Theor Chem Acc (2009) 124:1–10
DOI 10.1007/s00214-009-0608-y

OVERVIEW

Density fitting with auxiliary basis sets from Cholesky decompositions

Thomas Bondo Pedersen · Francesco Aquilante · Roland Lindh

Received: 8 June 2009 / Accepted: 9 July 2009 / Published online: 23 July 2009

Abstract Recent progress in the use of Cholesky decomposition techniques within the density fitting approximation of two-electron integrals is reviewed with emphasis on the theoretical background. Special attention is paid to the fact that errors due to the density fitting approximation can be controlled by constructing auxiliary basis sets by means of Cholesky decomposition of either the entire or certain subblocks of the molecular two-electron integral matrix. Finally, the prospects of trivial linear-scaling calculation of fitting coefficients in the Cholesky decomposition-based density fitting scheme are outlined.

Keywords Cholesky decomposition · Density fitting · Resolution of the identity · Ab initio methods · Linear scaling

1 Introduction

Combined with the Pauli principle, the Coulomb repulsion governs the correlated dynamics of electrons in atoms, molecules, and bulk material, and it is thus responsible for almost all phenomena in the everyday world. It is, however, also responsible for grand challenges in computational quantum chemistry [1], making the parametrization of the exact electronic wave function or, equivalently, the density functional of density functional theory (DFT) complicated.

In the second quantization formulation [1], the electronic Coulomb interaction is parametrized in terms of two-electron integrals which can be structured in a matrix with elements,

$$(\mu\nu|\lambda\sigma) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \chi_\mu(\mathbf{r}_1)\chi_\nu(\mathbf{r}_1) r_{12}^{-1} \chi_\lambda(\mathbf{r}_2)\chi_\sigma(\mathbf{r}_2), \quad (1)$$

where the basis functions are assumed real and $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$. (Atomic units are used throughout.) The basic challenge is the number of integrals: with N basis functions, there are $N^4/8$ two-electron integrals. In quantum chemistry, the basis set is normally composed of functions centered at the atomic positions and contracted Gaussians are most often chosen in order to facilitate the evaluation of the two-electron integrals [1]. With atom-centered basis functions, the number of two-electron integrals formally shows quartic scaling with system size (number of atoms). Regardless of the exact choice of basis functions, however, the two-electron integral matrix is symmetric positive semidefinite and the matrix elements, hence, satisfy the inequality [2, 3]

$$|(\mu\nu|\lambda\sigma)| \leq \sqrt{(\mu\nu|\mu\nu)(\lambda\sigma|\lambda\sigma)}. \quad (2)$$

For atom-centered basis functions with limited radial extent, it then follows that the number of significant

T. B. Pedersen · R. Lindh (✉)
Department of Theoretical Chemistry, Chemical Center,
University of Lund, P.O. Box 124, 221 00 Lund, Sweden
e-mail: roland.lindh@teokem.lu.se

T. B. Pedersen
e-mail: thomas.pedersen@teokem.lu.se

F. Aquilante
Department of Physical Chemistry, Sciences II,
University of Geneva, Quai E. Ansermet 30,
1211 Geneva 4, Switzerland
e-mail: francesco.aquilante@unige.ch

Present Address:
T. B. Pedersen
Center for Theoretical and Computational Chemistry,
Department of Chemistry, University of Oslo,
P.O. Box 1033, Blindern, 0315 Oslo, Norway

integrals should grow as the square of the number of atoms [4]. The recently developed multipole-based integral estimates by Ochsenfeld et al. [5–7] do not change the fundamental scaling behavior but offer significant improvements for identifying small integrals. Note, however, that the onset of quadratic scaling depends on the radial extent of the basis functions. As the basis set is increased toward the basis set limit, the radial extent increases (more diffuse basis functions) and the scaling will then be quadratic only for very large systems.

There have been several attempts to simplify the handling of two-electron integrals. The most widespread approach is the density fitting (DF) or resolution-of-the-identity (RI) approximation [2, 8–13] in which the product densities $|\mu\nu)$ are expanded in an auxiliary basis set. A straightforward integral approximation is then given by

$$(\mu\nu|\lambda\sigma) \approx \sum_{PQ} C^{P}_{\mu\nu}(P|Q)C^{Q}_{\lambda\sigma}, \tag{3}$$

where P, Q are indices of the auxiliary basis set and the expansion coefficients $C^{P}_{\mu\nu}$ are determined by least-squares fitting. Dunlap [14–17] has convincingly argued in favor of the robust integral representation,

$$\begin{aligned}(\mu\nu|\lambda\sigma) \approx & \sum_{P}(\mu\nu|P)C^{P}_{\lambda\sigma} + \sum_{P} C^{P}_{\mu\nu}(P|\lambda\sigma)\\ & - \sum_{PQ} C^{P}_{\mu\nu}(P|Q)C^{Q}_{\lambda\sigma},\end{aligned} \tag{4}$$

which is equivalent to Eq. 3 when certain conditions are satisfied, as we shall discuss in more detail below. The auxiliary basis functions are normally chosen as atom-centered functions, and the calculation of the expansion coefficients thus involves only three-center integrals. Thus, the immediate advantage of the DF approach is to avoid calculation of four-center integrals, but the scaling with system size is cubic (albeit with a very small prefactor) due to the matrix inversion needed to calculate the expansion coefficients. Local DF schemes are currently being developed for linear-scaling calculation of Coulomb and exchange contributions to the Fock matrix, see e.g. [18, 19] and references therein, and for linear-scaling calculation of local second-order Møller-Plesset (MP2) theory for molecules [20] as well as periodic (insulating) systems [21]. The accuracy of the DF approximation depends on the choice of auxiliary basis set. The standard approach is to optimize an auxiliary basis set for each atomic orbital (AO) basis set and quantum chemical contribution, such as Coulomb [22–24] and exchange [25, 26] contributions to the Fock matrix (including Kohn-Sham DFT), and MP2 and second-order coupled cluster (CC2) models [27–31].

A special case of LU factorization, Cholesky decomposition (CD) is defined for any symmetric positive semidefinite matrix [3, 32]. As the two-electron integral matrix is symmetric positive semidefinite, Beebe and Linderberg suggested the Cholesky integral representation [33–37]

$$(\mu\nu|\lambda\sigma) \approx \sum_{J=1}^{M} L^{J}_{\mu\nu}L^{J}_{\lambda\sigma}, \tag{5}$$

which is formally exact, but the number of Cholesky vectors M is often limited by an accuracy parameter, the decomposition threshold. The CD works by removing linear dependence (in the Coulomb metric) among the product densities $|\mu\nu)$ leaving M products for a given decomposition threshold. The linear dependence increases with the quality of the basis set, making CD a possible solution to some of the challenges faced by conventional reduced scaling techniques when diffuse functions are included in the basis set [38]. Unlike DF, however, CD typically requires calculation of four-center integrals and observed computational savings stem from the fact that only a fraction of the integral matrix must be calculated to generate the Cholesky representation [37]. As we will discuss here, CD and DF are completely equivalent when a specific auxiliary basis is chosen, opening new opportunities for reduced scaling quantum chemistry.

Gill et al. [39, 40] have recently proposed the “resolution of the Coulomb operator”,

$$r_{12}^{-1} = \sum_{n}^{\infty} \phi_n(\mathbf{r}_1)\phi_n(\mathbf{r}_2), \tag{6}$$

where the functions on the right-hand side form a complete set, which is orthonormal in the Coulomb metric. Inserting Eq. 6 into Eq. 1, we see that the two-electron integrals can be expressed as sums of products of overlap integrals in a form that resembles the Cholesky representation, Eq. 5. The main challenge in this approach is to truncate the infinite sum in a way that makes it computationally feasible without sacrificing accuracy. Preliminary tests are promising [40].

We have used the Cholesky integral representation, Eq. 5, for quadratic scaling calculation of Coulomb and exchange Fock matrices [41], quartic scaling evaluation of canonical scaled opposite spin MP2 [42], complete active space self-consistent field (CASSCF) wave functions [43], multiconfigurational second-order perturbation theory (CASPT2) [44], and coupled cluster methods [45, 46]. These developments have had a particularly positive impact on the size of systems that can be treated with multiconfigurational methods, even if the limitations on the size of the active space remain unchanged. For example, Pierloot et al. [47] have studied relative energies of spin states in large ferrous complexes and binding energies to heme [48] using the CD-based CASSCF/CASPT2 implementations in the MOLCAS quantum chemistry package [46, 49–51]. For the coupled cluster methods in MOLCAS, the main impact has been that larger basis sets can be employed due to decreased disk space demands [46, 52–54].

In this article we discuss the theoretical foundation of the CD-based development in MOLCAS. In particular, we will focus on the properties of auxiliary basis sets for the DF approximation generated by means of CD. As CD-based auxiliary basis sets are constructed with no other knowledge than the AO basis set (and molecular stoichiometry and possibly geometry), we refer to this class of DF approximations as "ab initio density fitting". We proceed as follows. The DF approximation, including a brief review of the construction of the most popular predefined auxiliary basis sets, is described in Sect. 2 along with an analysis of DF errors. Motivated by the DF error analysis, CD is introduced in Sect. 3 as a procedure for generating auxiliary basis sets on-the-fly. In Sect. 4 we discuss the inherent locality of CD-based fitting coefficients and its implications for local DF. Finally, concluding remarks are given in Sect. 5.

2 Density fitting

2.1 Fundamentals of density fitting

The DF approximation is based on an expansion of product densities in an auxiliary basis,

$$|\mu\nu) \approx |\widetilde{\mu\nu}) = \sum_P C^P_{\mu\nu}|P), \tag{7}$$

leading to the two-electron representation of Eq. 3 by straightforward substitution in Eq. 1. The expansion coefficients can be determined by minimizing the norm

$$d_{\mu\nu} = (\Delta_{\mu\nu}|\Delta_{\mu\nu})_g, \tag{8}$$

of the fitting error

$$|\Delta_{\mu\nu}) = |\mu\nu) - |\widetilde{\mu\nu}), \tag{9}$$

in a Hermitian positive definite metric $\hat{g}(\mathbf{r}_1, \mathbf{r}_2)$. The subscript g thus indicates the inner product

$$(f|h)_g = \int d\mathbf{r}_1 \int d\mathbf{r}_2\, f(\mathbf{r}_1)\hat{g}(\mathbf{r}_1, \mathbf{r}_2)h(\mathbf{r}_2). \tag{10}$$

Minimization of $d_{\mu\nu}$ leads to the linear fitting equations $(\Delta_{\mu\nu}|P)_g = 0$ or, equivalently,

$$\sum_Q C^Q_{\mu\nu}(Q|P)_g = (\mu\nu|P)_g. \tag{11}$$

Dunlap [14–17] has pointed out that this least-squares fitting procedure does not necessarily conserve the variational principle. Using the fitting equations as constraints, however, Lagrangian multipliers can be introduced to ensure variational integrals [19]. While the Lagrangian multiplier terms do not contribute to the value of the approximated integral itself, they are important for derivatives [19]. As we do not discuss integral derivatives in this work, the Lagrangian multiplier terms will not be considered any further here (they can be added in all relevant equations without affecting the conclusions).

Auxiliary basis sets are customarily composed of atom-centered functions. It is then clear from Eqs. 3 and 11 that at most three-center integrals need to be evaluated. In fact, a major computational advantage of the DF approximation is that expensive four-center integrals are eliminated. This elimination is not free, however, as the matrix inversion required to solve the DF equations (11) shows a formal cubic scaling with system size. Fortunately, the prefactor is small and large computational savings (often an order of magnitude or more) are observed for molecules of up to about a few hundred atoms. For larger systems, linear scaling techniques based on local fitting must be employed, see e.g. [18, 19, 55].

2.2 Predefined auxiliary basis sets

Successful application of the DF approximation clearly depends on the choice of auxiliary basis set. For Gaussian one-electron basis sets it is natural to use Gaussians also for the auxiliary basis set. Auxiliary basis function parameters (exponents, contraction patterns, and contraction coefficients) must be chosen such that the DF approximation becomes both accurate and efficient. The most commonly used Gaussian auxiliary basis sets seem to be those originating from Ahlrichs' quantum chemistry group in Karslruhe, who defined the balance between accuracy and efficiency as [22–30]

1. The error due to the DF approximation must be at least one order of magnitude smaller than the error arising from one-electron basis set incompleteness, and
2. The number of auxiliary basis functions must be a few (approximately 2–4) times larger than the number of AO basis functions.

In order to strike this balance, auxiliary basis sets are generally optimized for each quantum chemical model (to satisfy requirement 1) and each atom and AO basis set (to satisfy requirement 2).

Using ground state energies to quantify "the error due to the DF approximation", auxiliary basis sets for non-hybrid DFT with the split-valence SVP [56] and triple-ζ TZVP [57] one-electron basis sets were published by Eichkorn et al. [22, 23]. Test calculations on a number of chemical compounds have revealed that these so-called RI–J auxiliary basis sets may also be used in conjunction with other (smaller as well as larger) one-electron basis sets than SVP and TZVP [58]. The RI–J auxiliary basis sets have recently been improved by Weigend [24] who developed auxiliary basis sets for all atoms from hydrogen to radon. The improved sets are universal in the sense that they are sufficiently flexible to be used for non-hybrid DFT with all one-electron basis sets from split-valence to quadruple-ζ quality. For heavier elements these sets are intended for use with effective core potentials only [24] and lanthanides and actinides are not supported.

Weigend has also developed the RI–JK auxiliary basis sets for HF and hybrid DFT methods [25, 26]. The RI–JK sets are constructed to minimize the error in both Coulomb and exchange ground state energies due to the DF approximation. At the same time, the resulting MOs and MO energies are sufficiently accurate to allow MP2 calculations with the DF–HF wave function as 0th order approximation.

Separate auxiliary basis sets are needed for the correlated calculation, however, and the RI–C auxiliary basis sets have been constructed for a large number of one-electron basis sets [27–31]. The RI–C sets are constructed to minimize the error in MP2 ground state energies (although only the quadratic part of the error function is used for minimization), but they have also been used for CC2 calculations of excitation energies [59] and for local coupled cluster calculations [60].

The RI–X (X = J, JK, C) sets are optimized at atomic (or atomic ion) calculations with a few exceptions, where diatomic hydrides are used. The accuracy is subsequently evaluated using molecular test sets covering the majority of atoms in their most common oxidation states, leading to statistical knowledge of the typical ground state energy error per atom (see, e.g., [24, 26]).

2.3 Density fitting error analysis

When approximating the two-electron integrals it is important to realize that one is ultimately modifying the Hamiltonian operator and, hence, its eigenstates and eigenvalues. All molecular properties amenable to theoretical computation are thus affected by the integral approximation, not just ground state energies. The most general way to measure the error due to the DF approximation, therefore, is to study the integral representation error.

The error of the DF integral representation, Eq. 3, is given by

$$\begin{aligned} D_{\mu\nu,\lambda\sigma} &= (\mu\nu|\lambda\sigma) - (\widetilde{\mu\nu}|\widetilde{\lambda\sigma}) \\ &= (\Delta_{\mu\nu}|\Delta_{\lambda\sigma}) + (\Delta_{\mu\nu}|\widetilde{\lambda\sigma}) + (\widetilde{\mu\nu}|\Delta_{\lambda\sigma}), \end{aligned} \tag{12}$$

which, of course, vanishes when $|\Delta_{\mu\nu}) = 0$ for all $\mu\nu$. This requires that the auxiliary basis spans the same space as the original product densities. Applying Eq. 2 to each of the three terms in Eq. 12 we obtain an upper bound to the absolute error,

$$\begin{aligned} \left|D_{\mu\nu,\lambda\sigma}\right| \leq{}& \sqrt{(\Delta_{\mu\nu}|\Delta_{\mu\nu})(\Delta_{\lambda\sigma}|\Delta_{\lambda\sigma})} + \sqrt{(\Delta_{\mu\nu}|\Delta_{\mu\nu})(\widetilde{\lambda\sigma}|\widetilde{\lambda\sigma})} \\ &+ \sqrt{(\widetilde{\mu\nu}|\widetilde{\mu\nu})(\Delta_{\lambda\sigma}|\Delta_{\lambda\sigma})}, \end{aligned} \tag{13}$$

indicating that the most accurate fit for a given auxiliary basis is obtained by minimizing the self-energy of the fitting error, $(\Delta_{\mu\nu}|\Delta_{\mu\nu})$, for all $\mu\nu$. From Eq. 8 we thus find that minimum integral representation error is ensured by choosing the Coulomb metric, $\hat{g}(\mathbf{r}_1, \mathbf{r}_2) = r_{12}^{-1}$, as has also been firmly established numerically [13, 19, 61]. If, in addition to using the Coulomb metric, all product densities are expanded in the same set of auxiliary functions, it follows from Eq. 11 that $(\Delta_{\mu\nu}|\widetilde{\lambda\sigma}) = 0$ and the integral representation error becomes quadratic in the fitting error $|\Delta_{\mu\nu})$:

$$\left|D_{\mu\nu,\lambda\sigma}\right| = \left|(\Delta_{\mu\nu}|\Delta_{\lambda\sigma})\right| \leq \sqrt{d_{\mu\nu} d_{\lambda\sigma}}. \tag{14}$$

The actual value of $d_{\mu\nu}$ depends, of course, on the details of the auxiliary basis set. Examples of the variation of $d_{\mu\nu}$ with different auxiliary basis sets are given in [62] and [63].

If a non-Coulomb metric is used or if the product densities are not expanded in the same set of auxiliary functions (e.g., using local fitting domains), $(\Delta_{\mu\nu}|\widetilde{\lambda\sigma}) \neq 0$ and the integral representation error becomes linear in the fitting error $|\Delta_{\mu\nu})$ according to Eq. 12 with the upper bound of Eq. 13. If, instead of Eq. 3, we use the robust integral representation advocated by Dunlap [14–17],

$$\begin{aligned} (\mu\nu|\lambda\sigma) &\approx (\widetilde{\mu\nu}|\widetilde{\lambda\sigma}) + (\Delta_{\mu\nu}|\widetilde{\lambda\sigma}) + (\widetilde{\mu\nu}|\Delta_{\lambda\sigma}) \\ &= (\mu\nu|\widetilde{\lambda\sigma}) + (\widetilde{\mu\nu}|\lambda\sigma) - (\widetilde{\mu\nu}|\widetilde{\lambda\sigma}), \end{aligned} \tag{15}$$

the representation error becomes quadratic in the fitting error $|\Delta_{\lambda\sigma})$, as is easily seen by rearranging Eq. 12. As the self-energy $(\Delta_{\mu\nu}|\Delta_{\mu\nu})$ is not directly minimized when a non-Coulomb metric is used, the integral representation error may be larger than that obtained using the Coulomb metric.

3 Auxiliary basis sets from Cholesky decompositions

As is evident from the analysis of Sect. 2.3, the accuracy of the DF approximation depends on the ability of the auxiliary basis to span the same space as the product densities $|\mu\nu)$. Observing that the product densities become more and more linearly dependent as the one-electron basis set approaches the basis set limit, an auxiliary basis spanning the same space can be generated by simply removing linearly dependent products. Cholesky decomposition of the two-electron integral matrix offers a numerically stable procedure to do just that.

3.1 The Cholesky basis

The accuracy of the Cholesky integral representation, Eq. 5, is measured by means of the residual matrix,

$$D^{(J)}_{\mu\nu,\lambda\sigma} = (\mu\nu|\lambda\sigma) - \sum_{K=1}^{J} L^K_{\mu\nu} L^K_{\lambda\sigma}, \tag{16}$$

such that

$$(\mu\nu|\lambda\sigma) = \sum_{J=1}^{M} L^J_{\mu\nu} L^J_{\lambda\sigma} + D^{(M)}_{\mu\nu,\lambda\sigma}. \tag{17}$$

Being a positive semidefinite matrix [33], the residual satisfies an inequality analogous to Eq. 2,

$$\left| D^{(J)}_{\mu\nu,\lambda\sigma} \right| \leq \sqrt{D^{(J)}_{\mu\nu,\mu\nu} D^{(J)}_{\lambda\sigma,\lambda\sigma}}, \tag{18}$$

and, consequently, the accuracy of the Cholesky representation of the two-electron integrals can be controlled by requiring

$$\max_{\mu\nu} \left(D^{(M)}_{\mu\nu,\mu\nu} \right) \leq \delta, \tag{19}$$

where the decomposition threshold is a non-negative real number, $\delta \geq 0$. All integrals are thus reproduced with an accuracy of at least δ, and the Cholesky representation can be employed in conjunction with any quantum chemical method and becomes exact when $\delta = 0$, as all integrals are represented exactly in this case. The choice $\delta = 10^{-4}$ roughly corresponds to the accuracy obtained using DF with predefined auxiliary basis sets [64].

The CD is a recursive procedure in which the Cholesky vector at recursion J is calculated from the residual matrix of recursion $J-1$ according to [33, 37, 46]

$$L^J_{\mu\nu} = \left[D^{(J-1)}_{[\lambda\sigma]_J,[\lambda\sigma]_J} \right]^{-1/2} D^{(J-1)}_{\mu\nu,[\lambda\sigma]_J}. \tag{20}$$

Here, $[\lambda\sigma]_J$ is the index of the largest residual diagonal element at the $(J-1)$th recursion, i.e., the product density that gives rise to the Jth vector. The stop criterion is given by Eq. 19, i.e., the recursion stops when the largest residual diagonal element is smaller than or equal to δ, thus determining the total number of Cholesky vectors M. The CD, thus, identifies a linearly independent subset $\{|h_J)\} = \{|[\lambda\sigma]_J)\}$ of the product density set $\{|\mu\nu)\}$. We refer to this subset as the Cholesky basis.

Exploiting Eq. 18 for prescreening at each recursion, a general integral-direct implementation of CD was presented by Koch et al. [37]. A slightly modified algorithm has been implemented in the MOLCAS program as described in [46], and Røeggen and Johansen [65] have recently presented a parallel implementation based on family-type basis sets. Only a fraction of the integral matrix (M columns) needs to be computed in the recursive procedure to represent the integrals with finite accuracy. The bottleneck of the integral-direct algorithm is in most cases the calculation of the residual matrix, Eq. 16. As only M columns of the residual matrix are needed to generate the Cholesky vectors, Eq. 16 shows a formal complexity of $N_d M^2$, where N_d is the number of significant product densities as determined from Eq. 2. Increasing accuracy by decreasing the decomposition threshold is, thus, accompanied by a computational penalty scaling quadratically with the increase in the number of Cholesky vectors. Typically, M is 3–10 times the number of one-electron basis functions.

The significance of the Cholesky basis can be clarified by Gram-Schmidt orthonormalization in the Coulomb metric, i.e.

$$|Q_J) = \mathcal{N}_J^{-1/2} \left[|h_J) - \sum_{K=1}^{J-1} |Q_K)(Q_K|h_J) \right], \tag{21}$$

where the normalization constant is given by

$$\mathcal{N}_J = (h_J|h_J) - \sum_{K=1}^{J-1} (h_J|Q_K)^2, \tag{22}$$

such that $(Q_J|Q_K) = \delta_{JK}$. Noting that

$$\begin{aligned} (\mu\nu|Q_1) &= (h_1|h_1)^{-1/2}(\mu\nu|h_1) \\ &= L^1_{\mu\nu}, \end{aligned} \tag{23}$$

it is easy to verify by recursion that the components of the orthonormalized functions in the original product basis are equal to the corresponding Cholesky vector components. That is,

$$(\mu\nu|Q_J) = L^J_{\mu\nu}, \tag{24}$$

and CD is, therefore, equivalent to a Gram-Schmidt orthonormalization of the product density set $\{|\mu\nu)\}$ in the Coulomb metric. As the normalization constant, Eq. 22, is simply a diagonal element of the residual matrix, Eq. 16, the Cholesky threshold clearly plays the role of tolerance for linear dependence. The Cholesky representation of the two-electron integrals can now be written as

$$(\mu\nu|\lambda\sigma) \approx \sum_{J=1}^{M} (\mu\nu|Q_J)(Q_J|\lambda\sigma), \tag{25}$$

which is equivalent to the inner projection formulation of Beebe and Linderberg [33] using an orthonormal basis.

Using the orthonormalized Cholesky basis as auxiliary basis for the DF approximation, the solution of the fitting equations (11) with the Coulomb metric becomes

$$C^J_{\mu\nu} = L^J_{\mu\nu}, \tag{26}$$

by virtue of Eq. 24. The CD and DF approaches are, therefore, entirely equivalent when the Cholesky basis is used as auxiliary basis. Employing the nonorthonormal Cholesky basis, we may rewrite the Cholesky representation of the integrals in the DF form

$$(\mu\nu|\lambda\sigma) = \sum_{JK} C^J_{\mu\nu}(h_J|h_K)C^K_{\lambda\sigma} + D_{\mu\nu,\lambda\sigma}, \tag{27}$$

where

$$|D_{\mu\nu,\lambda\sigma}| \leq \delta, \tag{28}$$

is inherited from the CD procedure. The fitting coefficients in the nonorthonormal Cholesky basis are related to the Cholesky vectors according to

$$\sum_K C^K_{\mu\nu} L^J_K = L^J_{\mu\nu}. \tag{29}$$

While the Cholesky basis is formally exact (when $\delta = 0$), it generally contains both one- and two-center functions. Both the fitting Eqs. 11 and 27, therefore, contain some four-center integrals and a major computational advantage of DF is partially lost. It is, therefore, important to explore approximate Cholesky bases containing one-center functions only.

The Cholesky basis depends in a nontrivial way on molecular geometry. Besides the explicit dependence on atomic positions of the Cholesky basis product functions, molecular geometry dictates which product functions are selected by the CD procedure to enter the Cholesky basis. Although one might expect this feature to give rise to discontinuities in potential energy surfaces, numerical tests have not revealed such issues [66].

3.2 One-center Cholesky decomposition

The one-center CD (1C-CD) approximation [62] is a straightforward way to avoid two-center functions: simply deny two-center functions entry into the Cholesky basis during the recursive procedure defined by Eq. 20. The recursion stops when the largest one-center diagonal residual element is below δ regardless of the size of the two-center diagonal residual elements. While one-center fitting errors are bounded by the decomposition threshold, $d_{\mu_A\nu_A} \leq \delta$, the two-center errors $d_{\mu_A\nu_B}$ are not bounded but rather minimized to the extent that they can be expanded in the one-center Cholesky basis. Consequently, only one- and two-center integrals of the types $(\mu_A\nu_A|\lambda_A\sigma_A)$ and $(\mu_A\nu_A|\lambda_B\sigma_B)$ are represented with an accuracy of at least δ according to Eq. 14. The errors in the remaining integrals approach a constant value as the decomposition threshold is reduced, but they will not generally vanish in the limit $\delta \to 0$.

Like the full Cholesky basis, the 1C-CD basis depends on geometry, although this does not seem to lead to discontinuity issues in practice [66]. In fact, we have utilized Eq. 27 in conjunction with the 1C-CD approximation to implement analytic gradients and shown that the accuracy of bond lengths and angles can be controlled by adjusting the decomposition threshold [66].

3.3 Atomic Cholesky decomposition

The geometry-dependence of the 1C-CD basis can easily be removed by generating a Cholesky basis for each atom. This idea is closely related to Ten-no and Iwata's [67, 68] RI with linear combination of atomic electron distributions (RI-LCAD) method, in which the atomic integral matrix is diagonalized and the eigenfunctions corresponding to eigenvalues above a given threshold are used as auxiliary basis functions. In the atomic CD (aCD) approach [62] each unique atomic block, $(\mu_A\nu_A|\lambda_A\sigma_A)$, of the integral matrix is decomposed to give an atomic Cholesky basis, which contains exactly those product densities that are linearly independent within the decomposition threshold. The atomic Cholesky basis, therefore, does not necessarily contain all angular components and the aCD auxiliary basis set is obtained by completing the shell structure [62, 63] (i.e., by adding "missing" angular components). Constructed as products of AOs, the aCD set may contain an exceedingly large number of primitive Gaussian functions, making the calculation of two- and three-center integrals expensive. In order to reduce the number of primitives, we have developed the atomic compact CD (acCD) auxiliary basis set [63]. The acCD set is obtained from the aCD by decomposing an "angular free" two-electron integral matrix to remove linear dependence among the primitive Gaussians, followed by a fit of the contraction coefficients, as described in more detail in [63]. Extensive tests have shown that the differences between total energies calculated with the aCD and acCD sets are one to two orders of magnitude smaller than the error relative to conventional calculations [63, 64].

In principle, the aCD and acCD sets can be generated once for each atom in the periodic table (and each one-electron basis set and decomposition threshold) and the results stored in an auxiliary basis set library. This approach is not worthwhile, however, as the computational overhead of generating the aCD sets on-the-fly is minute. In addition, the atomic natural orbital (ANO) basis sets of Widmark et al. [69, 70] allow tailoring of the contraction scheme to the problem at hand. An aCD and acCD set for each possible contraction scheme would therefore have to be stored in the auxiliary basis set library, making the on-the-fly approach preferable.

As for 1C-CD, only one- and two-center integrals of the types $(\mu_A\nu_A|\lambda_A\sigma_A)$ and $(\mu_A\nu_A|\lambda_B\sigma_B)$ are represented with an accuracy of at least δ with the aCD and acCD sets according to Eq. 14.

4 Local density fitting

The cubic scaling with system size of the DF approach is due to the expansion of all product densities in the full set of auxiliary functions according to Eq. 7. The dimension of the fitting equations (11) thus grows quadratically with system size. Two approaches have been used to tackle this problem. In partitioning schemes [55, 71–74], the expansion is limited to auxiliary functions in the vicinity of the product density. The dimension of the fitting equations for a given product density, thus, becomes independent of system size. This can also be achieved by exploiting the sparsity of the $(Q|P)_g$ and $(\mu\nu|P)_g$ matrices arising through the use of a local metric g [19, 61]. The drawback of both types of approaches is that the accuracy is reduced compared to full expansions with the Coulomb metric.

As discussed in Sect. 2.3, Dunlap's robust integral representation ensures that the error becomes quadratic in the fitting error and thus bounded from above according to

$$\begin{aligned}\left|D_{\mu_A\nu_B,\lambda_C\sigma_D}\right| &= \left|(\Delta_{\mu_A\nu_B}|\Delta_{\lambda_C\sigma_D})\right| \\ &\leq \sqrt{(\Delta_{\mu_A\nu_B}|\Delta_{\mu_A\nu_B})(\Delta_{\lambda_C\sigma_D}|\Delta_{\lambda_C\sigma_D})},\end{aligned} \tag{30}$$

regardless of the chosen partitioning and metric. Evidently, one should use the Coulomb metric in order to obtain the smallest possible error. The question then becomes if all auxiliary functions are needed to accurately represent a given product density.

The partitioning scheme employed for charge densities by Fonseca Guerra et al. [72] is defined such that only auxiliary functions on A and B are used for expanding the product densities $|\mu_A\nu_B)$. This approach is justified by the intuitively reasonable statement that if the auxiliary basis functions on centers A and B are sufficiently flexible for calculations on the diatomic molecule AB, they should be sufficient for expanding $|\mu_A\nu_B)$ in polyatomic molecules as well. By the same token, the accuracy of the expansion becomes independent of chemical environment, i.e., the fitting error for a given product density becomes transferable from one system to another.

In this approach, product densities are expanded according to

$$|\widetilde{\mu_A\nu_B}) = \sum_{P_{AB}} C^{P_{AB}}_{\mu_A\nu_B}|P_{AB}), \tag{31}$$

where P_{AB} denotes auxiliary functions on atoms A and B ($A \geq B$). Defining a Cholesky basis for each interacting atom pair by decomposing the corresponding diagonal block $(\mu_A\nu_B|\lambda_A\sigma_B)$ of the integral matrix, we have

$$(\Delta_{\mu_A\nu_B}|\Delta_{\mu_A\nu_B}) \leq \delta, \tag{32}$$

and hence from Eq. 30,

$$\left|D_{\mu_A\nu_B,\lambda_C\sigma_D}\right| \leq \delta. \tag{33}$$

Combining the partitioning scheme of Fonseca Guerra et al. [72] with the concept of the Cholesky basis, we thus obtain a local DF (LDF) approach with complete error control through the decomposition threshold and linear-scaling calculation of fitting coefficients with the Coulomb metric.

One drawback is that once the fitting coefficients have been calculated, evaluation of integrals according to Eq. 15 requires calculation of four-center integrals. This can be avoided by using the aCD or acCD classes of auxiliary basis sets such that

$$|\widetilde{\mu_A\nu_A}) = \sum_{P_A} C^{P_A}_{\mu_A\nu_A}|P_A), \tag{34}$$

$$|\widetilde{\mu_A\nu_B}) = \sum_{X=A,B}\sum_{P_X} C^{P_X}_{\mu_A\nu_B}|P_X). \tag{35}$$

Note, however, that we are trading accuracy for efficiency since Eq. 32 now only holds for $A = B$.

The aCD sets are tailored for accurate fitting of one-center product densities. It is therefore reasonable to expect that even if a given one-center product density is expanded in the full aCD set of functions, there will be only significant contributions from auxiliary functions on the same center. This is indeed what we observed in recent numerical tests [63]. The "inherent" locality depends, of course, on the decomposition threshold defining the aCD set, but locality is achieved even with the Coulomb metric [63]. For low enough threshold, therefore, Eq. 34 is practically equivalent to the full expansion. Moreover, the prescreening techniques used, e.g., by Reine et al. [19] would produce the local expansion of Eq. 34.

For two-center product densities, on the other hand, the aCD sets cannot be expected to be inherently local. To demonstrate this, we have performed a series of calculations on the N_2—benzene van der Waals complex using a development version of MOLCAS [46, 49–51] and the AO basis set ANO-RCC-VTZP of Roos et al. [75]. Fitting coefficients for the two-center product densities on the nitrogen atoms are calculated for full expansions in the aCD-n, $n = -\log\delta = 4,6,8,10,12$, series of auxiliary basis sets. In Fig. 1 we have plotted the largest "contaminant",

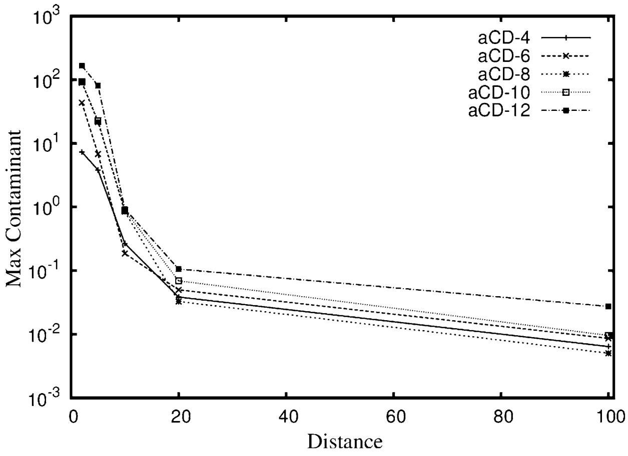

Fig. 1 Contaminants in fitting two-center product densities on the nitrogen atoms of the N_2—benzene van der Waals complex as a function of the distance (in Å) between the centers of mass of the moieties

i.e. fitting coefficient corresponding to auxiliary functions centered on the benzene ring, as a function of the distance between the centers of mass of N_2 and benzene. The N_2 molecule is oriented parallel to an axis through two opposite carbon atoms of the benzene ring. The slow decay behavior of the contaminant is in agreement with the results of Jung et al. [61] and confirms that the two-center fitting coefficients obtained with the aCD-*n* sets are not inherently local. In contrast to the behavior observed for one-center product densities [63], we find that lowering the decomposition threshold does not improve locality. It should be noted, however, that this does not imply that the local *ansatz* of Eq. 35 is too inaccurate to be useful. It does show that accuracy will improve if the fitting domain is enlarged, although there is no obvious way to choose the domain size by means of a distance cut-off. In cases where it is needed, control of the error in integrals involving two-center product densities can be regained by augmenting the atom-centered auxiliary basis functions with two-center functions obtained by decomposing the corresponding diagonal block of the residual matrix. This will lead to fewer two-center auxiliary functions, and hence fewer four-center integrals, than the "brute-force" Cholesky basis used in Eq. 31. Only practical experience can tell if this augmentation is necessary. Work along these lines is currently in progress and our results will be reported in the near future.

5 Concluding remarks

We have reviewed the theoretical foundation for using CD as a generator of auxiliary basis sets for the DF approximation of two-electron integrals. Starting from an explicit demonstration of the equivalence of CD and DF when the Cholesky basis is used as auxiliary basis set, we have reviewed three types of CD-based techniques: 1C-CD is based on the molecular integral matrix, while aCD and acCD are based on the atomic integral matrix. The main features of these CD-based approaches can be summarized as:

- Full CD
    - Decomposition of the molecular integral matrix.
    - Cholesky vectors are computed according to Eq. 20.
    - Subsequent calculations may use either the CD form, Eq. 5, or the DF form, Eq. 3, with fitting coefficients determined by Eq. 29.
    - All integrals are represented with an accuracy of at least δ.
    - Depends on molecular geometry.
- 1C-CD
    - Decomposition of the molecular integral matrix.
    - Cholesky vectors are computed according to Eq. 20.
    - Only residual diagonal elements corresponding to one-center product densities are allowed to produce Cholesky vectors.
    - Subsequent calculations may use either the CD form, Eq. 5, or the DF form, Eq. 3, with fitting coefficients determined by Eq. 29.
    - Integrals that solely involve one-center product densities are represented with an accuracy of at least δ. All other integrals may have larger errors.
    - Depends on molecular geometry.
- aCD
    - Decomposition of the atomic integral matrix produces a Cholesky basis for each unique atom and AO basis set.
    - The aCD auxiliary basis set is obtained by adding missing angular components to the Cholesky basis.
    - Subsequent calculations use the DF form, Eq. 3, with fitting coefficients determined by Eq. 11.
    - Integrals that solely involve one-center product densities are represented with an accuracy of at least δ. All other integrals may have larger errors.
    - Independent of molecular geometry.
- acCD
    - The acCD auxiliary basis set is obtained from the aCD set by removing linear dependence in the primitive product space (by means of CD).
    - Subsequent calculations use the DF form, Eq. 3, with fitting coefficients determined by Eq. 11.

- Integrals that solely involve one-center product densities are represented with an accuracy of approximately δ or better. All other integrals may have larger errors.
- Independent of molecular geometry.

In practice, full CD and 1C-CD have been implemented using Eq. 5 in place of the integrals, whereas aCD and acCD have been implemented using Eq. 3 in a form that resembles the CD expression. The details of our implementation in the MOLCAS software can be found in [46] and references therein.

Using CD to generate auxiliary basis sets for the DF approach offers a number of advantages, including:

- automatic generation of the auxiliary basis set,
- control of the error in individual two-electron integrals, and
- inherent locality.

The DF approach is first and foremost a technique for speeding up quantum chemical calculations. From the viewpoint of a user of quantum chemical software, it is clearly advantageous to have the auxiliary basis set generated automatically with as little user input as possible. The only input needed to generate CD-based auxiliary basis sets is the decomposition threshold for which a suitable default value can be established on the basis of benchmark calculations [46].

The specific type of Cholesky decomposition (full CD, 1C-CD, aCD, and acCD) along with the decomposition threshold control the maximum error in individual two-electron integrals. This means that the error of the Hamiltonian operator is controllable and that the CD-based auxiliary basis sets are unbiased in the sense that they can be used in conjunction with any quantum chemical method and any AO basis set for the calculation of ground and excited states [37, 38, 41–44, 46, 63, 64, 66, 76] as well as static and frequency-dependent electromagnetic properties [45, 77]. The error in computed quantities, such as total energies become controllable through the decomposition threshold [64].

The unbiased nature and controllable error of the CD-based auxiliary basis sets are also the source of the main disadvantage, namely that they are more computationally demanding than predefined auxiliary basis set [64, 78]. We find that this is a reasonable price to pay for an integral representation with uniform error and inherent locality. It should be stressed, however, that method-specific CD has recently been proposed as a procedure for generating compact auxiliary basis sets tailored for calculating individual contributions, such as Coulomb and exchange terms [38].

The inherent locality of CD-based auxiliary basis sets is perhaps the greatest advantage, as it allows trivial linear-scaling calculation of the fitting coefficients. While maintaining complete error control, linear-scaling calculation of the inactive and active Fock matrices needed for multiconfigurational wave functions become possible. This paves the way for truly large-scale applications of multiconfigurational methods, making it possible to address chemical problems without having to resort to small model systems. It must be added, of course, that the CD-based techniques do not affect the current limitations on the size of the active orbital space. This emerging technology is, therefore, mostly aimed at systems, such as transition metal complexes with large ligands, where only a small active space (orbitals around the central transition metal atom) is needed. Our progress in this direction will be reported in the near future.

Acknowledgments Funding from the Swedish Research Council (VR) and the Swiss National Science Foundation (SNF; Grant No. 200020-120007) is gratefully acknowledged.

References

1. Helgaker T, Jørgensen P, Olsen J (2000) Molecular electronic-structure theory. Wiley, Chichester
2. Whitten JL (1973) J Chem Phys 58:4496
3. Golub GH, Van Loan CF (1996) Matrix computations, 3 edn. Johns Hopkins University Press, Baltimore
4. Häser M, Ahlrichs R (1989) J Comput Chem 10:104
5. Lambrecht DS, Ochsenfeld C (2005) J Chem Phys 123:184101
6. Lambrecht DS, Doser B, Ochsenfeld C (2005) J Chem Phys 123:184102
7. Doser B, Lambrecht DS, Ochsenfeld C (2008) Phys Chem Chem Phys 10:3335
8. Baerends EJ, Ellis DE, Ros P (1973) Chem Phys 2:41
9. Sambe H, Felton RH (1975) J Chem Phys 62:1122
10. Dunlap BI, Connolly JWD, Sabin JR (1979) J Chem Phys 71:3396
11. Dunlap BI, Connolly JWD, Sabin JR (1979) J Chem Phys 71:4993
12. Feyereisen M, Fitzgerald G, Komornicki A (1993) Chem Phys Lett 208:359
13. Vahtras O, Almlöf J, Feyereisen M (1993) Chem Phys Lett 213:514
14. Dunlap BI (2000) J Mol Struct: Theochem 501–502:221
15. Dunlap BI (2000) J Mol Struct: Theochem 529:37
16. Dunlap BI (2000) Phys Chem Chem Phys 2:2113
17. Dunlap BI (2003) J Phys Chem A 107:10082
18. Sodt A, Head-Gordon M (2008) J Chem Phys 128:104106
19. Reine S, Tellgren E, Krapp A, Kjærgaard T, Helgaker T, Jansík B, Høst S, Sałek P (2008) J Chem Phys 129:104101
20. Werner HJ, Manby FR, Knowles PJ (2003) J Chem Phys 118:8149
21. Maschio L, Usvyat D, Manby FR, Casassa S, Pisani C, Schütz M (2007) Phys Rev B 76:075101
22. Eichkorn K, Treutler O, Öm H, Häser M, Ahlrichs R (1995) Chem Phys Lett 240:283
23. Eichkorn K, Weigend F, Treutler O, Ahlrichs R (1997) Theor Chem Acc 97:119
24. Weigend F (2006) Phys Chem Chem Phys 8:1057
25. Weigend F (2002) Phys Chem Chem Phys 4:4285

26. Weigend F (2008) J Comput Chem 29:167
27. Weigend F, Häser M, Patzelt H, Ahlrichs R (1998) Chem Phys Lett 294:143
28. Weigend F, Köhn A, Hättig C (2002) J Chem Phys 116:3175
29. Hättig C (2005) Phys Chem Chem Phys 7:59
30. Hellweg A, Hättig C, Höfener S, Klopper W (2007) Theor Chem Acc 117:587
31. Hill JG, Platts JA (2008) J Chem Phys 128:044104
32. Benoit CE (1924) Bulletin Geodesique 7:67
33. Beebe NHF, Linderberg J (1977) Int J Quantum Chem 12:683
34. Røeggen I, Wisløff-Nielsen E (1986) Chem Phys Lett 132:154
35. O'Neal DW, Simons J (1989) Int J Quantum Chem 36:673
36. Wilson S (1990) Comput Phys Commun 58:71
37. Koch H, Sánchez de Merás A, Pedersen TB (2003) J Chem Phys 118:9481
38. Boman L, Koch H, Sánchez de Merás A (2008) J Chem Phys 129:134107
39. Varganov SA, Gilbert ATB, Deplazes E, Gill PMW (2008) J Chem Phys 128:201104
40. Gill PMW, Gilbert ATB (2009) Chem Phys 356:86
41. Aquilante F, Pedersen TB, Lindh R (2007) J Chem Phys 126:194106
42. Aquilante F, Pedersen TB (2007) Chem Phys Lett 449:354
43. Aquilante F, Pedersen TB, Lindh R, Roos BO, Sánchez de Merás A, Koch H (2008) J Chem Phys 129:024113
44. Aquilante F, Malmqvist PÅ, Pedersen TB, Ghosh A, Roos BO (2008) J Chem Theory Comput 4:694
45. Pedersen TB, Sánchez de Merás AMJ, Koch H (2004) J Chem Phys 120:8887
46. Aquilante F, De Vico L, Ferré N, Ghigo G, Malmqvist PÅ, Neogrády P, Pedersen TB, Pitoňák M, Reiher M, Roos BO, Serrano-Andrés L, Urban M, Veryazov V, Lindh R (2009) J Comput Chem (in press). doi:10.1002/jcc.21318
47. Pierloot K, Vancoillie S (2008) J Chem Phys 128:034104
48. Radon M, Pierloot K (2008) J Phys Chem A 112:11824
49. MOLCAS 7 (2007) University of Lund, Sweden. See http://www.molcas.org
50. Karlström G, Lindh R, Malmqvist PÅ, Roos BO, Ryde U, Veryazov V, Widmark PO, Cossi M, Schimmelpfennig B, Neogrády P, Seijo L (2003) Comput Mater Sci 28:222
51. Veryazov V, Widmark PO, Serrano-Andrés L, Roos BO (2004) Int J Quantum Chem 100:626
52. Pitoňák M, Neogrády P, Řezáč J, Jurečka P, Urban M, Hobza P (2008) J Chem Theory Comput 4:1829
53. Pitoňák M, Riley KE, Neogrády P, Hobza P (2008) ChemPhysChem 9:1636
54. Pluháčková K, Grimme S, Hobza P (2008) J Phys Chem A 112:12469
55. Sodt A, Subotnik JE, Head-Gordon M (2006) J Chem Phys 125:194109
56. Schäfer A, Horn H, Ahlrichs R (1992) J Chem Phys 97:2571
57. Schäfer A, Huber H, Ahlrichs R (1994) J Chem Phys 100:5829
58. Skylaris CK, Gagliardi L, Handy NC, Ioannou AG, Spencer S, Willets A (2000) J Mol Struct: Theochem 501:229
59. Hättig C, Weigend F (2000) J Chem Phys 113:5154
60. Schütz M, Manby FR (2003) Phys Chem Chem Phys 5:3349
61. Jung Y, Sodt A, Gill PMW, Head-Gordon M (2005) Proc Natl Acad Sci USA 102:6692
62. Aquilante F, Lindh R, Pedersen TB (2007) J Chem Phys 127:114107
63. Aquilante F, Gagliardi L, Pedersen TB, Lindh R (2009) J Chem Phys 130:154107
64. Boström J, Aquilante F, Pedersen TB, Lindh R (2009) J Chem Theory Comput 5:1545
65. Røeggen I, Johansen T (2008) J Chem Phys 128:194107
66. Aquilante F, Lindh R, Pedersen TB (2008) J Chem Phys 129:034106
67. Ten-no S, Iwata S (1995) Chem Phys Lett 240:578
68. Ten-no S, Iwata S (1996) J Chem Phys 105:3604
69. Widmark PO, Malmqvist PÅ, Roos BO (1990) Theor Chim Acta 77:291
70. Widmark PO, Persson BJ, Roos BO (1991) Theor Chim Acta 79:419
71. Gallant RT, St-Amant A (1996) Chem Phys Lett 256:569
72. Fonseca Guerra C, Snijders JG, te Velde G, Baerends EJ (1998) Theor Chem Acc 99:391
73. Polly R, Werner HJ, Manby FR, Knowles PJ (2004) Mol Phys 102: 2311
74. Sałek P, Høst S, Thøgersen L, Jørgensen P, Manninen P, Olsen J, Jansík B, Reine S, Pawlowski F, Tellgren E, Helgaker T, Coriani S (2007) J Chem Phys 126:114110
75. Roos BO, Lindh R, Malmqvist PÅ, Veryazov V, Widmark PO (2004) J Phys Chem A 108:2851
76. Aquilante F, Todorova TK, Gagliardi L, Pedersen TB, Roos BO (2009) J Chem Phys (Accepted for publication)
77. Pedersen TB, Koch H, Boman L, Sánchez de Merás AMJ (2004) Chem Phys Lett 393:319
78. Weigend F, Kattannek M, Ahlrichs R (2009) J Chem Phys 130:164106