Density Fitting notes

June 14, 2017

DOI: 10.1063/1.4864755
The author is David Hollman, Schaefer, and Valeev

The high computational cost of LCAO comes from computing the two-electron integrals

$$g_{\mu\nu\lambda\sigma} = \int \chi_\mu(\vec{r_1})\chi_\nu(\vec{r_1})\frac{1}{\vec{r_{12}}}\chi_\lambda(\vec{r_2})\chi_\sigma(\vec{r_2})$$

scaling to $\mathcal{O}(N^4)$. The number of significant entries need to compute can be reduced using Schwarz inequality.

$$|(\mu\nu|\lambda\sigma)| \leq (\mu\nu|\mu\nu)^{1/2}(\lambda\sigma|\lambda\sigma)^{1/2}$$

But the prefactor to this calculation is still large growing with basis set so it is still difficult to use. There are efforts to find a more aggressive screening technique.

Local approximations have increased since 2009 with PNO's, local PNO, and domain-based PNO. Essential to the computation of LPNO's is the decomposition of the two electron integrals **g** using density fitting or resolution of the identity.

The density fitting approximation expresses the two electron densities, composing the bra and ket product densities components of **g** as a linear combination of basis function from an auxiliary basis:

$$|\mu\nu) = \sum_X C^X_{\mu\nu}|X)$$

where X, Y, ... indicate indices in an auxiliary basis set. The expansion is exact if $\{|X)\}$ Spans the same space as $\{\mu\nu\}$. In practice it is an an approximation for molecules, since atom-centered Gaussian basis functions are typically used for both the orbitals and auxiliary basis thus the products will, in general, be sums of gaussian centered between the centers of μ and ν. The auxiliary basis is typically 3 times the size of the orbital basis to reduce fitting errors to smaller than other sources of error.

The coefficients are solved for by left-projection by an auxiliary basis bra $(Y|$ and the coulomb operator yielding a system of linear equations,

$$(Y|\mu\nu) = \sum_X C^X_{\mu\nu}(Y|X)$$

This is solved efficiently by decomposing the (positive semidefinite) matrix $(Y|X)$ using standard linear algebra techniques and $(Y|\mu\nu)$ includes the implicit application of the coulomb kernel $\frac{1}{r_{12}}$
The decomposition of the $(Y|X)$ and computation of the $(Y|\mu\nu)$ require $\mathcal{O}(N^3)$ and can be reduced using Schwarz screening.

This paper describes the means of reducing the cost of solving to linear in N without decreasing the accuracy beyond other sources of error.

Theoretical Background

This paper made choices on 2 generalizations: The fitting metric $\hat{M}$ and the fitting basis $\{|X)\}$ thus the equation

$$(Y_{[\mu\nu]}|\hat{M}|\mu\nu) = \sum_{X\in[\mu\nu]} C_{\mu\nu}^{X_{[\mu\nu]}} (Y_{[\mu\nu]}|\hat{M}|X_{[\mu\nu]})$$

where $[\mu\nu]$ is some domain specific to a given μ and ν for which the $X_{[\mu\nu]}$ are drawn.

The overlap metric is good because it decays instantly as apposed to the inverse linear behavior of the coulomb metric but many test have shown that the overlap metric gives errors about an order of magnitude worse than the coulomb metric.

An alternative approach to reducing the complexity is to fit a given product $|\mu\nu)$ with some subset of the auxiliary basis. Later proposed was the construction of fitting domains based on distance cut-offs, which lead to discontinuities in the potential energy surface. To alleviate this issue methods have been proposed in which the fitting functions for a given density are attenuated using a continuous and n-differentiable "bump" function that smoothly excludes auxiliary functions near the domain boundaries. This leads to PES artifacts, though they are smooth and continuous.

A second approach is to eliminate domain boundaries and include only fitting functions for densities $|\mu\nu)$ that are on the same centers as μ and ν. Using the notation $|\mu_a\nu_b)$ to indicate the functions μ is on a center a and ν is on a center b and $|X_{(ab)})$ to indicate that X is either on center a or b. Referred to as concentric atomic density fitting

The series expansion of the **g** as

$$g_{\mu\nu,\lambda\sigma} \equiv (\mu\nu|\lambda\sigma) \approx \sum_{XY} C_{\mu\nu}^{X} (X|Y) C_{\lambda\sigma}^{Y}$$

is only correct to first order in the density error, unless the coulomb metric is used. A more robust formula, accurate to second order error with the non-coulombic metrics:

$$g_{\mu\nu,\lambda\sigma} \approx \sum_{X} C_{\mu\nu}^{X} (X|\lambda\sigma) + \sum_{Y} (\mu\nu|Y) C_{\lambda\sigma}^{Y} - \sum_{XY} C_{\mu\nu}^{X} (X|Y) C_{\lambda\sigma}^{Y}.$$

The first use of this robust formulation in the context of CADF is the pair-atomic resolution of the identity in which the auxiliary basis functions were constructed via an atomic Cholesky decomposition.
A problem in the development of PARI they found that using a local or non-coulomb metrix the approximate **g** tensor is no longer manifestly positive semidefinite. this can converge on a density **D** that charaterizes unphysical "attractive electrons" the unphysical convergence is exceedingly rare near the PES minima. One solution was to add fitting functions on a line between the centers a and b. But this extension slowed down their computation by a factor of 20-70 for their largest computations. This paper present an extension to robust CADF that only requires about 10% extra effort for similarly sized computations with less percentage additional effort for larger molecules.

To address the loss of positive semidefiniteness in the approximate **g** tensor, a mixed approach in which atom-wise semidiagonal integrals $(\mu_a\nu_b|\lambda_a\sigma_b)$ are included exactly and other integrals are approximated using robust concentric atomic density fitting (see paper for formulation)

results to come later

DF Approximations to the ERI
David Sherrill

1 Introduction

The paper is based on the paper by Werner, the first reference and will is the next review.
Density fitting is a way to approximate the 2-electron integrals,

$$(pq|rs) = \int dr_1 \int dr_2 \phi_p(r_1)\phi_q(r_1)\frac{1}{r_{12}}\phi_r(r_2)\phi_s(r_2)$$

where we have assumed real orbitals. One can consider this electron repulsion integral as the repulsion between 2 generalized electron densities

$$(pq|rs) = \int dr_1 \int dr_2 \rho_{pq}(r_1)\frac{1}{r_{12}}\rho_{rs}(r_2)$$

where $\rho_{pq}(r) = \phi_p(r)\phi_q(r)$ and $\rho_{rs} = \phi_r(r)\phi_s(r)$. The densities may be approximated using an auxiliary basis set as (equation 3)

$$\bar{\rho}_{pq}(r) = \sum_P^{N_{fit}} d_P^{pq}\chi_P(r)$$

where d_p^{pq} are fitting coefficients.

There are various methods for obtaining these fitting coefficients. If one minimizes the following (positive definite) functional with a $1/r_{12}$ weight factor

$$\Delta_{pq} = \int dr_1 \int dr_2 \frac{[\rho_{pq} - \bar{\rho}_{pq}(r_1)][\rho_{pq}(r_2) - \bar{\rho}_{pq}(r_2)]}{r_{12}}$$

then it has been shown that this minimizes the error in the electric field and leads to the following fitting coefficients:

$$d_Q^{pq} = \sum_P (pq|P)[[J]^{-1}]_{PQ}$$

where

$$(pq|P) = \int dr_1 \int dr_2 \phi_p(r_1)\phi_q(r_1)\frac{1}{r_{12}}\chi_P(r_2)$$

and

$$J_{PQ} = \int dr_1 \int dr_2 \chi_P(r_1)\frac{1}{r_{12}}\chi_Q(r_2).$$

Therefore the 4 index integrals is replaced with

$$(pq|rs) = \int dr_1 \int dr_2 \rho_{pq}(r_1)\frac{1}{r_{12}}\rho_{rs}(r_2)$$

$$= \int dr_1 \int dr_2 \sum_Q d_Q^{pq}\chi_Q(r_1)\frac{1}{r_{12}}\phi_r(r_2)\phi_s(r_2)$$

$$= \sum_Q d_Q^{pq}(Q|rs)$$

$$= \sum_{PQ}(pq|P)[[J]^{-1}]_{PQ}(Q|rs)$$

Although the methods look like resolution of the identity and are called RI approximations, they are not the standard RI expressions.

The number of possible choices for the fitting coefficients d_P^{pq} in equation 3 and depends on the functional form which is minimized. Instead of equation 4, which uses the coulomb operator $g(\mathbf{r}_1, \mathbf{r}_2) = 1/r_{12}$ one may use an expression similar to an overlap $g(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$, an anti-coulomb operator $g(\mathbf{r}_1, \mathbf{r}_2) = -|\mathbf{r}_1 - \mathbf{r}_2|$ this is optimal for representing the potential caused by ρ_{pq} or other choices.
one disadvantage of the coulomb metric is that the fitting coefficients decay slowly wrt the distance between $|pq > and|P >$, this is also a bad choice under periodic boundary condition. The overlap metric does not have these difficulties but is not as accurate. To overcome these difficulty there has been investigation into the attenuated Coulomb operator, $g(\mathbf{r}_1, \mathbf{r}_2) = erfc(\omega|\mathbf{r}_1 - \mathbf{r}_2|)/|\mathbf{r}_1 - \mathbf{r}_2|$ which can be tuned between the coulomb and overlap metric depending on the size of ω
The authors of these papers mention that the number of significant quantities in the sum grows quadratically with system size for any of the three metrics they considered and thus the square root formulation appears to be unsuitable for linear-scaling algorithms for large systems.

In DF, one normally uses atom-centered Gaussians as the auxiliary basis set the size of the auxiliary basis in density fitting is usually 3-4 times the size of the standard basis. A small disadvantage is that different basis set work better depending on the type of density being fit. Somewhat larger basis sets are required to fit exchange integrals and also work well to fit coulomb integrals. Such are called "JK-fit" basis sets. Unfortunately these basis sets are not optimal for HF procedures. Auxiliary basis sets have also been developed for use in conjunction with correlation consistent basis sets for density fitted MP2 and CC2 computations.

Closely related is the pseudospectral (from tensor hypercontraction) which uses a real-space grid instead of aatom centered Gaussians as the auxiliary basis set. The Cholesky decomposition when applied to electron repulsion integrals also expresses four index integrals as sums of products of three index integrals, though it avoids the use of the coulomb metric.

It is more likely that the 3 centered integrals with fit into memory than the four centered. It also takes less I/O to read three index integrals from disc than 4 index integrals. An obvious disadvantage is that, unless there is a favorable factorization, 4 index integrals must be constructed from three index integrals like

$$(pq|rs) = \sum_{PQ} (pq|P)[\mathbf{J}^{-1}]_{PQ}(Q|rs)$$

In AO $(pq| = |rs)$ and $\mathbf{J}$ is the coulomb metric
the floating point procedure is faster than I/O needed to read 4 index integrals.
to reduce the cost of constructing the 4 index integrals it can be advantageous to use a symmetric expression which splits the $\mathbf{J}^{-1}$ into a product of $\mathbf{J}^{-1/2}\mathbf{J}^{-1/2}$

$$(pq|rs) = \sum_{PQR} (pq|P)[\mathbf{J}^{-1/2}]_{PQ}[\mathbf{J}^{-1/2}]_{QR}(R|rs)$$

this factorizes into

$$(pq|rs) = \sum_{Q} b_{pq}^{Q} b_{rs}^{Q}$$

where

$$b_{pq}^{Q} = \sum_{P} (pq|P)[\mathbf{J}^{-1/2}]_{PQ}$$

2 DF-MP2

The most time consuming step of MP2 is the integral transformations

$$(ia|jb) = \sum_{\mu\nu\rho\sigma} C_\mu^i C_\nu^a C_\rho^j C_\sigma^b (\mu\nu|\rho\sigma)$$

This method has scaling N^8 but one can break up the transformation into parts and reduce the scaling to N^5. See below

$$(i\nu|\rho\sigma) = \sum_{\mu} C_{\mu}^{i}(\mu\nu|\rho\sigma)$$

$$(ia|\rho\sigma) = \sum_{\nu} C_{\nu}^{a}(i\nu|\rho\sigma)$$

$$(ia|j\sigma) = \sum_{\rho} C_{\rho}^{j}(ia|\rho\sigma)$$

$$(ia|jb) = \sum_{\sigma} C_{\sigma}^{b}(ia|j\sigma)$$

With the density fitting, one can factorize the work more effectively:

$$b_{i\nu}^{Q} = \sum_{\mu} C_{\mu}^{i} b_{\mu\nu}^{Q}$$

$$b_{ia}^{Q} = \sum_{\nu} C_{\nu}^{a} b_{i\nu}^{Q}$$

$$(ia|jb) = \sum_{Q} b_{ia}^{Q} b_{jb}^{Q}$$

The most expensive step of the DF transformation is the last one with $\mathrm{O}(N^5)$

Fast linear scaling second-order MP2 using local and DF approximations
DOI: 10.1063/1.1564816

Fast linear scaling second-order MP2 using local and DF approximations
DOI: 10.1063/1.1564816 Goals to create high-level local electron correlation method with low-order scaling as a function of molecular size. Local correlation approach linear scaling has been achieved for local second order MP2, local CCSD and local CCSD(T) and related methods, (Look to sources). Other methods are Laplace transform MP2 methods from Ayala and Scuseria, AO based LMP2 methods and pseudospectral LMP@ method. These methods have extended the applications of wave function based correlation methods to systems with 100 atoms with basis sets of double-zeta plus polarization quality.

Basis sets of triple-zeta quality are needed to obtain sufficiently accurate results. Methods require computational time and disk-space requirement of forth order of the basis set size per atom. This is the limiting factor in accurate calculations for larger systems. When cc-pVnZ basis sets are used the basis set size increases 2^n, thus an increase of the CPU time by an order of magnitude is expected when going to the larger basis set. Another problem is linear scaling is reached only for rather extended systems so saving achieved by linear scaling methods are less for compact 3D molecules than for extended systems, such as one dimensional alkane or peptide chains.

The bottleneck in large basis set calculations is the 2-electron integrals and their transformation from the AO to local orbital basis(MO?) (we see this in LMP2, LCCSD and Laplace-transform linear scaling MP2 method) An alternative to exact calculations of the 2-electron integrals is their approximation with density fitting.
Density fitting mathematically resembles a resolution of the identity $(|y) = \sum(y|X)|y))$ when fitting criterion and target integral type coincides. How it differs is RI involves a summation over states and an implied overlap metric, DF does not. DF is better though of minimizing the coulomb energy of a fitting residual. The advantages of DF are as follows: The computational cost is reduced from N_{AO}^4 to N_{AO}^3. the three index integrals have a low pre-factor for medium sized molecules. DF-MP2 still scales N^5 because it cannot be reduced wiht scaling while integral direct can be reduced to N^3.

This paper shows that this bottleneck can be reduced by local approximations. This includes the use of individual excitation subspaces (domains) for electron pairs and the use of multipole expansions for generating transformed 2-electron integrals for distant pairs. Reducing scaling order to N^2. Introducing different fitting bases for each e- pair the scaling can be reduced to O(N)

3 Theory

Summarize conventional DF-MP2 using canonical orbital basis.
Required for MP2 are 2-electron integrals $K_{ab}^{ij} = (ai|bj)$ over occupied and virtual orbitals $(\phi_i\phi_j, \phi_a\phi_b)$. They can be defined as the electrostatic repulsion between two orbital product densities

$$K_{ab}^{ij} = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \frac{\rho_{ai}(\mathbf{r}_1)\rho_{bj}\mathbf{r}_2}{r_{12}}$$

the one electron densities are $\rho_{ai}(\mathbf{r}_1) = \phi_a(\mathbf{r}_1)\phi_i(\mathbf{r}_1)$ and are approximated as

$$\hat{\rho}_{ai}(\mathbf{r}) = \sum_A^{N_{fit}} d_A^{ai}\chi_A(\mathbf{r})$$

where $\chi_A(\mathbf{r})$ are fitting basis functions. the expansion coefficients (d) are obtained by minimizing the positive definite functional

$$\Delta_{ai} = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \frac{[\rho_{ai}(\mathbf{r}_1) - \hat{\rho}_{ai}(\mathbf{r}_1)][\rho_{ai}(\mathbf{r}_2) - \hat{\rho}_{ai}(\mathbf{r}_2)]}{r_{12}}$$

this leads to

$$d_B^{ai} = \sum_A (ai|A)[\mathbf{J}^{-1}]_{AB}$$

$$\hat{K}_{ab}^{ia} = \sum_B d_B^{ai}(B|bj) = \sum_{AB} (ai|A)[\mathbf{J}^{-1}]_{AB}(B|bj)$$

Where

$$J_{AB} = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \frac{\chi_A(\mathbf{r}_1)\chi_B(\mathbf{r}_2)}{r_{12}}$$

$$(ai|A) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \frac{\phi_a(\mathbf{r}_1)\phi_i(\mathbf{r}_1)\chi_A(\mathbf{r}_1)}{r_{12}}$$

If the MOs are expanded in a basis of GROs $\{\chi_\mu\}$, the 3-index integrals in the MO basis are obtained by a two step transormation of the 3-index integrals $(\mu\nu|A)$ in the AO basis (same as in the Sherrill paper for MP2)

$$(\mu i|A) = \sum_\nu C_{\nu_1}(\mu\nu|A)$$

$$(ai|A) = \sum_\mu C_{\mu a}(\mu i|A)$$

The size of an integral $(\mu\nu|A)$ decreases exponentially with the square distance between the basis functions χ_μ and χ_ν the number of non-negligible integrals scales asymptotically as molecule $size^2$
The number of occupied (correlated??) orbitals is proportional to the molecular size so the first transformation scales as $size^3$. Since the canonical MOs are delocalized over the whole molecule, first half transformation is not sparse and thus the second half transformations scales as $size^4$ The next step is to solve for linear equations for the weighting factors d_B^{ai} since $\mathbf{J}^{-1}$ cannot be expected to be sparse this equation also scales $size^4$ the final step of forming $\hat{K}_{ab}^{ia}$ requires $N_{occ}(N_{occ}+1)N_{fit}N_{virt}^2$ equivalent to $size^5$ This final step dominates the total computational cost in calculations for large molecules.

4 local DF MP2

Occupied space is spanned by localized molecular orbitals obtained from the canonical orbitals by standard localization procedure. The virtual space is spanned by a basis of nonorthogonal projected atomic orbitals(PAO) obtained from the AO basis by projecting out the occupied orbital space, labeled r,s. The inherent locality allows for 2 approximations: excitations from a pair of occupied LMP can be restricted to subsets of PAOs that are spatially close to the two LMOs. The number of functions $N_{[ij]}$ in each of these subsets (pair domains) is independent of the molecular size, and the number of excitations for each electron pair reduces from N_{virt}^2 to $N_{[ij]}^2$. Second the integrals $(ri|sj)$ for distant orbitals i and j can be approximated by multipole expansions or neglected. the remaining number of non-distant orbital pairs (ij), and therefor the total number of excitations, scales linearly with molecular size. There is a 1-1 correspondence between # of excitations and integrals so number of integrals also scales linearly with molecular size. The range of r,s (in $(ri|sj)$ is restricted to the pair domain [ij]).
This reduces the scaling of the last step of LMP2 to order $size^2$. Since PAOs r must be in a finite range of LMOs i and j, one only requires those 3-index integral transforms $(ri|A)$ with r in the united pair domain of the associated orbital i. This domain has all PAOs that belong to any pair domain [ij] in which orbital i occurs. For large molecules the size of united pair domains is independent of molecular size and the

integrals scale $size^2$ reducing the effort to solve for d from $size^4$ to $size^3$ Finally since occupied orbitals i are local the AO to LMO transformations should scale as $size^2$, with prescreening. linear scaling can be achieved by using domains also for the fitting basis.

There are 2 cases for fitting. In the first case a different fitting basis is used for each electron pair and the linear eq solved for each pair individually. The fitting basis for a given pair is all fitting fucntions at the atom belonging to the pair domain and includes all functions at all atoms within some distance R. For large molecules the size of these pair fitting sets is independent for the molecular size, so there is linear scaling provided the number of electron pairs scales linearly. For a given orbital i the 3 body integrals are needed only for orbital fit domain $[i]_{fit}$ which is the union of all pair fit domains $[ij]_{fit}$ containing i. For a given LMO i, the number of PAOs r and of fitting functions A is asymptotically independent of molecular size, linear scaling is achieved for CPU time and disk space. The prefactor depends on the size of the fitting domains.
The second case the fit is not performed for each pair individually but only once for each orbital i, using the orbital fit domains $[i]_{fit}$ as a fitting basis. As before the size of this fitting basis is asymptotically independent of the molecular size and therefore linear scaling is achieved for the linear equations and integral assembly. The number of sets of linear equations solved is typically 15 times smaller than the first method, but the number of fitting functions in the orbital domains and number of PAOs r is larger therefore providing a higher cost for solving each set of equations. This second method is more accurate since it includes more functions in the fit for each pair and does not require domain extensions (R=0 bohr).

In order to achieve linear scaling for the integral transformation a further fitting domain is used since the in the int assembly step the orbital fit domain $A \in [i]_{fit}$is used to multiply d with the integrals $(sj|A)$ the integrals for a fixed j must be available for A belonging to the union of all $[i]_{fit}$ of orbitals i forming pairs with j. They are larger than the orbital fit domains but their size is asymptotically independent of the molecular size.

The final problem is the evaluation of the 3-index integrals $(\mu\nu|A)$. Before it is noted that for each i only a constant number of A is needed. One can use this by extimating the magnitude of the integral by the Schwarz inequality

$$(\mu\nu|A) <= (\mu\nu|\mu\nu)^{1/2}(A|A)^{1/2}$$

For the list of orbitals i, a lookup table is used to test if A is needed. If not the integral transformation is skipped reduceing the scaling to linear. Screening is performed using blocks of integrals. The integrals $(\mu\nu|A)$ for each fitting index A are split into blocks of AOs μ, ν. Typically, one block comprises all AOs at one atom.

5 tests

MOLPRO. fitting basis picked so completeness of .985, least squares risidual of .015)Largest error from the local approximation in the molecules that can be well localized and fitting error the largest for these cases as well. reason unclear. Previous work has shown that local approximations don't really affect properties like equilibrium and harmonic frequencies. though reaction enthalpies can be strongle affected.
Tested on C21H36O2 cc-pVDZ and cc-pVTZ, going form canonical to df speeds up MP2 by 5.5 times. introducing local DFMP2 reduces again by a factor of 3-4. cpu time reduced by a factor of 150 1.5 hrs to 36 seconds. The solving time is sensitive to the domain size and significant increases if the domains are extended. DF-LMP2 provides 3% error for CC-PVDz and 1.5% with the extension to R=3. In cc-pVTZ the error is only 2% and 1% Fitting errors are larger for cc-pVTZ than DZ basis