*	11.31	0.49	11.80	
		GW	11.27	1.34	12.61	
Min.	Nonlocal	Gauss	3.97	1.47	5.44	
		Lorentz	3.77	1.40	5.17	5.48
	Local	*B-H*	3.80	0.75	4.55	
		GW	4.21	1.12	5.33(5.6)	

TABLE III. Interband transition energies at main symmetry points of the silicon band structure and for the indirect gap ("Min."). The notation is the same as in Tables I and II. The experimental data are taken from the paper by Godby and Schlüter (Ref. 13). The *GW* results listed with and without parentheses are due to Louie (Ref. 19) and to Godby and Schlüter (Ref. 13), respectively. Energies are in eV.

	Gap	Exchange-correlation potential V_{XC}		Kohn-Sham band gap ϵ_g	Band-gap correction Δ_{fi}	Excitation energy Theory $\epsilon_g+\Delta_{fi}$	Experiment
Si	Γ	Nonlocal	Gauss	3.05	0.63	3.68	
			Lorentz	2.52	0.72	3.24	3.40
		Local	*B-H*	2.82	0.69	3.51	
			GW	2.68	0.62	3.30(3.35)	
	X	Nonlocal	Gauss	3.68	1.11	4.79	
			Lorentz	3.15	1.19	4.34	4.25
		Local	*B-H*	3.51	0.97	4.48	
			GW	3.64	0.63	4.27	
	L	Nonlocal	Gauss	2.24	0.27	2.51	
			Lorentz	2.46	0.17	2.63	3.3±0.2
		Local	*B-H*	2.55	0.26	2.81	
			GW	2.83	0.66	3.49(3.54)	
C	Min.	Nonlocal	Gauss	0.32	1.09	1.41	
			Lorentz	0.28	1.11	1.39	1.17
		Local	*B-H*	0.54	0.91	1.45	
			GW	0.66	0.58	1.24(1.29)	

C. Alkali halide crystals

Ionic crystals are well known for their large "gap discrepancies" and therefore provide another important testing ground for our gap-correction formula. To gain some confidence in the capability of our approach we have performed self-consistent LDA calculations on nine alkali halides. The results are shown in Table IV. To reduce the computational effort in determining the band-gap corrections we have replaced $2\bar{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r})-V_{\mathrm{XC}}^{(0)}(\mathbf{r})$ in Eq. (12) with the pertinent local approximation, which is consistent with calculating the band structure at the LDA level. The interband transition links a p-type anion state to an s-type cation state. As in the case of the solid rare gases, it is hence qualitatively clear that the gap corrections must be positive. But there is also a relatively fair quantitative agreement with the experiments considering the rather crude approximations that have been made. Apart from using a local (von Barth–Hedin) approximation to $\bar{\epsilon}_{\mathrm{XC}}^{(0)}(\mathbf{r})$ and $V_{\mathrm{XC}}^{(0)}(\mathbf{r})$ we have not made any attempt to refine the subdivision of the NaCl structure into atomic spheres by using different sizes to match the different ionic radii. Instead we have used identical spheres for cations and anions in performing the LMTO-ASA calculations. Nevertheless, the results compare quite favorably with those obtained by Kunz,[22] who starts from a Hartree-Fock (HF) level and calculates "correlation corrections" using Toyozawa's electronic polaron formalism. His derivation is predicated on the assumption that the HF total-energy difference for an excitation across the gap yields only the HF band gap, i.e., the difference of the associated one-particle energies. It remains unclear whether or there is definitely no analogous term in the HF total-energy difference that would correspond to our expression Δ_{fi}. The occurrence of Δ_{fi} for such a transition is associated with the invalidation of Janak's theorem which is just the analogue of Koopman's theorem. A

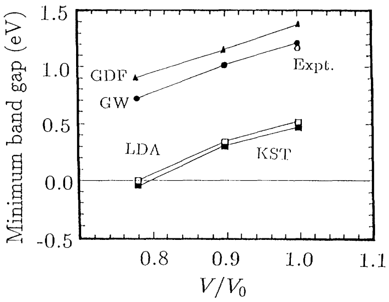

FIG. 1. Minimum band gap of silicon in the diamond structure as a function of the unit-cell volume. The present results (marked by triangles) have been connected by a dashed line and are marked "GDFT." The other results are taken from the paper of Godby and Needs (Ref. 21). The curves labeled "LDA" and "KST," respectively, refer to the Kohn-Sham gaps which have been calculated within LDA and, alternatively, by using a true exchange-correlation potential. The *GW* results correspond to quasiparticle excitation energies. The experimental gap value for zero pressure is indicated by an open circle.

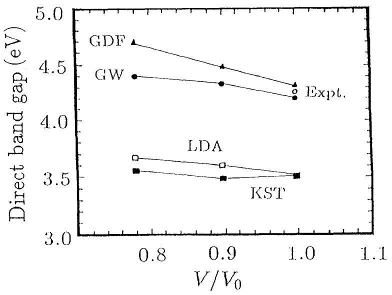

FIG. 2. Results analogous to those in Fig. 1 for the direct band gap of silicon in the diamond structure at the X point.

TABLE IV. Interband transition energies at the Γ point of the alkali halides. The calculations have, in that case, only been performed on the local *B-H* level explained in Table I. The experimental data are identical to those quoted in the paper by Kunz (Ref. 22). Energies are in eV.

	Exchange-correlation potential V_{XC}		Kohn-Sham band gap ϵ_g	Band-gap correction Δ_{fi}	Excitation energy Theory $\epsilon_g+\Delta_{fi}$	Experiment
LiF	Local	*B-H*	9.9	6.1	16.0	14.2
LiCl	Local	*B-H*	6.5	3.4	9.9	9.4
LiBr	Local	*B-H*	5.2	3.5	8.7	7.6
NaF	Local	*B-H*	7.4	6.4	13.8	11.5
NaCl	Local	*B-H*	5.4	3.6	9.0	9.0
NaBr	Local	*B-H*	4.5	3.7	8.2	7.1
KF	Local	*B-H*	5.7	7.2	12.9	10.8
KCl	Local	*B-H*	4.8	3.8	8.6	8.7
KBr	Local	*B-H*	5.5	4.4	9.9	7.4

clarification of this point appears to be very desirable. As regards comparative *GW* studies on alkali halides, we are only aware of one gap value, viz., 9.1 eV for LiCl obtained by Louie.[19] This value agrees also very satisfactorily with our result.

V. CONCLUSIONS

The results obtained in the present paper lend considerable credence to the validity of GDF theory in general, and to our gap-correction formula in particular. In view of the simplicity of the latter, the overall agreement with the experimental data is very gratifying. The discrepancies that still exist for the solid rare gases and the semiconductors are very likely due to insufficient accuracy of our LMTO-ASA one-particle states whose square moduli enter directly into the gap correction. Moreover, the function $\tilde{\epsilon}^{(0)}_{XC}(\mathbf{r})$, which occurs as a factor of the difference of those square moduli, is based on our approximate form of the correlation factors. It appears to be unlikely that the error introduced by this can be reliably estimated. However, these correlation factors are determined such that they definitely exclude electronic self-interaction as they are rigorously subject to the so-called sum rules and may therefore be expected to give rise to only minor inaccuracies of the gap correction. As already stated, the band-gap corrections for the alkali halides have deliberately been calculated at a lower level of approximation and are clearly open to improvement. All in all, the preliminary results of the present study may be taken as an encouraging basis for further work in this direction.

From a conceptual point of view, it appears to be very satisfying that our gap correction does not contain any "dynamical effects," as opposed to the *GW* expression, which involves the full frequency-dependent inverse dielectric function. This is a consequence of the underlying quasiparticle picture, which describes the motion of an extra electron in a dynamically responding background of $N-1$ or N electrons, respectively. The gap energy is defined as the difference between the lowest removal energy of that extra electron (leaving $N-1$ electrons behind) and the maximum energy gained by adding an extra electron to the N-electron system in its ground state. Clearly, none of the two situations corresponds exactly to the initial and final states in the actual experiment. In our treatment that refers to N fully indistinguishable particles, dynamical (i.e., frequency-dependent) effects cannot occur, since each of the N-electron states considered (either the ground state or some excited state) is constructed as a solution to the time-independent Schrödinger equation. Hence, the associated total-energy difference for excitations across a gap refers to the actually established situation in a threshold photoconductivity experiment where one is dealing with a system of N electrons before and after the photoabsorption.

ACKNOWLEDGMENT

We are greatly indebted to Professor N. Christensen, University of Aarhus, Denmark, for having furnished us with his LMTO computer code. We are also grateful to Dr. J. Cordes and H. Grieb for their assistance in using some of their subroutines concerning nonlocal exchange-correlation potentials. The present study has been financially supported by the German Ministry of Research and Technology (BMFT).

*Permanent address: Department of Physics, University of Science and Technology of China, Hefei, Anhui, The People's Republic of China.

[1] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1333 (1965).

[2] L. Fritsche, Phys. Rev. B **33**, 3976 (1986).

[3] J. Cordes and L. Fritsche, Z. Phys. D **13**, 345 (1989).

[4] L. Fritsche, Physica B **172**, 7 (1991).

[5] L. Fritsche, C. Kroner, and Th. Reinert, J. Phys. B **25**, 4287 (1992).

[6] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[7] L. Fritsche and H. Gollisch, Z. Phys. B **48**, 209 (1982); in *Local Density Approximations in Quantum Chemistry and Solid State Physics*, edited by J. P. Dahl and J. Avery (Plenum, New York, 1984), p. 245.

[8] J. F. Janak, Phys. Rev. B **18**, 7165 (1978).

[9] U. von Barth and L. Hedin, J. Phys. C **5**, 1629 (1972).

[10] L. Hedin, Phys. Rev. **139**, A796 (1965).

[11] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **32**, 7005 (1985); **34**, 5390 (1986); Phys. Rev. Lett. **58**, 1551 (1987).

[12] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. B **37**, 10 159 (1988).

[13] R. W. Godby and M. Schlüter, in *Proceeding of the 18th International Conference on the Physics of Semiconductors, Stockholm, 1986,* edited by O. Engström (World Scientific, Singapore, 1987), p. 1103.

[14] J. Harris, Phys. Rev. A **29**, 1648 (1984).

[15] J. E. Harriman, Phys. Rev. A **24**, 680 (1981).

[16] R. McWeeny, Rev. Mod. Phys. **32**, 335 (1960).

[17] O. K. Andersen, Phys. Rev. B **12**, 3060 (1975).

[18] N. C. Bacalis, D. A. Papaconstantopoulos, and W. E. Pickett, Phys. Rev. B **38**, 6218 (1988).

[19] S. G. Louie, in *Proceeding of the 18th International Conference on the Physics of Semiconductors, Stockholm, 1986* (Ref. 13), p. 1095.

[20] H. Armon and J. P. F. Sellschop, Phys. Rev. B **26**, 3289 (1982).

[21] R. W. Godby and R. J. Needs, Phys. Rev. Lett. **62**, 1169 (1989).

[22] A. B. Kunz, Phys. Rev. B **26**, 2056 (1982).