Fractional charge perspective on the band-gap in density-functional theory

Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang
Department of Chemistry, Duke University, Durham, North Carolina 27708
(Dated: February 5, 2008)

The calculation of the band-gap by density-functional theory (DFT) methods is examined by considering the behavior of the energy as a function of number of electrons. It is found that the incorrect band-gap prediction with most approximate functionals originates mainly from errors in describing systems with fractional charges. Formulas for the energy derivatives with respect to number of electrons are derived which clarify the role of optimized effective potentials in prediction of the band-gap. Calculations with a recent functional that has much improved behavior for fractional charges give a good prediction of the energy gap and also $\varepsilon_{\text{homo}} \approx -I$ for finite systems. Our results indicate it is possible, within DFT, to have a functional whose eigenvalues or derivatives accurately predict the band-gap.

PACS numbers: 71.10.-w, 31.15.Ew, 71.15.Mb

One of the many important uses of density-functional theory (DFT) is the calculation of the bandstructure which has many applications throughout physics, for example in semiconductors, electron transport and reactions at surfaces. The first step in achieving accuracy in the band structure is to understand the band-gap which standard functionals have long been known to systematically underestimate by as much as ~50%. Recent efforts have focused on use of the optimized effective potential (OEP) method, which can often give an improvement in the prediction of band-gaps for small-gap semiconductors, but has problems with wider gap semiconductors and insulators [1, 2, 3]. In many cases it has proved necessary to move to the quasi-particle GW theory to calculate the band-gap of solids accurately [4]. Conventionally, the band-gap problem, has been related to the so-called “derivative discontinuity” in the exchange correlation potential: even with an accurate Kohn-Sham potential, the energy-gap is still different from the true gap by an amount of the derivative discontinuity [5, 6]. This perspective, however, does not offer the understanding or the mechanism needed for making progress for band-gap prediction with DFT.

In this Letter, a new perspective is offered: the band-gap problem is shown to be related to the behavior of approximate density functionals for fractional numbers of electrons, an issue which has drawn considerable recent interest [7, 8, 9]. This enables us to understand the problem with band-gap calculations and offers ideas to develop functionals which predict the band-gap correctly. Examples will be given for molecules where the energy gap can be compared with explicit calculations of systems with fractional charges.

The fundamental band-gap for an $N$-electron system in an external potential $v(r)$, is given by

$$E_g = \left[ E_v(N-1) - E_v(N) \right] - \left[ E_v(N) - E_v(N+1) \right] = I - A,$$

(1)

where $E_v(N)$ is the ground-state energy of the $N$ particle system and $I$ is the ionization energy and $A$ is the electron affinity. For a system with a fractional number of electrons $N + \delta N$, with $0 < \delta N < 1$, it has been shown that the energy is a straight line connecting the total energies at integer numbers of electrons; namely, $E_v(N + \delta N) = \delta N E_v(N) + (1 - \delta N) E_v(N + 1) \ [10, 11]$. This linear relation means that the energy gap in Eq. (1) can be given by the derivative difference

$$E^\text{der}_g = \lim_{\delta N \to 0} \left\{ \frac{\partial E_v}{\partial N} \right|_{N+\delta N} - \frac{\partial E_v}{\partial N} \right|_{N-\delta N} \right\} \ (2)$$

If we substitute in the DFT total energy expression $E_v = T_s[\rho] + V_{\text{ext}}[\rho] + J[\rho] + E_{xc}[\rho]$, we have

$$E_g = \varepsilon_{\text{lumo}} - \varepsilon_{\text{homo}} + \Delta_{xc} = E^\text{KS}_g + \Delta_{xc} \ (3)$$

where the $E^\text{KS}_g$ is the gap in a Kohn-Sham calculation and the $\Delta_{xc}$ is the derivative discontinuity [5, 6].

In this work we identify the problem with calculations using approximate density functionals by considering the basic assumption in Eq. (2), that the energy at $N + 1$ and at $N - 1$ can be given simply from the derivatives at $N$. This is true for exact DFT but it may or may not be true for approximate functionals. The key is to investigate the behavior of the total energy as a function of numbers of electrons. To do this we consider a non-interacting ground state reference system where we allow the occupation numbers of the orbitals to vary the number of electrons smoothly; the first-order reduced density matrix of the reference system is given by

$$\rho_s(r, r') = \sum_i n_i \phi_i(r) \phi_i^*(r') \ (4)$$

where $n_i = 1$ for $i < f$, $n_i = \delta N$ for $i = f$, and $n_i = 0$, for $i > f$, and $f$ is the index for the frontier orbital. The electron density is just its diagonal $\rho_s(r) = \rho_s(r, r)$. The behavior of three qualitatively different exchange-correlation functionals: the local density approximation (LDA), Hartree-Fock (HF) and MCY3 [12] is shown in
Fig. 1. MCY3 was constructed as a functional of $\rho_s(r,r')$ to give a much improved description of fractional numbers of electrons and we can see that it gives a straight line interpolation between the integers. The interpolation from the other functionals is far from linear, HF curves in a concave manner and LDA in a very convex manner. There are two main problems with approximate density functionals: First, they can have a remarkably different behavior from the exact functional in fractional charge systems, compared to their behavior for the corresponding integer charge systems. Second, the error in the integer charge systems can also be significant. The combined effects lead to the error in the band-gap prediction from derivative information.

For molecules, LDA has a very reasonable description of the integer values ($I$ and $A$ are given well) but a much worse description in between the integers. The use of the first derivatives for LDA will clearly not give the $I$ and $A$ from the integer calculations. Because of the convexity of the curve, LDA will give too small a value for $I$ and too large a value for $A$, meaning that the band-gap $I - A$ will be too small as shown by the dotted lines in Fig 1. The case for HF is very different as the integer values are not as good because of the lack of correlation, and also it curves in a concave manner. These two errors cancel each other to some extent in the prediction of $I$ but add together in the calculation of $A$. For HF the value of $I$ will be about right and $A$ too small, meaning that the band-gap $I - A$ will be too large as shown in Fig 1. For MCY3, as it is very straight, the use of the derivatives will give a prediction very similar to the integer calculations for $I$ and $A$, and the band-gap $I - A$ should be accurate as the integer values are good. For functionals that have a linear behavior for fractional charge systems (e.g. MCY3 and the exact functional) the initial derivative is all that is needed to calculate $I$ and $A$ and the band-gap as in Eq. (2). Hence, we consider analytic expressions for $\partial E_v/\partial N$ for some families of exchange-correlation functionals. The main ideas and results are summarized here, with further details presented in the supplementary material [13].

In the fractional charge non-interacting system, Eq. (4), the orbitals are the eigenstates of an one-electron local potential $v_s(r)$

$$\left(-\frac{1}{2}\nabla^2 + v_s\right)\phi_i = \varepsilon_i\phi_i, \quad (5)$$

or a non-local potential $v_s^{NL}(r,r')$.

$$\left(-\frac{1}{2}\nabla^2 + v_s^{NL}\right)\phi_i = \varepsilon_i^{GKS}\phi_i. \quad (6)$$

The former is the original Kohn-Sham (KS) reference system and the latter has been called the Hartree-Fock-Kohn-Sham (HFKS) [14] or the generalized Kohn-Sham (GKS) method [15].

For the Kohn-Sham reference system with local potential $v_s(r)$, we here use the potential-functional formulation [16]. The electron density $\rho_s(r)$ can be represented as the set of orbitals and occupation numbers $\{\phi_i, n_i\}$, or equivalently as the local KS potential and total particle number $\{v_s(r), N\}$. Thus the total energy functional, formally in term of density as $E_v[\rho_s(r)]$, can be expressed as $E_v[v_s(r), N]$. The ground state energy is the minimum of the KS energy functional, expressed (explicitly or implicitly) in terms of the local potential $v_s(r)$: $E_v(N) = \min_{v_s} E[v_s(r), N] = E[v_s^{eq}, N]$, where the minimizer $v_s^{eq}$ is the optimized effective potential (OEP), as established recently [16]. The variational nature of $v_s^{eq}$ simplifies the calculation of the
derivative: $\frac{\partial E_v(N)}{\partial N} = \frac{\partial E_v[\phi^{\nu_0}, N]}{\partial N} |_{\phi^{\nu_0}}$. Consider a change in the total number of electrons $N = N_0 + \delta N$, where $N_0$ is an integer and $|\delta N| < 1$. At fixed $\nu_0$, all the orbitals $\{\phi^{\nu_0}\}$, as its eigenstates, are fixed. Since $\rho_{s}(r)$ is the ground state density of the reference potential $\nu_0$, only the frontier level occupation $n_f$ is allowed to change $\delta N = \delta n_f$, thus

$$\frac{\partial E_v(N)}{\partial N} = \left( \frac{\partial E_v[\{\phi^{\nu_0}, n_f\}]}{\partial n_f} \right)_{\{\phi^{\nu_0}\}}, \quad (7)$$

where the frontier orbital is either the LUMO, $n_f = n_{lumo}$, if $\delta N > 0$, or the HOMO, $n_f = n_{homo}$, if $\delta N < 0$. We consider three cases for which the analytic derivatives can be obtained [13]:

**Case A:** $E_{xc}[\rho_s(r)]$, an explicit functional of $\rho_s$ (e.g. LDA or GGA):

$$\frac{\partial E_v(N)}{\partial N} = \varepsilon_f \quad (8)$$

where $\varepsilon_f$ is the KS eigenvalue for the frontier orbital in the local potential $v_s(r) = v(r) + v_J(r) + v_{xc}(r)$. This is exactly the combination of Eq. (7) with Janak’s theorem for $n_f$ [17].

**Case B:** $E_{xc}[\rho_s(r, r')]$, a functional of the first order density matrix minimized with a local potential as in Eq. (5) (e.g. OEP exact exchange, EXX).

$$\frac{\partial E_v(N)}{\partial N} = \varepsilon_f + \langle \phi_i | v + v_J + v_{xc}^{NL}(r, r') | \phi_i \rangle, \quad (9)$$

where the non-local potential $v_{xc}^{NL}(r, r') = \frac{\delta E_{xc}[\rho_s(r, r')] }{\delta \rho_s(r, r')}$. Eq. (2) is a key result, showing that for general orbital functionals, $\frac{\partial E_v}{\partial N}$ is not given by the frontier OEP eigenvalue, $\varepsilon_f$, but with a correction term $\Delta_{xc}^I$. This general result agrees with [1] in the case of exact exchange, and is related to the results of [18] from the self energy.

**Case C:** $E_{xc}[\rho_s(r, r')]$, with the energy minimized with respect to the orbitals $\phi_i$ (e.g. HF):

$$\frac{\partial E_v(N)}{\partial N} = \varepsilon^{GKS}_f \quad (10)$$

where $\varepsilon^{GKS}_f$ is the eigenvalue of the frontier orbital of the non-local potential $v + v_J + v_{xc}^{NL}(r, r')$, as in Eq. (6).

All three cases can be unified in the expression

$$\frac{\partial E_v(N)}{\partial N} = \langle \phi_i | H_{eff} | \phi_i \rangle, \quad (11)$$

where $H_{eff} = -\frac{1}{2} \nabla^2 + v + v_J + v_{xc}(r)$ for Case A where $E_{xc} = E_{xc}[\rho(r)]$, and $H_{eff} = -\frac{1}{2} \nabla^2 + v + v_J + v_{xc}^{NL}(r, r')$ for case B and C where $E_{xc} = E_{xc}[\rho_s(r, r')]$, which are often called orbital functionals. In cases A and C, $\frac{\partial E_v(N)}{\partial N}$ is equal to the corresponding eigenvalue, but not in case B. The only difference in between cases B and C is the orbitals used to evaluate the overall expression.

Combining Eqs. (2) and (11) thus leads to the general expression for the band-gap from derivatives for an N-electron system:

$$E_{\bar{g}}^{der} = \langle \phi_{N+1} | H_{eff} | \phi_{N+1} \rangle - \langle \phi_N | H_{eff} | \phi_N \rangle. \quad (12)$$

We consider a few illustrative atomic and molecular systems for which we have performed self-consistent calculations using a cc-pVQZ basis set in an modified version of CADPAC. We compare $-\varepsilon_{homo}$ with the experimental $I$, $-\varepsilon_{lumo}$ with the experimental $A$ and also their corresponding differences. The results for LDA, HF and MCY3 are given in Table 1. MCY3 gives very good agreement between $-\varepsilon_{homo}$ and $I$ which is to be expected from its straight line behavior. We should emphasize that this has not been seen before for calculations with approximate exchange-correlation functionals. The error is relatively small, 0.5eV, and is roughly similar in $I$, $A$ and also the difference $I - A$. LDA eigenvalues have a large error, with a consistent underestimation of $I$ by about 5 eV and overestimation of $A$ by about 4 eV and a poor prediction of the gap. LDA does well for the explicit calculation of the $N + 1$ and $N - 1$ systems, it is just the use of the derivatives at $N$ that lead to large errors, this is now clearly understood from the fractional charge picture. The HF $-\varepsilon_{homo}$ is often close to $I$, however there are larger errors for $A$ and also for the gap.

The band-gap issue is well understood for calculations with local density functionals (case A) or orbitals func-
Table II: Comparison of GKS and OEP eigenvalues, and \(\partial E/\partial n\) for HF and MCY3, see [13].

| Mol  | MCY3 | MCY3 | HF | HF | HF | Expt |
|------|------|------|----|----|----|------|
| OEP  | GKS  | OEP  | GKS | GKS | GKS |       |
| ε   | \(\partial E/\partial n\) | ε   | \(\partial E/\partial n\) | ε   | \(\partial E/\partial n\) | ε   |
| C   | 0.70 | 10.16 | 0.13 | 14.73 | 13.49 | 12.76 | 10.00 |
| I   | 10.58 | 11.11 | 11.12 | 11.97 | 11.94 | 11.14 | 11.27 |
| A   | 9.88 | 0.96 | 1.09 | 10.50 | -1.55 | -0.82 | 1.27 |
| F   | 4.06 | 13.74 | 13.75 | 5.62 | 20.49 | 20.44 | 14.40 |
| G   | 14.67 | 15.16 | 15.17 | 15.94 | 18.11 | 18.13 | 15.70 |
| A   | 10.61 | 1.42 | 1.42 | 10.32 | -2.37 | -2.31 | 1.30 |

The OEP (MCY3) \(-\varepsilon_{\text{homo}} \approx I\) (as is proven for the exact functional [21]) however the \(E_g\) is much smaller than the exact gap. The inclusion of \(\Delta_{xc}\) gives a much better agreement between the derivatives and the GKS eigenvalues. This brings us on to the nature of this term; it is only the difference between a KS and GKS calculation and is needed to correctly give the derivative at \(N\). It does not, however, address the question of whether the functional used for the calculation has the correct straight line behavior for fractional numbers of electrons, which is the key question in the evaluation of the band-gap.

In conclusion we have carried out analysis and calculations on systems with fractional numbers of electrons to gain understanding of the band-gap problem in DFT. We show that the band-gap is only given by the eigenvalue difference if the functional has the correct linear behavior for systems with a fractional charge. We have recently developed a functional with this linear behavior giving \(-\varepsilon_{\text{homo}} \approx I\) and \(-\varepsilon_{\text{lumo}} \approx A\) and a good prediction of the band-gap in molecules. We have also considered OEP calculations, in which the derivative of the energy with respect to number of electrons is not given by the OEP eigenvalue. When the derivative is correctly evaluated, it gives practically the same as in GKS calculations. Our work thus provides the new insight: it is possible to have a functional which gives the correct band-gap from the eigenvalues or derivative information, so long as it has the correct fractional charge behavior and accurate energies for integer systems. Such possible functionals include explicit functionals of the electron density \(E_{xc}[\rho(r)]\). We have only considered the explicit calculation of molecules but the same ideas are undoubtedly of key importance in solids. The understanding gained in this Letter offers a new perspective and way forward for accurate calculations of the band-gap in DFT.

Support from NSF is greatly appreciated.

[1] M. Städele, J. A. Majewski, P. Vogl, and A. Görling, Phys. Rev. Lett. 79, 2089 (1997).
[2] M. Gruning, A. Marini, and A. Rubio, J. Chem. Phys. 124, 154108 (2006).
[3] M. Gruning, A. Marini, and A. Rubio, Phys. Rev. B 74, 161103(R) (2006).
[4] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. B 37, 10159 (1988).
[5] J. P. Perdew and M. Levy, Phys. Rev. Lett. 51, 1884 (1983).
[6] L. J. Sham and M. Schlüter, Phys. Rev. Lett. 51, 1888 (1983).
[7] Y. Zhang and W. Yang, J. Chem. Phys. 109, 2604 (1998).
[8] P. Mori-Sánchez, A. J. Cohen, and W. Yang, J. Chem. Phys. 125, 201102 (2006).
[9] A. Ruzsinszky, J. P. Perdew, G. I. Csonka, I. O. A. Vydrov, and G. E. Scuseria, J. Chem. Phys. 126, 104102 (2006).
[10] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz Jr., Phys. Rev. Lett. 49, 1601 (1982).
[11] W. Yang, Y. Zhang, and P. Ayers, Phys. Rev. Lett. 84, 5172 (2000).
[12] A. J. Cohen, P. Mori-Sánchez, and W. Yang, J. Chem. Phys. 126, 191109 (2007).
[13] See EPAPS Document No. ? for details of the derivation of Equations (10)-(13) and additional calculations on molecules. This document can be reached via a direct link in the online article’s HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
[14] R. G. Parr and W. Yang, "Density-Functional Theory of Atoms and Molecules" (Oxford University Press, New York, 1989).
[15] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, Phys. Rev. B 53, 3764 (1996).
[16] W. Yang, P. W. Ayers, and Q. Wu, Phys. Rev. Lett. 92, 146404 (2004).
[17] J. F. Janak, Phys. Rev. B 18, 7165 (1978).
[18] M. E. Casida, Phys. Rev. B 59, 4694 (1999).
[19] W. Yang and Q. Wu, Phys. Rev. Lett. 89, 143002/1 (2002).
[20] Q. Wu and W. Yang, J. Theo. Comp. Chem. 2, 627 (2003).
[21] J. P. Perdew and M. Levy, Phys. Rev. B 56, 16021 (1997).