Why the Generalized Gradient Approximation Works and How to Go Beyond It

KIERON BURKE,* JOHN P. PERDEW, AND MATTHIAS ERNZERHOF
Department of Physics and Quantum Theory Group, Tulane University, New Orleans, Louisiana 70118

Received September 15, 1995; revised manuscript received December 26, 1995; accepted December 28, 1995

ABSTRACT

The local spin density (LSD) approximation, while of only moderate accuracy, has proven extremely reliable over three decades of use. We argue that any gradient-corrected functional should preserve the correct features of LSD, even if the system under study contains no regions of small density gradient. The Perdew–Wang 1991 (PW91) functional respects this condition, while, e.g., the Lee–Yang–Parr (LYP) correlation functional violates it. We extend this idea to the next generation of density functionals, those which incorporate exact exchange via the optimized effective potential (OEP), with a model in which the correlation hole is constructed from the exact exchange hole. The resulting exchange-correlation hole is deeper and less diffuse than the exact exchange hole. We denote such a functional as “locally correlated Hartree–Fock” and list a variety of conditions such a functional should satisfy. We demonstrate the promise of this approach with a crude simple model. © 1997 John Wiley & Sons, Inc.

Introduction and Summary of Conclusions

For many years, the local spin density (LSD) approximation [1] has been the mainstay of electronic structure calculations in solid-state physics [2–4]. This approximation may be written

\[ E^{\text{LSD}}_{\text{xc}}[n_\uparrow, n_\downarrow] = \int d^3r \, n(r) \epsilon_{\text{xc}}^{\text{unif}}[n_\uparrow(r), n_\downarrow(r)], \]

where \( \epsilon_{\text{xc}}^{\text{unif}}(n_\uparrow, n_\downarrow) \) is the exchange-correlation energy per particle of a uniform electron gas (jellium) [5–7]. The LSD exchange-correlation energies are insufficiently negative (by about 10%) for almost all atoms, molecules, and solids [2]. The LSD is a reliable, moderate-accuracy approximation. (By the
"reliability" of an approximation, we mean that errors are regular and chemical trends are reproduced.) For many solid-state purposes, the LSD level of accuracy is sufficient.

However, LSD is not accurate enough for most chemical applications, which require the determination of energy differences with considerable precision. Hence the disinterest of the quantum chemistry community toward density functional methods [8] until recently. New, gradient-corrected functionals, of the form

\[ E_{xc}^{GA} [n_\uparrow, n_\downarrow] = \int d^3 f\left( n_\uparrow (r), n_\downarrow (r), \nabla n_\uparrow, \nabla n_\downarrow \right) \]

have reduced LSD atomization energy errors by about a factor of 5 [9]. Unfortunately (and in marked contrast to LSD) a plethora of functions \( f \) are in use in the literature [10], each yielding different energies for the same system. These functionals may be divided into two broad classes: "locally based" functionals, whose construction starts from the uniform electron gas [11–16], and semiempirical functionals [17, 18], which contain one or more parameters fitted to a particular finite system (or class of systems). Such functionals are called generalized gradient approximations (GGAs) in the literature (although the term originally [12] meant only the locally based functionals). The culmination of many years of theoretical work has produced the most modern locally based functional, Perdew–Wang 1991 (PW91), which contains no empirical input, while Becke exchange [17] and Lee–Yang–Parr (LYP) correlation [18] (BLYP) are perhaps the most popular semiempirical formulas.

How should one choose between the various GGAs currently being used? The answer to this question lies in the unique choice of LSD which has been used for many years. The particular form of exchange-correlation energy per particle is always chosen as that of the uniform electron gas. (VWN [5], PZ [6], and PW [7] are almost identical parametrizations of this quantity.) If the form Eq. (1) is applied to a set of atomic and molecular energies, and the function \( e_{xc}(n_\uparrow, n_\downarrow) \) fitted thereto, the result will differ from \( e_{xc}^{\text{unif}}(n_\uparrow, n_\downarrow) \).

But, as explained in detail in the following section, LSD is reliable because it respects conditions on the exact universal density functional. Thus the fitted form of \( e_{xc}(n_\uparrow, n_\downarrow) \), while more accurate for the specific systems it was fitted to (and, possibly, for similar systems), may fail when applied to different systems, while \( e_{xc}^{\text{unif}}(n_\uparrow, n_\downarrow) \) will (almost) always yield a moderately accurate energy. Thus the semiempirical approach yields essentially a limited-range interpolation scheme, while LSD produces a controlled and justified extrapolation scheme, being expected to give moderately accurate results, and systematic errors, for systems not yet even imagined.

We can extend this logic to the choice of GGA. If a user is confident that her or his system is covered by the interpolation scheme of the semiempirical formulas, and has no need to understand the origin of the error made by the functional, then such semiempirical functionals should prove useful. The hybrid functionals which mix GGA with Hartree–Fock are typically based on this philosophy [19]. Otherwise, the locally based functionals, which incorporate those features which LSD gets right (see the third section), are best, with PW91 being the latest of these.

Note that a given energy evaluation may have little or no contribution from regions where the gradient is small and therefore be insensitive to how a functional behaves in this limit. (Zupan and co-workers are currently studying these kinds of questions with a new density gradient analysis [20].) On this basis, one might say that the behavior of a functional is not important in this limit. However, a good locally based functional should incorporate all the correct features of LSD, including the correct energy per particle of the uniform gas. Thus, by construction, a locally based functional reduces to LSD in the uniform density limit, and this is an identifying characteristic of such functionals. Thus LYP (see the third section) and the crude version of locally correlated Hartree–Fock (see the fourth section) do not fall in this category.

These arguments can be extended to the next generation of functionals, those based on exact exchange, as incorporated in the optimized effective potential [21] (OEP). Once OEP calculations become part of the standard repertoire of quantum chemical techniques, there will doubtless be an explosion of semiempirical formulas. However, a functional which can combine the good features of both LSD and GGA with those of OEP, such as the locally correlated Hartree–Fock (LCHF) discussed in the fourth section, can be hoped to achieve another factor of 5 reduction in energy errors, and so yield chemical accuracy, for all those systems for which LSD yields moderate accuracy!
A useful aid in the construction of such a functional would be if results were reported using all levels of functionals, especially LSD. Such results give guidance as to how well functionals have been constructed, and what is still missing in them. Achievement of an accurate LCHF functional might be if results were reported using all levels of functionals, especially LSD and LSD has dominated solid-state electronic structure calculations as thoroughly as LSD has dominated solid-state electronic structure calculations [22].

**WHY LSD IS RELIABLE**

The exchange-correlation energy may be written in terms of the exchange-correlation hole [23]:

\[ E_{xc} = \frac{1}{N} \int d\lambda \int d^3r d^3r' \frac{n(r) n_{xc,\lambda}(r, r')}{2|\mathbf{r} - \mathbf{r}'|} \]  

where \( n_{xc,\lambda}(r, r') \) is the exchange-correlation hole at coupling strength \( \lambda \) [4, 24]. The on-top value of the hole is especially well approximated in LSD [23, 25, 26]:

\[ n_{xc,\lambda}(r, r) \approx n_{xc,\lambda}^{\text{uni}}[n_\uparrow(r), n_\downarrow(r); |\mathbf{r} - \mathbf{r}'| = 0], \]  

because the local approximation for the hole works best in the vicinity of the electron. In fact, Eq. (4) is exact for the exchange hole (\( \lambda = 0 \)) when the Kohn–Sham wave function is a single determinant; even at \( \lambda = 1 \), Eq. (4) is accurate over most of space, including both valence and intershell regions [26]. Furthermore, since both the LSD and the exact angle-averaged exchange-correlation holes,

\[ n_{xc,\lambda}(r, u) = \frac{1}{4\pi} \int n_{xc,\lambda}(r, \mathbf{r} + \mathbf{u}), \]  

satisfy the cusp condition [27, 28] at \( u = 0 \),

\[ n_{xc,\lambda}(r, u = 0) = \lambda [n_{xc,\lambda}(r, u = 0) + n(r)], \]  

where \( n_{xc,\lambda}(r, u = 0) = \partial n_{xc,\lambda}(r, u) / \partial u |_{u=0} \) this accuracy is extended [23] to the neighborhood \( r' \) close to \( r \). The hole also satisfies the charge conservation rule [29]:

\[ \int d^3r' n_{xc,\lambda}(r, r') = -1, \]  

both exactly and in LSD. Thus LSD describes both the size and the shape of this hole quite well. Finally, we note that only the system and angle averages [30, 31] of the hole appear in Eq. (3), i.e.,

\[ \langle n_{xc,\lambda}(u) \rangle = \frac{1}{N} \int d^3r n(r) n_{xc,\lambda}(r, u), \]  

and

\[ E_{xc} = N \int_0^1 d\lambda \int_0^\infty du \, 2\pi u \langle n_{xc,\lambda}(u) \rangle. \]  

The system-averaging unweights regions of small \( r \) (near the nucleus) in an atom, where LSD does less well, due to large higher-order reduced density gradients [26]. In Figure 1, we plot this average hole [32] at \( \lambda = 1 \) for Hooke’s atom [33], a model system consisting of two electrons in an external oscillator potential, with spring constant \( k = \frac{1}{2} \), both exactly and in LSD. This analytically solved model system has electron density values typical of those for real valence electrons, and so provides a fairly realistic test of density functional approximations. We see that the above restrictions ensure that the area under the curve, \( E_{xc,\lambda=1} \), also referred to as the potential contribution to \( E_{xc} \) [34], will almost always be moderately well approximated in LSD. (In fact, LSD does better at \( \lambda = 1 \) than at \( \lambda = 0 \) or for the coupling-constant average [26].) The reason that LSD satisfies all these constraints is because it approximates the hole by that of another physical system, the uniform electron gas. (Another choice, the hydrogen atom hole, has been successfully invoked as a model for the spherically averaged hole by Becke and Roussel [35, 36].)

**FIGURE 1.** Full-coupling strength system-averaged exchange-correlation hole, weighted by \( 2\pi u \), for the \( k = 1 / 4 \) Hooke’s atom, both exactly and within LSD. The area under the curve is \( E_{xc,\lambda=1} \) (see text).
**How a GGA Should Work**

From the preceding section, it should be clear that the reliability of LSD depends on more than just the particular choice of local energy per particle, \( e_{\text{unif}}^{\text{unif}}(n, n_z) \). Any GGA that hopes to emulate this reliability should retain the good features of LSD. The simplest form, already suggested by Kohn and Sham [1], is to expand the energy in a gradient expansion and include the \( |\nabla n|^2 \) terms. However, this gradient expansion approximation (GEA) often does worse than LSD [12] because the associated hole is not the hole of any physical system, and so violates Eq. (7) and other exact conditions.

Perdew [12] devised a simple procedure to overcome this difficulty, by removing all obviously unphysical contributions to the GEA hole, thereby restoring Eq. (7). This defines a numerical GGA which retains the good features of LSD, while improving the description of the average hole (and therefore the energy) by using the gradient. PW91 is an analytic parameterization of this numerical functional, which incorporates several further exact conditions. Thus PW91 is expected to improve LSD accuracy while preserving LSD reliability.

Clearly, functionals that have not been constructed in this way should not be expected to be as reliable. However, at the exchange (or \( \lambda = 0 \)) level, the Becke 88 functional [17] is quantitatively very similar to numerical GGA, and so should be reasonably reliable. Such agreement can be ascribed to the universal nature of the functional, to Becke’s use of the correct \( e_{\text{unif}}^{\text{unif}}(n, n_z) \), and to the relatively simple nature of exchange. However, LYP [18] underestimates the correlation energy of the uniform electron gas by about a factor of 2 [37], and so cannot (and is not) close to the results of numerical GGA. Any successes of this functional for specific systems should not be expected to transfer to other systems. The ground-state structure of \( C_{20} \) may be a case in point [38].

---

**Locally Correlated Hartree–Fock Approach**

It is now possible to incorporate exact exchange via construction of the optimized effective potential (OEP) [21] for an orbital-dependent functional. (In practice, one often uses the Krieger–Lee–Lafrate (KLI) approximation [39], which introduces negligible errors.) However, exact exchange is too non-local to be combined with any of the locally based correlation functionals discussed above. The exact exchange hole in a molecule is often much more diffuse than the exchange-correlation hole, as discussed in Ref. [40]. To overcome this difficulty, we suggest modeling the angle-averaged exchange-correlation hole of Eq. (5) by

\[
n_{\text{xc}, \lambda}(r, u) = n_{s}(r, u) K_{\text{xc}, \lambda}(r, u),
\]

where \( n_{s} = n_{\text{xc}, \lambda = 0} \), and \( K_{\text{xc}, \lambda}(r, u) \) is a local “correlation factor,” to be modeled using the restrictions already discussed. By including exact exchange, such a model retains the benefits of OEP [41]: It is self-interaction free, the exchange-correlation potential has the correct asymptotic behavior (\( \sim -1/r \)) far from infinite systems, it contains the derivative discontinuity with respect to particle number, etc. It has the further advantage of confining the correlation hole to regions of the system where the density is significant, a property not shared by LSD. This approach is similar in spirit to the Colle–Salvetti approximation for the wave function (Eqs. (1) and (2) of Ref. [42]).

The conditions given in the second section may be written as conditions on the correlating factor: Eq. (4) implies

\[
K_{\text{xc}, \lambda}(r, u = 0) = K_{\text{xc}, \lambda}^{\text{unif}}[n_{s}(r), n_{s}(r); u = 0],
\]

Eq. (6) implies (when the Kohn–Sham wave function is a single determinant)

\[
K'_{\text{xc}, \lambda}(r, u = 0) = \lambda \left[ K_{\text{xc}, \lambda}(r, u = 0) - \frac{2}{1 + \xi^{2}(r)} \right],
\]

where \( \xi(r) = [n_{s}(r) - n_{s}(r)]/n(r) \) is the local relative spin polarization, and the sum rule of Eq. (7) requires

\[
\int d^{3}u \ n_{s}(r, u) K_{\text{xc}, \lambda}(r, u) = -1.
\]

The first two of these conditions are simple enough to fulfill, but the last may require solving an integral equation at every point in the system. Finding a useful approximation for \( K_{\text{xc}, \lambda} \) is a challenge for the next generation of density functionals.

To demonstrate the power of this approach, we make a crude first gaff at the correlating factor,
with the form

\[ K_{c,\lambda}(\mathbf{r}, u) \equiv K_{xc,\lambda}(\mathbf{r}, u) - 1 \]

\[ \approx K_{c,\lambda}^{\text{unif}}\left[n_{\uparrow}(\mathbf{r}), n_{\downarrow}(\mathbf{r}); u = 0\right] e^{-u/d_x(\mathbf{r})} - C_x(\mathbf{r}) u^2 / \left[ u^2 + d_x^2(\mathbf{r}) \right]. \tag{14} \]

Here \( d_x(\mathbf{r}) \) is chosen to satisfy the cusp condition of Eq. (6),

\[ d_x(\mathbf{r}) = -\frac{n_{xc,\lambda}^{\text{unif}}\left[n_{\uparrow}(\mathbf{r}), n_{\downarrow}(\mathbf{r}); u = 0\right]}{n_{xc,\lambda}^{\text{unif}}\left[n_{\uparrow}(\mathbf{r}), n_{\downarrow}(\mathbf{r}); u = 0\right]}, \tag{15} \]

and \( C_x(\mathbf{r}) \) is chosen to satisfy the sum rule, Eq. (7). Thus this crude interpolation consists of a short-ranged contribution taken from LSD and a long-ranged contribution which restores the sum rule. Because of the linear sum of long- and short-ranged contributions, \( C_x(\mathbf{r}) \) may be found simply by evaluating two integrals over the spherically averaged exchange hole at each point in the system:

\[ C_x(\mathbf{r}) = K_{c,\lambda}^{\text{unif}}\left[n_{\uparrow}(\mathbf{r}), n_{\downarrow}(\mathbf{r}); u = 0\right] \times \frac{\int_0^\infty du 4\pi u^2 n_{\uparrow}(\mathbf{r}, u) \exp[-u/d_x(\mathbf{r})]}{\int_0^\infty du 4\pi u^2 n_{\uparrow}(\mathbf{r}, u) / \left[ u^2 + d_x^2(\mathbf{r}) \right]} \tag{16} \]

We are not suggesting this approximation for use in real electronic structure calculations, but merely as a pedagogical demonstration of the potential of this approach. In fact, we have shown elsewhere \[43\] that \( K_{xc,\lambda}(\mathbf{r}, u) \to \) a finite value as \( u \to \infty \), just as in Eq. (14). However, since it is unclear if this limit is achieved at any energetically meaningful value of \( u \), we do not use it to determine \( C_x(\mathbf{r}) \) in Eq. (16).

In Figure 2 we plot the spherically averaged correlation hole at full coupling strength for Hooke’s atom with spring constant \( k = 1/4 \), for several values of \( r \), both exactly and within the approximation of Eq. (14). Clearly this approximation is not too bad. We can define a correlation energy density via

\[ \epsilon_{c,\lambda}(\mathbf{r}) = 2\pi \int_0^\infty du u n_{c,\lambda}(\mathbf{r}, u), \tag{17} \]

which is plotted in Figure 3 within LSD, from Eq. (14), and exactly. We see that this function is very well described by Eq. (14).

In order to get the coupling-constant-averaged quantities, which yield the usual definition of the correlation energy, one could imagine applying Eq. (14) for each value of \( \lambda \), and integrating the results over \( \lambda \). However, a much simpler approach, and probably no less accurate, is to do only one calculation, in which all the quantities have already been averaged over \( \lambda \), i.e.,

\[ K_c(\mathbf{r}, u) \equiv K_{xc}(\mathbf{r}, u) - 1 \]

\[ \approx K_{c}^{\text{unif}}\left[n_{\uparrow}(\mathbf{r}), n_{\downarrow}(\mathbf{r}); u = 0\right] e^{-u/d_x(\mathbf{r})} - C(\mathbf{r}) u^2 / \left[ u^2 + d_x^2(\mathbf{r}) \right], \tag{18} \]

where the lack of explicit \( \lambda \)-dependence implies the averaged quantities. In this case, we have no

**Figure 2.** Angle-averaged full-coupling strength correlation energy per electron, as defined in Eq. (3), for the \( k = 1/4 \) Hooke’s atom; exactly, within LSD, and within Eq. (14)

**Figure 3.** Full-coupling strength correlation energy per electron, as defined in Eq. (3), for the \( k = 1/4 \) Hooke’s atom; exactly, within LSD, and within Eq. (14).
exact calculations of holes to compare with, but we can compare the energies. In Table I, we compare several results for the $k = \frac{1}{4}$ Hooke's atom. The crude model of Eq. (14) does better than even PW91 for both the full-coupling strength correlation energy and for the coupling-constant average. Even when exchange and correlation are added together, a fairer test against functionals which incorporate some cancellation of errors, Eq. (14) does best. Only when the full-coupling strength exchange-correlation energy is used, where the cancellation of errors in the LSD and PW91 functionals is greatest [26, 44], do they do better than the crude model. We make a similar comparison in Table II for He, finding here that the crude model is comparable to PW91.

Note that the crude model of Eq. (14) has not been designed to recover the uniform gas limit, a crucial component of a locally based functional. So we can test its accuracy on the uniform gas itself. In Figure 4, we compare the coupling-constant-averaged correlation energy in the unpolarized uniform electron gas from Eq. (18) with the exact value [45]. Clearly this crude approximation does not do well here, being particularly poor in the high-density ($r_s \rightarrow 0$) regime. In fact, it does even worse than LYP and so should not be considered as accurate enough for state-of-the-art calculations.

To understand the origin of this failure, we plot in Figure 5 the (coupling-constant-averaged) correlation factor of Eq. (18) for both a high-density ($r_s = 0.5$) and a low-density case ($r_s = 5$), as a

![Graph](image1)

**Figure 4.** Correlation energy in the spin-unpolarized uniform electron gas both exactly [7] and within Eq. (18).

![Graph](image2)

**Figure 5.** Correlation factor in the spin-unpolarized uniform electron gas both exactly [45] (solid lines) and within Eq. (18) (circles) for a high density ($r_s = 0.5$) and a low density ($r_s = 5$).

---

**Table I**

<table>
<thead>
<tr>
<th>Energy</th>
<th>LSD</th>
<th>PW91</th>
<th>Eq. (14)</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{c,\lambda=1}$</td>
<td>-0.133</td>
<td>-0.091</td>
<td>-0.058</td>
<td>-0.068</td>
</tr>
<tr>
<td>$E_c$</td>
<td>-0.086</td>
<td>-0.052</td>
<td>-0.032</td>
<td>-0.038</td>
</tr>
<tr>
<td>$E_{xc}$</td>
<td>-0.527</td>
<td>-0.546</td>
<td>-0.547</td>
<td>-0.554</td>
</tr>
<tr>
<td>$E_{xc,\lambda=1}$</td>
<td>-0.574</td>
<td>-0.581</td>
<td>-0.573</td>
<td>-0.583</td>
</tr>
</tbody>
</table>

**Table II**

<table>
<thead>
<tr>
<th>Energy</th>
<th>LSD</th>
<th>PW91</th>
<th>Eq. (14)</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{c,\lambda=1}$</td>
<td>-0.180</td>
<td>-0.083</td>
<td>-0.074</td>
<td>-0.079</td>
</tr>
<tr>
<td>$E_c$</td>
<td>-0.112</td>
<td>-0.046</td>
<td>-0.039</td>
<td>-0.042</td>
</tr>
<tr>
<td>$E_{xc}$</td>
<td>-0.996</td>
<td>-1.062</td>
<td>-1.063</td>
<td>-1.067</td>
</tr>
<tr>
<td>$E_{xc,\lambda=1}$</td>
<td>-1.034</td>
<td>-1.099</td>
<td>-1.098</td>
<td>-1.103</td>
</tr>
</tbody>
</table>
proximate correlation factor to drop very quickly at small \( y \), with a value of \( C = 0.05 \), producing far too little correlation energy. Thus, much work remains to be done to construct a reliable locally correlated Hartree–Fock functional.

**ACKNOWLEDGMENTS**

This work has been supported by National Science Foundation grant DMR95-21353 and by the Deutsche Forschungsgemeinschaft. After this work was performed, we learned that van Leeuwen and von Barth were also pursuing a similar line of reasoning [46].

**References**

43. M. Ernzerhof, K. Burke, and J. P. Perdew, unpublished.
46. R. van Leeuwen and U. von Barth, private communication.