PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2016, **18**, 21024

Avoiding the 4-index transformation in one-body reduced density matrix functional calculations for separable functionals

Klaas J. H. Giesbertz

Received 14th January 2016,
Accepted 12th February 2016

DOI: 10.1039/c6cp00303f

www.rsc.org/pccp

One of the major computational bottlenecks in one-body reduced density matrix (1RDM) functional theory is the evaluation of approximate 1RDM functionals and their derivatives. The reason is that more advanced approximate functionals are almost exclusively defined in the natural orbital basis, so a 4-index transformation of the two-electron integrals appears to be unavoidable. I will show that this is not the case and that so-called separable functionals can be evaluated much more efficiently, *i.e.* only at cubic cost in the basis size. Since most approximate functionals are actually separable, this new algorithm is an important development to make 1RDM functional theory calculations feasible for large electronic systems.

1 Introduction

Though density functional theory (DFT) is formally exact, practical density functionals have great difficulty in capturing strongly correlated phenomena such as the breaking of chemical bonds.[1–3] The calculation of excitation energies along the bond-breaking coordinate with the current approximations results in an even bigger disaster.[4–6] Also long-range charge transfer excitations pose a serious challenge for semi-local density functionals,[7,8] though some improvements have been reported with the help of range-separated hybrids,[9,10] direct modifications of the kernel[7,11,12] or the variational approach by Ziegler *et al.*[13–15]

One-body reduced density matrix (1RDM) functional theory provides a promising route for alleviating most of these problems existent in practical DFT. It has been demonstrated that the ground state energy of small singly bonded molecular systems can be reasonably well reproduced along the full bond-breaking coordinate.[16–20] It has been shown for two-electron systems that the time-dependent extension of the 1RDM functional is much more capable of dealing with bond breaking excitations and charge transfer excitations even within the adiabatic approximation, and also a significant amount of double excitations is captured.[21–24] Attempts are currently made to extend these results to general N-electron systems.[25–27] A more extensive overview of the current status of 1RDM functional theory can be found in ref. 28.

Section of Theoretical Chemistry, Faculty of Exact Sciences, VU University, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands.
E-mail: k.j.h.giesbertz@vu.nl

Though 1RDM functional theory has some appealing advantages compared to DFT, its practical use is currently limited due to two major computational bottlenecks. One computational hurdle is the excruciatingly slow self-consistent field (SCF) convergence to obtain the ground state energy. Although several algorithms have been proposed,[29–32] a significant breakthrough in this difficulty has not yet been achieved. Another computational complication is the evaluation of the 1RDM functionals themselves. Only the simplest approximate 1RDM functionals are explicitly defined in terms of the 1RDM. More advanced approximations are defined implicitly *via* the natural orbitals (NOs) and (natural) occupation numbers, which are defined as the eigenfunctions and eigenvalues of the 1RDM, respectively

$$\gamma(\mathbf{x},\mathbf{x}') = \sum_k n_k \phi_k(\mathbf{x}) \phi_k^*(\mathbf{x}'), \tag{1}$$

where $\mathbf{x} := \mathbf{r}\sigma$ is a combined space-spin coordinate. Current implementations therefore rely on a 4-index transformation of the two-electron integrals to the NO basis which is a very costly operation and impairs any calculation on systems with a large number of electrons. The situation is far worse than in correlated methods such as coupled cluster (CC) and configuration interaction (CI), since the 4-index transformation needs to be performed at each step of the SCF procedure.

I will show in this article that the 4-index transformation can actually be avoided for the so-called separable functionals. Separability allows a functional to be evaluated directly in the atomic orbital (AO) basis, or any other basis employed in the computer code. This reduces the computational cost from formal m^5 to a formal m^4 scaling, where m is the size of the basis set. Most integral routines make use of screening techniques to

only calculate significant two-electron integrals which further reduces the scaling to m^3 or even less. It turns out that most current approximate 1RDM functionals are separable functionals, with only a few exceptions.

This paper is organized as follows. First the algorithm for the evaluation of separable 1RDM functionals and the first order derivatives is explained in detail. Particular attention is needed for the so-called diagonal corrections which are sometimes called 'self-interaction corrections'. Special care needs to be taken to avoid excessive computational cost and memory imprint for these corrections. The evaluation of some particular functionals will be discussed to illustrate the algorithm. The next section presents benchmark results for the new algorithm using alkanes of varying length as test systems. A significant speed-up is obtained for the evaluation of separable functionals and is most drastic for functionals without diagonal corrections.

2 Algorithm

To avoid the 4-index transformation we generalize the direct approach in Hartree–Fock (HF) theory to calculate the contribution of the two-body interaction to the energy and the Fock matrix.[33] For example, consider the Hartree (classical Coulomb) contribution to the energy which in terms of the HF orbitals has the following simple form

$$W^{\mathrm{H}} = \frac{1}{2}\sum_{ij} n_i n_j [ii|jj], \tag{2}$$

where n_i are the occupation numbers of the HF orbitals, *i.e.* the NOs of the HF system. The two-electron integrals are denoted in the chemist's notation, so they are defined as

$$[ij|kl] := \int \mathrm{d}\mathbf{x}_1 \int \mathrm{d}\mathbf{x}_2 \phi_i^*(\mathbf{x}_1)\phi_j(\mathbf{x}_1) w(|\mathbf{r}_1 - \mathbf{r}_2|)\phi_k^*(\mathbf{x}_2)\phi_l(\mathbf{x}_2). \tag{3}$$

The interaction between the particles will be usually the Coulomb interaction $w(r) = 1/r$, but could also have some other form, *e.g.* the long-range part of the Coulomb interaction in a range-separated scheme $w(r) = \mathrm{erf}(\mu r)/r$.[34–36]

In an arbitrary basis, $\{\chi_\mu\}$, the Hartree term can be expressed as

$$W^{\mathrm{H}} = \frac{1}{2}\sum_{\mu\nu\kappa\lambda} \gamma_{\nu\mu}\gamma_{\lambda\kappa}[\mu\nu|\kappa\lambda], \tag{4}$$

where the matrix elements of the 1RDM are defined as

$$\gamma_{\mu\nu} := \int \mathrm{d}\mathbf{x} \int \mathrm{d}\mathbf{x}' \chi_\mu^*(\mathbf{x})\gamma(\mathbf{x},\mathbf{x}')\chi_\nu(\mathbf{x}'). \tag{5}$$

Throughout the text I will always use Greek indices to refer to this arbitrary basis in which the two-electron integrals are supplied. The latin indices will exclusively be used for the NO basis.

Suppose now that all the two-electron integrals are available in the basis $\{\chi_\mu\}$. Typically this will be an atomic orbital basis or a plane wave basis. The Hartree contribution can be calculated without transforming any 4-index quantity by first performing the contraction over only one 1RDM as

$$v^{\mathrm{H}}_{\mu\nu} = \sum_{\kappa\lambda} [\mu\nu|\kappa\lambda]\gamma_{\lambda\kappa}. \tag{6}$$

The summations are generally performed by looping over integrals stored on file or by calculating all integrals on the fly. In a second step the contraction with the other 1RDM is performed

$$W^{\mathrm{H}} = \frac{1}{2}\mathrm{Tr}\{\gamma \cdot \mathbf{v}^{\mathrm{H}}\} = \frac{1}{2}\sum_{\mu\nu} \gamma_{\nu\mu} v^{\mathrm{H}}_{\mu\nu}. \tag{7}$$

The only essential feature of the Hartree contribution to allow for this trick is that it is a Coulomb-type separable functional. With a Coulomb-type separable functional I mean a functional which can be expressed as a linear combination of a few terms of the form

$$W^{\mathrm{J}}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\sum_{\mu\nu\kappa\lambda} f_{\nu\mu} g_{\lambda\kappa}[\mu\nu|\kappa\lambda]. \tag{8}$$

This is a Coulomb-type separable functional which can efficiently be evaluated in the same fashion as the Hartree term. Typically we will have $\mathbf{f} = \mathbf{g}$ and $\mathbf{f} = \mathbf{f}^\dagger$, but this is not necessary for the trick to work. Its evaluation proceeds again *via* a Coulomb-like potential (6), which is generalized now to

$$v^{\mathrm{J}}_{\mu\nu}(\mathbf{g}) = \sum_{\kappa\lambda} [\mu\nu|\kappa\lambda] g_{\lambda\kappa}. \tag{9}$$

Likewise, an exchange-type separable functional can be expressed as a linear combination of a few terms of the form

$$W^{\mathrm{K}}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\sum_{\mu\nu\kappa\lambda} f_{\lambda\mu} g_{\nu\kappa}[\mu\nu|\kappa\lambda], \tag{10}$$

so just two indices are swapped around compared to the Coulomb case. The exchange-type separable functionals allow for a similar efficient evaluation to the Coulomb-type separable functionals by first forming an exchange-type potential

$$v^{\mathrm{K}}_{\mu\nu}(\mathbf{g}) = \sum_{\kappa\lambda} g_{\lambda\kappa}[\mu\lambda|\kappa\nu] \tag{11}$$

and subsequently performing the final contraction

$$W^{\mathrm{K}}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\sum_{\mu\nu} f_{\nu\mu} v^{\mathrm{K}}(\mathbf{g})_{\mu\nu} = \frac{1}{2}\mathrm{Tr}\{\mathbf{f} \cdot \mathbf{v}^{\mathrm{K}}(\mathbf{g})\}. \tag{12}$$

When the basis functions are real, the Coulomb- and exchange-type versions are the only two possible distinct forms of a separable functional. In the case one uses complex basis functions, *e.g.* plane waves, one also might need to consider the versions where the complex conjugation has been interchanged in the last pair of the two-electron integral, *i.e.* $[\mu\nu|\kappa\lambda] \rightarrow [\mu\nu|\lambda\kappa]$ in (8) and (10).

Not only the contribution to the energy can be calculated in this manner, but also the projected orbital derivatives

$$W_{kl} := \int \mathrm{d}\mathbf{x} \frac{\partial W}{\partial \phi_k(\mathbf{x})}\phi_l(\mathbf{x}), \tag{13}$$

needed for the construction of the Fock matrix[31,32] or direct orbital optimization[29] can easily be calculated. Only the final contraction over the potentials needs to be left out

$$W^{\mathrm{J/K}}_{\mu\nu}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\sum_{\kappa}\left(f_{\mu\kappa}v^{\mathrm{J/K}}_{\kappa\nu}(\mathbf{g}) + g_{\mu\kappa}v^{\mathrm{J/K}}_{\kappa\nu}(\mathbf{f})\right) \tag{14}$$

and subsequently a transformation to the NO basis needs to be made. For most approximate functionals, **f** and **g** will be diagonal in the NO basis. So it is computationally beneficial for transforming first the potentials to the NO basis and only then for multiplying them by the diagonal **f** and **g** matrices

$$W^{\mathrm{J/K}}_{kl}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\left(f_k v^{\mathrm{J/K}}_{kl}(\mathbf{g}) + g_k v^{\mathrm{J/K}}_{kl}(\mathbf{f})\right), \tag{15}$$

where f_k and g_k denote the diagonal entries of the **f** and **g** matrices in the NO representation. The energetic contribution can now cheaply be obtained by taking the trace

$$W^{\mathrm{J/K}}(\mathbf{f},\mathbf{g}) = \frac{1}{2}\mathrm{Tr}\left\{\mathbf{W}^{\mathrm{J/K}}(\mathbf{f},\mathbf{g})\right\}. \tag{16}$$

If the matrices **f** and **g** are not only diagonal in the NO basis, but if also their diagonal elements only depend on the occupation numbers with the same index, *i.e.* $f_k(\gamma) = f_k(n_k)$ and $g_k(\gamma) = g_k(n_k)$, the derivatives with respect to the occupation numbers become particularly simple

$$\frac{\partial W^{\mathrm{J/K}}}{\partial n_k} = \frac{1}{2}\left(\frac{\partial f_k}{\partial n_k}v^{\mathrm{J/K}}_{kk}(\mathbf{g}) + \frac{\partial g_k}{\partial n_k}v^{\mathrm{J/K}}_{kk}(\mathbf{f})\right). \tag{17}$$

Most approximate 1RDM functionals exhibit such a simple dependence on the occupation numbers. More complicated dependencies need to be worked out for each approximate functional separately.

Though separability might seem to be a very stringent condition on a general 1RDM functional, many approximate functionals used in 1RDM functional calculations are actually separable. The only non-separable functionals I am aware of are the empirical functional by Marques and Lathiotakis,[37] the automated version of the BBC3 functional by Rohr *et al.*[17] and the PNOF4 by Piris *et al.*[18] To use the proposed scheme to avoid the 4-index transformation, one only needs to rewrite the approximate 1RDM functional in a separable form, *i.e.* as a linear combination of terms of the form (8) and (10). An obvious example is the Müller functional,[38–40] since it was originally published in its separable form

$$W^{\mathrm{Müller}} = \frac{1}{2}\mathrm{Tr}\left\{\gamma\cdot\mathbf{v}^{\mathrm{J}}(\gamma) - \sqrt{\gamma}\cdot\mathbf{v}^{\mathrm{K}}(\sqrt{\gamma})\right\}. \tag{18}$$

A function of the 1RDM, the square root in this case, is defined in the usual manner *via* its diagonal representation

$$f(\gamma)_{\mu\nu} := \sum_i \langle\chi_\mu|\phi_i\rangle f(n_i)\langle\phi_i|\chi_\nu\rangle. \tag{19}$$

Approximate functionals which simply modify the square root to some other power, $\sqrt{\gamma}\to\gamma^\alpha$, also belong to this class of functionals in which the 4-index transformation can trivially be avoided.[41–44]

An explicit expression in terms of the 1RDM itself is not available for more advanced 1RDM functionals which classify NOs in strongly and weakly occupied groups and/or contain 'diagonal corrections'. Though these functionals are only explicitly defined in terms of occupation numbers and NOs, many of these functionals can still be rewritten in a separable form, allowing for the previously described tricks. These non-trivial separable forms of more advanced 1RDM functionals are best explained with an example. To this end we will use the BBC2 functional, which is defined in the NO representation as[16]

$$W^{\mathrm{BBC2}} := W^{\mathrm{H}} + \frac{1}{2}\sum_{ij}F(n_i,n_j)[ij|ji], \tag{20}$$

where W^{H} is the usual Hartree term introduced earlier (2) and $F(n_i, n_j)$ is defined as

$$F(n_i,n_j) := \begin{cases} \sqrt{n_i n_j} & \text{for } i\neq j \text{ and } n_i,n_j < 1/2 \\ -n_i n_j & \text{for } i\neq j \text{ and } n_i,n_j \geq 1/2 \\ -\sqrt{n_i n_j} & \text{otherwise.} \end{cases} \tag{21}$$

The involved expression for $F(n_i,n_j)$ renders an explicit expression in terms of γ virtually impossible. Nevertheless, the BBC2 functional can still be expressed in a separable form by using other (auxiliary) matrices. To this end we first neglect the $i \neq j$ conditions in (21). This 'non-diagonal part' of the BBC2 functional can now be expressed in a separable form as

$$\begin{aligned} W^{\mathrm{BBC2}}_{\text{no diag}} = \frac{1}{2}\mathrm{Tr}\big\{&\gamma\cdot\mathbf{v}^{\mathrm{J}}(\gamma) + \sqrt{\gamma}^{\mathrm{virt}}\cdot\mathbf{v}^{\mathrm{K}}\big(\sqrt{\gamma}^{\mathrm{virt}}\big) \\ &-2\sqrt{\gamma}^{\mathrm{virt}}\cdot\mathbf{v}^{\mathrm{K}}\big(\sqrt{\gamma}^{\mathrm{occ}}\big) - \gamma^{\mathrm{occ}}\cdot\mathbf{v}^{\mathrm{K}}\big(\gamma^{\mathrm{occ}}\big)\big\}, \end{aligned} \tag{22}$$

where

$$f(\gamma)^{\mathrm{occ}}_{\mu\nu} := \sum_{n_i\leq 1/2}\langle\chi_\mu|\phi_i\rangle f(n_i)\langle\phi_i|\chi_\nu\rangle, \tag{23a}$$

$$f(\gamma)^{\mathrm{virt}}_{\mu\nu} := \sum_{n_i > 1/2}\langle\chi_\mu|\phi_i\rangle f(n_i)\langle\phi_i|\chi_\nu\rangle. \tag{23b}$$

The remaining diagonal part (correction) is now of the form

$$W^{\mathrm{BBC2}}_{\text{diag}} = \frac{1}{2}\sum_{n_i\leq 1/2}\left(n_i^2 - n_i\right)[ii|ii] - \sum_{n_i > 1/2} n_i[ii|ii]. \tag{24}$$

Unfortunately the diagonal correction cannot be straightforwardly be written in a separable form and it seems that we still need to resort to a 4-index transformation of the two-electron integrals for their evaluation. However, the proposed trick can still be used by first constructing 1RDMs in which only one NO is occupied

$$\bar{\gamma}^{(i)}_{\mu\nu} := \langle\chi_\mu|\phi_i\rangle\langle\phi_i|\chi_\nu\rangle. \tag{25}$$

The next step is to form the contraction with the two-electron integrals as

$$v^{(i)}_{\mu\nu} = \sum_{\kappa\lambda}[\mu\nu|\kappa\lambda]\bar{\gamma}^{(i)}_{\lambda\kappa}. \tag{26}$$

The last step is to transform the potentials $\mathbf{v}^{(i)}$ back to the NO representation and to form the final contraction

$$W^{\mathrm{diag}}(\mathbf{d}) = \frac{1}{2}\sum_i d_i v_{ii}^{(i)}, \tag{27}$$

where the elements d_i depend on the particular form of the approximate 1RDM functional under consideration. For the BBC2 functional we have

$$d_i = \begin{cases} n_i^2 - n_i & \text{for } n_i \geq 1/2 \\ -2n_i & \text{for } n_i < 1/2 \end{cases}. \tag{28}$$

The corresponding projected orbital derivatives required for the SCF can be obtained from the off-diagonal elements

$$W_{kl}^{\mathrm{diag}}(\mathbf{d}) = d_k v_{kl}^{(k)}. \tag{29}$$

Though we could avoid the use of a 4-index transformation in this manner, the additional computational cost to calculate the diagonal correction of the BBC2 functional (24) is significant compared to the cost to calculate the non-diagonal part (22). The non-diagonal part only needs the contraction with 4 different auxiliary matrices (1 Coulomb and 3 exchange), whereas the diagonal part requires a contraction with m auxiliary matrices, so comprises a significantly more expensive part of the functional to evaluate. Furthermore, the complete construction of the orbital 1RMDs $\bar{\gamma}^{(i)}$ and the corresponding potentials $\mathbf{v}^{(i)}$ gives a significant memory imprint, since both scale cubically with the number of basis functions. Fortunately, the special structure of the diagonal correction allows one to avoid the explicit construction of these large matrices. Avoiding the explicit construction of the orbital 1RDMs, $\bar{\gamma}^{(i)}$, is readily achieved by not constructing them explicitly. Instead, the required elements are only constructed on a need-to-be basis when looping over the two-electron integrals.

Now let us consider the high memory imprint of the potentials $\mathbf{v}^{(i)}$. Note that in the expression for the energy contribution (27) and for the orbital derivative (29), at least one of the lower indices is always equal to the upper index, so we can avoid the construction of many unnecessary elements. We do this by transforming the potentials $\mathbf{v}^{(i)}$ partially to the NO basis immediately during their construction as

$$\begin{aligned} \bar{v}_{i\nu} := v_{i\nu}^{(i)} &= \sum_{\mu\kappa\lambda} \langle \phi_i | \chi_\mu \rangle [\mu\nu|\kappa\lambda] \bar{\gamma}_{\lambda\kappa}^{(i)} \\ &= \sum_{\mu\kappa\lambda} \langle \phi_i | \chi_\mu \rangle [\mu\nu|\kappa\lambda] \langle \chi_\lambda | \phi_i \rangle \langle \phi_i | \chi_\kappa \rangle. \end{aligned} \tag{30}$$

The projected orbital derivatives and the contribution to the energy can now easily be calculated by transforming the last index also to the NO basis and forming

$$W_{kl}^{\mathrm{diag}}(\mathbf{d}) = d_k \bar{v}_{kl}, \tag{31a}$$

$$W^{\mathrm{diag}}(\mathbf{d}) = \frac{1}{2}\sum_k d_k \bar{v}_{kk} = \frac{1}{2}\mathrm{Tr}\{\mathbf{W}^{\mathrm{diag}}\}. \tag{31b}$$

Though we have now avoided the explicit construction of the $\bar{\gamma}^{(i)}$ matrices and the corresponding potentials $\mathbf{v}^{(i)}$, the operation count for the calculation of the diagonal correction is significantly higher than that for the non-diagonal part of the functional. For separable functionals with a diagonal correction, the evaluation of the diagonal part therefore remains the computational bottleneck in the evaluation of their values and derivatives.

The formal scaling of the proposed algorithm for functionals without diagonal corrections is still of order m^4 due to the loop over the two-electron integrals to construct the Coulomb potentials (9) and exchange potentials (11). The situation is even worse for functionals with diagonal corrections, since the construction of the intermediate potentials $\bar{\mathbf{v}}$ (30) makes it formally an m^5 process. The main advantage of the proposed algorithm is that integral screening becomes way more effective than in a calculation *via* a 4-index transformation. Screening of the two-electron integrals is a very common strategy in quantum chemistry software to avoid the calculation of many insignificant two-electron integrals. The number of significant two-electron integrals turns out to be only of the order m^2 for the larger molecular systems of interest,[45] so the scaling can in principle be reduced by an additional factor of 2. This is readily done by using the Schwarz inequality on the two-electron integrals[46]

$$0 \leq [ij|kl] \leq \sqrt{[ij|ij]}\sqrt{[kl|kl]}. \tag{32}$$

The integrals on the right-hand side are pre-calculated and only require m^2 storage. Only when the product of the square roots on the right-hand side is larger than some tolerance ε, the program will actually calculate the integrals on the left. Screening of two-electrons is particularly effective when the integrals are evaluated in a basis with a strong local character. This method and more advanced techniques have been implemented in virtually all quantum chemistry software packages, so can directly be exploited to reduce the scaling significantly as I will demonstrate in the following section.

It should be mentioned that the importance of separability for an efficient evaluation of energy expressions and derivatives has already been recognised for a few decades in quantum chemistry. The most famous example is the 2nd order Müller–Plesset (MP2) energy correction, which in its usual form is non-separable

$$E^{\mathrm{MP2}} = -\frac{1}{4}\sum_{ijab} \frac{|[ai|bj] - [aj|bi]|^2}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j}, \tag{33}$$

where the indices i, j refer to the occupied Hartree–Fock (HF) orbitals, the indices a, b to the unoccupied HF orbitals and ε_{r} are the HF orbital energies. By writing the denominator as a Laplace transform, the MP2 energy correction can be turned into a separable form[47–49]

$$E^{\mathrm{MP2}} = -\frac{1}{4}\int_0^\infty \mathrm{d}t \sum_{ijab} \mathrm{e}^{-(\varepsilon_a+\varepsilon_b-\varepsilon_i-\varepsilon_j)t} |[ai|bj] - [aj|bi]|^2. \tag{34}$$

The summation over the HF orbitals is now a separable expression and the strategy described before can now be used to evaluate the integrant efficiently. The price to pay is that the contraction needs to be evaluated for several values of t to

evaluate the integral numerically with some suitable numerical integration scheme, but this only increases the computational cost by a prefactor. A similar trick can be used to bring the random phase approximation (RPA) in the linear scaling regime *via* the spherical Laplace transform.[50] Similar tricks will probably be useful to rewrite additional approximate 1RDM functionals in a separable form.

3 Benchmarking

To test the new strategy to evaluate the energy of separable functionals, I have constructed a modular Fortran 2003/2008 implementation, interfaced to a modified version of the Gamess-US program package.[51–53] The build-system of Gamess-US has been replaced by foray,[54] to avoid figuring out module dependencies by hand and allow for parallel compilation. The implementation is currently only intended for serial runs. A parallel implementation will be considered later. The cut-off criterion for the two-electron integrals has been set to $\varepsilon = 10^{-10}$.

The evaluation of the energy and gradients has been implemented in 3 different manners

(1) The straightforward version using the 4-index transformation of the two-electron integrals.

(2) Evaluation in the AO basis using integrals stored on disk.

(3) A direct option which does not store the two-electron integrals, but recalculates them each time when they are needed.

In Fig. 1 I plot timings for the Müller functional (18). For the input 1RDM I have used the HF orbitals as NOs. Typically, the occupation numbers are fractional in 1RDM calculations. This has been mimicked to some extend by setting the occupations of the occupied HF orbitals to 0.9 and the occupations of the virtual HF orbitals to $0.1N/(m - N)$. As test systems I used (linear) alkanes C_nH_{2n+2} in a spherical cc-pVTZ basis.[55] The highest point group symmetry of each system has been used to calculate only the symmetry unique integrals. For even n the point group is C_{2h} and for odd n this is C_{2v}. The two smallest alkanes allow for the use of higher point groups (T_d for methane and D_{3d} for ethane). This additional reduction in symmetry unique two-electron integrals is reflected by the strong deviation from the general trend in the plot in Fig. 1. The asymptotic scaling behavior (exponent) of the different strategies has been estimated by fitting a power function, $Ae^{\alpha m}$, to the results for a large enough basis size. The points taken into account for the fit vary per calculation and have been indicated by the straight line connecting the points in Fig. 1 and 2.

The naïve implementation *via* the 4-index transformation of the two-electron integrals is clearly the most expensive one in Fig. 1. Its main cost is the 4-index transformation, which is reflected in its high asymptotic scaling of order $m^{4.8}$. The evaluation in the AO basis is clearly more efficient. Reading the integrals from the file is slightly faster than recalculating them, but the asymptotic scaling is basically the same. In this benchmark, the file reading is particularly fast due to the solid state drive (SSD) and the direct option is actually slow, since only one core is used. It is therefore expected that a parallel implementation would easily beat the 'AO file' option, since the different processes would only start to compete for disk access. An additional bottleneck for the 'AO file' option is that

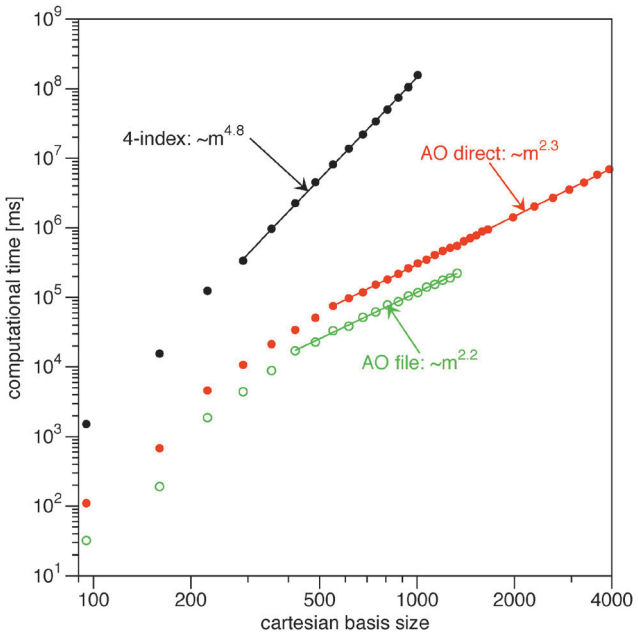

Fig. 1 Log–log plot of the computational time for the evaluation of the energy and the gradient from the Müller functional (18) against the number of primitive basis functions in the calculation. The straight lines indicate which points have been included in the fits to extract the exponents.

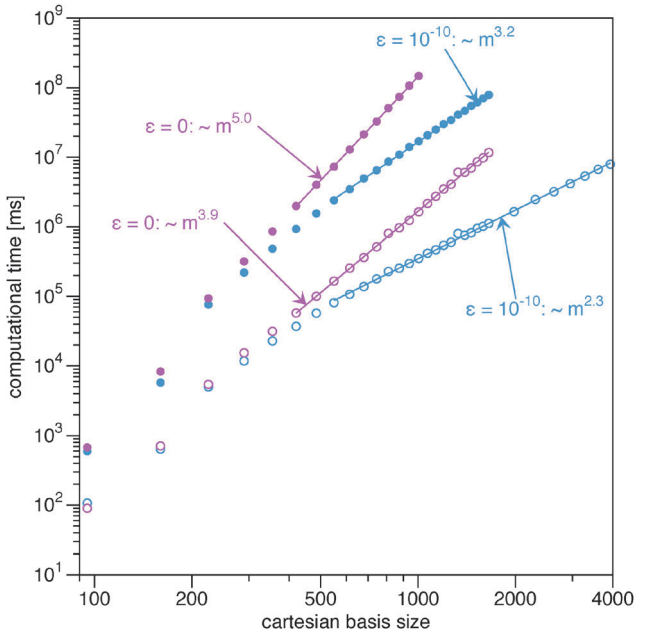

Fig. 2 Log–log plot of the computational time for the evaluation of the energy and the gradient from the BBC2 functional. The filled circles correspond to the full BBC2 functional and the open circles to only the non-diagonal part (22). The straight lines indicate which points have been included in the fits to extract the exponents.

the two-electron integral file becomes excessively large. The last reported point for the 'AO file' option ($C_{20}H_{42}$: 1330 cartesian basis functions) has already a two-electron file of 250 GB. The AO direct option does not have this bottleneck and can easily handle larger systems. The scaling of the 'AO direct' option has been estimated to be of order $m^{2.3}$. The exponent has not converged however, since the formation of the required 1RDMs in the AO basis (γ and $\sqrt{\gamma}$) involves a matrix–matrix product which should scale worse. Assuming the BLAS implementation is based on Strassen's algorithm,[56] an asymptotic scaling of at least order $m^{\log_2 7} \approx m^{2.8}$ would be expected.

In Fig. 2 the computational time for the evaluation of the BBC2 functional with and without its diagonal part (24) is shown when using the 'AO direct' option. The first thing to notice is the effectiveness of screening by the Schwarz inequality. Both for the full BBC2 functional and only the non-diagonal part, the use of screening leads to a drastic reduction in the asymptotic scaling.

Now let us concentrate on the calculations with only the non-diagonal part of the BBC2 functional (22). Though the non-diagonal part of the BBC2 functional requires the contraction with 4 different 1RDMs, instead of only 2 as required for the Müller functional, the increase in computational cost is marginal. Only the prefactor is increased slightly by $(A_{\mathrm{BBC2}} - A_{\mathrm{Müller}})/A_{\mathrm{Müller}} = 5\%$, but the asymptotic scaling remains the same, $m^{2.3}$.

On the other hand, the calculation of the diagonal corrections poses a significant increase in computational cost. They add an order of magnitude to the computational cost. If no screening is used, this even means that the asymptotic scaling is similar to the scaling of the 4-index transformation. Avoiding the use of small integrals is therefore crucial for the algorithm to improve over the standard evaluation by the 4-index transformation. If screening is used, however, the AO based evaluation of the diagonal correction provides a major increase in efficiency compared to the traditional approach *via* a 4-index transformation. The asymptotic computational cost has been reduced from order m^5 to $m^{3.2}$, so much larger quantum systems are accessible with the new implementation, especially when the algorithm is additionally parallelized.

The effect of the cutoff criterion on the total energies and gradients has also been investigated. In the upper panel of Fig. 3 the relative absolute error in the energy, $|(E_\varepsilon - E_0)/E_0|$, has been plotted as a function of the cartesian basis size for the alkane series. For the full range of alkanes, the error in the energy can be systematically reduced by lowering the value of the cutoff parameter, ε. Only for the small alkanes the reduction in the error is more erratic, probably due to more or less error cancelations. In the lower panel of Fig. 3 the averaged absolute error (open circles) and maximum error (filled circles) are shown for the alkane series. Again, the error in the gradient can be systematically reduced by lowering the cutoff parameter, ε.

In Fig. 4 I show the same errors for the full BBC2 functional for the different cutoff parameters, $\varepsilon = \{10^{-9}, 10^{-10}, 10^{-11}\}$. We see that the inclusion of the diagonal contributions to the BBC2 functional (24) looks very similar for the larger alkanes, so the error in the energy and the gradient in the same ballpark when the diagonal part of the BBC2 functional is included. Only for the smallest alkanes the errors are somewhat larger, probably due to less fortuitous error cancellations.

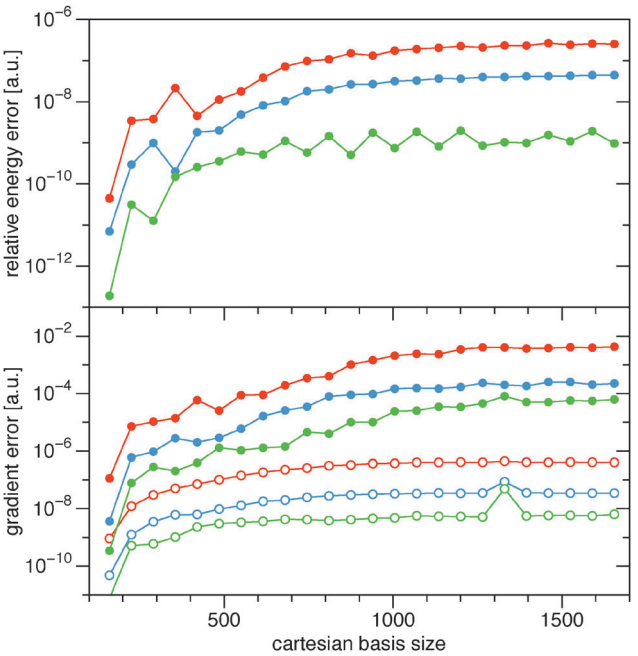

Fig. 3 Plot of the relative absolute errors in the calculation of the energy (upper panel) and the gradient (lower panel) of the non-diagonal part of the BBC2 functional for different cutoff criteria: upper curve (red) $\varepsilon = 10^{-9}$, middle curve (blue) $\varepsilon = 10^{-10}$, and lower curve (green) $\varepsilon = 10^{-11}$. The full circles in the gradient panel are the maximum errors in the gradient and the open circles are the averaged absolute error in the gradient.

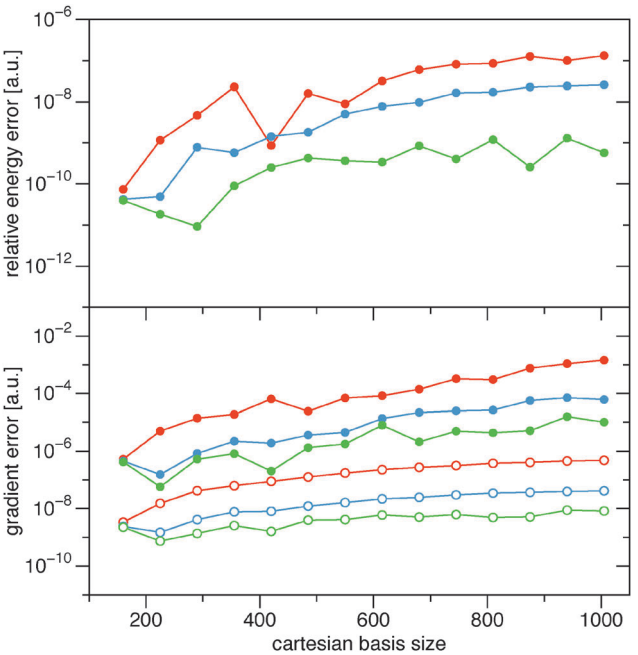

Fig. 4 Similar to that in Fig. 3, though now for the full BBC2 functional. Since the full BBC2 functional is much more expensive, only calculations up to $C_{15}H_{32}$ have been included in the comparison.

4 Conclusions

An algorithm to avoid the 4-index transformation for non-trivial separable 1RDM functionals has been presented, which reduces the formal computational cost to m^4. Diagonal corrections could be handled using a similar strategy, though additional modifications were needed to avoid the excessive use of memory. The formal scaling of the diagonal corrections remains m^5 unfortunately. The main advantage of the newly proposed strategy for the evaluation of 1RDM functionals and their derivatives is that now integral screening techniques can be used more effectively, leading to a significant reduction in the asymptotic scaling of the computational cost. Using screening by the Schwarz inequality, I have shown that the asymptotic scaling can be reduced to $m^{3.2}$ for separable functionals including diagonal corrections. For separable functionals without diagonal corrections the asymptotic scaling is even further reduced to $m^{2.3}$. It is expected, however, that for larger systems the scaling for functionals without diagonal corrections should be of about order $m^{2.8}$ due to the matrix–matrix products involved. All in all, a significant boost in the speed of the evaluation of approximate 1RDM functionals has been achieved, as the evaluation was one of the major computational bottlenecks in practical 1RDM functional calculations. This is an important step to make use of 1RDM functionals feasible for large molecular systems.

Acknowledgements

I am happy to dedicate this paper to Evert Jan Baerends, who has been a great PhD supervisor. I especially appreciate his confidence and trust that I would find my own way and develop my interests. Our discussions have always been very helpful to me to structure and organize my ideas well beyond my PhD. Support from the Netherlands Foundation for Research NWO (722.012.013) through a VENI grant is gratefully acknowledged.

References

1 J. P. Perdew, R. G. Parr, M. Levy and J. L. Balduz, *Phys. Rev. Lett.*, 1982, **49**, 1691–1694.
2 M. M. Ossowski, L. L. Boyer, M. J. Mehl and M. R. Pederson, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2003, **68**, 245107.
3 A. Ruzsinszky, J. P. Perdew, G. I. Csonka, O. A. Vydrov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 194112.
4 N. T. Maitra and D. G. Tempel, *J. Chem. Phys.*, 2006, **125**, 184111.
5 K. J. H. Giesbertz and E. J. Baerends, *Chem. Phys. Lett.*, 2008, **461**, 338.
6 J. I. Fuks, A. Rubio and N. T. Maitra, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2011, **83**, 042501.
7 M. E. Casida, F. Gutierrez, J. Guan, F.-X. Gadea, D. Salahub and J.-P. Daudey, *J. Chem. Phys.*, 2000, **113**, 7062.
8 A. Dreuw, J. L. Weisman and M. Head-Gordon, *J. Chem. Phys.*, 2003, **119**, 2943.
9 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51.
10 M. J. G. Peach, P. Benfield, T. Helgaker and D. J. Tozer, *J. Chem. Phys.*, 2008, **128**, 044118.
11 O. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2004, **121**, 655.
12 J. Neugebauer, O. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2006, **124**, 214102.
13 T. Ziegler, M. Seth, M. Krykunov and J. Autschbach, *J. Chem. Phys.*, 2008, **129**, 184114.
14 T. Ziegler, M. Seth, M. Krykunov, J. Autschbach and F. Wang, *J. Chem. Phys.*, 2009, **130**, 154102.
15 T. Ziegler and M. Krykunov, *J. Chem. Phys.*, 2010, **133**, 074104.
16 O. V. Gritsenko, K. Pernal and E. J. Baerends, *J. Chem. Phys.*, 2005, **122**, 204102.
17 D. R. Rohr, K. Pernal, O. V. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2008, **129**, 164105.
18 M. Piris, J. M. Matxain, X. Lopez and J. M. Ugalde, *J. Chem. Phys.*, 2010, **133**, 111101.
19 Ł. M. Mentel, R. van Meer, O. V. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2014, **140**, 214105.
20 M. Piris, *J. Chem. Phys.*, 2014, **141**, 044107.
21 K. J. H. Giesbertz, E. J. Baerends and O. V. Gritsenko, *Phys. Rev. Lett.*, 2008, **101**, 033004.
22 K. J. H. Giesbertz, K. Pernal, O. V. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2009, **130**, 114104.
23 K. J. H. Giesbertz, O. V. Gritsenko and E. J. Baerends, *Phys. Rev. Lett.*, 2010, **105**, 013002.
24 K. J. H. Giesbertz, O. V. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2010, **133**, 174119.
25 K. Pernal, *J. Chem. Phys.*, 2012, **136**, 184105.
26 K. Chatterjee and K. Pernal, *J. Chem. Phys.*, 2012, **137**, 204109.
27 R. van Meer, O. V. Gritsenko and E. J. Baerends, *J. Chem. Phys.*, 2014, **140**, 024101.
28 K. Pernal and K. J. H. Giesbertz, *Density-Functional Methods for Excited States*, Springer, Berlin, Heidelberg, 2015, vol. 368, pp. 125–183.
29 J. Cioslowski and K. Pernal, *J. Chem. Phys.*, 2001, **115**, 5784.
30 A. J. Cohen and E. J. Baerends, *Chem. Phys. Lett.*, 2002, **364**, 409–419.
31 K. Pernal, *Phys. Rev. Lett.*, 2005, **94**, 233002.
32 M. Piris and J. M. Ugalde, *J. Comput. Chem.*, 2009, **30**, 2078–2086.
33 J. Almlöf, K. Faegri, Jr. and K. Korsell, *J. Comput. Chem.*, 1982, **3**, 385–399.
34 A. Savin, *Int. J. Quantum Chem.*, 1988, **34**, 59–69.
35 A. Savin, in *Recent Advances in Density Functional Methods*, ed. D. P. Chong, World Scientific, Singapore, 1995, ch. Beyond the Kohn–Sham determinant.
36 D. R. Rohr, J. Toulouse and K. Pernal, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2010, **82**, 052502.
37 M. A. L. Marques and N. N. Lathiotakis, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2008, **77**, 032509.
38 A. M. K. Müller, *Phys. Lett. A*, 1984, **105**, 446–452.
39 M. Buijse, PhD thesis, Vrije Universiteit, De Boelelaan 1105, Amsterdam, The Netherlands, 1991.

40 M. Buijse and E. J. Baerends, *Mol. Phys.*, 2002, **100**, 401–421.
41 A. Holas, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1999, **59**, 3454–3461.
42 G. Csányi and T. A. Arias, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2000, **61**, 7348–7352.
43 S. Sharma, J. K. Dewhurst, N. N. Lathiotakis and E. K. U. Gross, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2008, **78**, 201103(R).
44 N. N. Lathiotakis, S. Sharma, J. K. Dewhurst, F. G. Eich, M. A. L. Marques and E. K. U. Gross, *Phys. Rev. A: At., Mol., Opt. Phys.*, 2009, **79**, 040501.
45 T. Helgaker, P. Jørgensen and J. Olsen, *Molecular Electronic-Structure Theory*, John Wiley & Sons, LTD, West Sussex, 2000.
46 J. L. Whitten, *J. Chem. Phys.*, 1973, **58**, 4496–4501.
47 J. Almlöf, *Chem. Phys. Lett.*, 1991, **181**, 319–320.
48 M. Häser and J. Almlöf, *J. Chem. Phys.*, 1992, **96**, 489.
49 M. Häser, *Theor. Chim. Acta*, 1993, **87**, 147–173.
50 H. F. Schurkus and C. Ochsenfeld, *J. Chem. Phys.*, 2016, **144**, 031101.
51 M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, *J. Comput. Chem.*, 1993, **14**, 1347–1363.
52 M. S. Gordon and M. W. Schmidt, *Theory and Applications of Computational Chemistry*, Elsevier, Amsterdam, 2005, ch. 41, pp. 1167–1189.
53 *Gamess-US*, http://www.msg.chem.iastate.edu/GAMESS/GAMESS.html.
54 D. A. McCormack, *Foray Tool*, https://bitbucket.org/siege/foraytool/downloads.
55 T. H. Dunning Jr., *J. Chem. Phys.*, 1989, **90**, 1007.
56 V. Strassen, *Numer. Math.*, 1969, **13**, 354–356.