Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Tensor-structured factorized calculation of two-electron integrals in a general basis

(revised version: November 2012)

by

Venera Khoromskaia, Boris N. Khoromskij, and Reinhold Schneider

Preprint no.: 29 2012

Tensor-structured factorized calculation of two-electron integrals in a general basis

V. KHOROMSKAIA,* B. N. KHOROMSKIJ,** R. SCHNEIDER◇

Abstract

In this paper, the problem of efficient grid-based computation of the two-electron integrals (TEI) in a general basis is considered. We introduce the novel multiple tensor factorizations of the TEI unfolding matrix which decrease the computational demands for the evaluation of TEI in several aspects. Using the reduced higher-order SVD the redundancy-free product-basis set is constructed that diminishes dramatically the initial number $O(N_b^2)$ of 3D convolutions, defined over cross-products of N_b basis functions, to $O(N_b)$ scaling. The tensor-structured numerical integration with the 3D Newton convolving kernel is performed in 1D complexity, thus enabling high resolution over fine 3D Cartesian grids. Furthermore, using the quantized approximation of long vectors ensures the logarithmic storage complexity in the grid-size. Finally, we present and analyze two approaches to compute the Cholesky decomposition of TEI matrix based on two types of precomputed factorizations. We show that further compression is possible via columnwise quantization of the Cholesky factors. Our "black-box" approach essentially relaxes limitations on the traditional Gaussian-type basis sets, giving an alternative choice of rather general low-rank basis functions represented only by their 1D samplings on a tensor grid. Numerical tests for some moderate size compact molecules demonstrate the expected asymptotic performance.

Key words: Hartree-Fock equation, Coloumb and exchange matrices, two-electron integrals, tensor-structured approximation, truncated Cholesky decomposition, reduced higher order SVD, quantized approximation of vectors.
AMS Subject Classification: 65F30, 65F50, 65N35, 65F10

1 Introduction

Two-electron integrals (TEI) tensor, also known as the Fock integrals, is the principal ingredient in electronic and molecular structure calculations. In particular, the corresponding coefficient tensor arises in ab-initio Hartree-Fock (HF) calculations, in post Hartree-Fock

*Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (vekh@mis.mpg.de).

**Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (bokh@mis.mpg.de).

◇TU Berlin, Strasse des 17 Juni 136, D-10623 Berlin, Germany (schneidr@tu-berlin.de).

models (MP2, CCSD, Jastrow factors, etc.) as well as in the core Hamiltonian appearing in FCI-DMRG calculations [2, 41, 34, 15, 28].

Given the finite basis set $\{g_\mu\}_{1\le\mu\le N_b}$, $g_\mu \in H^1(\mathbb{R}^3)$, the associated fourth order two-electron integrals tensor, $\mathbf{B} = [b_{\mu\nu\kappa\lambda}] \in \mathbb{R}^{N_b\times N_b\times N_b\times N_b}$, is defined entrywise by

$$b_{\mu\nu\kappa\lambda} = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{g_\mu(x)g_\nu(x)g_\kappa(y)g_\lambda(y)}{\|x-y\|}dxdy, \quad \mu,\nu,\kappa,\lambda \in \{1,...,N_b\} =: I_b. \tag{1.1}$$

The fast and accurate evaluation of the 4th order TEI tensor $\mathbf{B}$ of size N_b^4, is the challenging computational problem since it includes multiple 3D convolutions of the Newton kernel $1/\|x-y\|$, $x, y \in \mathbb{R}^3$ with strongly varying product-basis functions. Hence, in the limit of large N_b, the efficient numerical treatment and storage of the TEI tensor is considered as one of the central tasks in electronic structure calculations.

The traditional analytical integration using the representation of electronic orbitals in a Gaussian-type basis is the basement of most *ab-initio* quantum chemical packages. Hence, the choice of a basis set, $\{g_\mu\}_{1\le\mu\le N_b}$, is essentially restricted by the "analytic" integrability for efficient computations of the tensor entries represented by 6D integrals in (1.1). This approach possesses intrinsic limitations concerning the non-alternative constraint to the Gaussian-type basis functions, which may become unstable and redundant for higher accuracies, larger molecules or when considering heavy nuclei.

It is known since long time in quantum chemistry simulations [5, 41, 43] that, in case of compact molecules, the (pivoted) incomplete Cholesky factorization of the $N_b^2 \times N_b^2$ TEI matrix unfolding,

$$B = [b_{\mu\nu;\kappa\lambda}] := mat(\mathbf{B}) \quad \text{over} \quad (I_b \otimes I_b) \times (I_b \otimes I_b), \tag{1.2}$$

reduces the asymptotic storage of the resultant low-rank approximation to $O(N_b^3)$. It was observed in numerical experiments that the particular rank-bound in the Cholesky decomposition scales linearly in N_b, depending mildly, say logarithmically, on the error in the rank truncation. We refer to [16, 8, 13, 4, 37, 35] for more details on the algebraic aspects of matrix Cholesky decomposition and the related ACA techniques. The Cholesky decomposition is applicable since the TEI matrix B is, indeed, the symmetric Gramm matrix of the product basis set $\{g_\mu g_\nu\}$ in the Coulomb metric $\langle\cdot, \frac{1}{\|x-y\|}\cdot\rangle$, ensuring its positive semidefiniteness. In some cases it is possible to reduce the storage even to $O(N_b \log N_b)$ taking into account the pointwise sparsity of the matrix B in calculation of large rather extended systems [41].

In this paper, the Cholesky decomposition of a matrix B is calculated using the precomputed factorizations of this matrix in two different forms. The first kind of factorizations can be interpreted as the Galerkin representation (1.1) discretized in the complete product basis that incorporated the full set of 3D convolutions applied to $\{g_\mu g_\nu\}$, $1 \le \mu,\nu \le N_b$ over a large $n \times n \times n$ 3D tensor grid. We provide the theoretical rank estimates for the quantized-canonical approximation to large grid-supported data arrays leading to the $O(\log n)$ storage cost. In the second approach, we construct the algebraically optimized redundancy-free factorization to the TEI matrix B, based on the reduced higher-order SVD [22] to obtain the low-rank separable representation of the discretized basis functions $\{g_\mu g_\nu\}$. Numerical experiments show that this minimizes the dimension of dominating subspace in $span\{g_\mu g_\nu\}$ to $R_G \le N_b$, which allows to reduce the number of 3D convolutions by the

order of magnitude, from $O(N_b^2)$ to R_G. Combined with the quantized-canonical tensor decompositions of long spacial n-vectors this leads to the logarithmic scaling in n for storage, $O(R_G \log n + N_b^2 R_G)$. An essential compression rate is observed in numerical experiments even for compact molecules, becoming stronger for more stretched compounds.

Computation of the rank-R_B Cholesky decomposition employs only $R_B = O(N_b)$ selected columns in the TEI matrix B calculated from precomputed factorizations of this matrix. We show by numerical experiments that each long N_b^2-vector of the L-factor in the Cholesky LL^T-decomposition can be further compressed using the quantized-TT (QTT) approximation reducing the total storage from $O(N_b^3)$ to $O(N_b N_{orb}^2)$, where the number of electron orbitals, N_{orb}, usually satisfies $N_b \sim 10 N_{orb}$.

The presented grid-based approach benefits from the fast $O(n \log n)$ tensor-product convolution with the 6D Newton kernel over a large $n^3 \times n^3$ grid, [25], which has already proved the numerical efficiency in the evaluation of the Hartree and exchange integrals [22, 19, 23]. In these papers, describing the numerical solution of the Hartree-Fock equation in the multilevel (multigrid) tensor-structured format, both the Coulomb and exchange operators are calculated directly "on the fly" at each DIIS iteration, thus the use of TEI was avoided at the expense of time loops. This initial approach was the crucial point to test and analyze the validity of the tensor-structured methods in Hartree-Fock calculations.

The beneficial feature of the grid-based tensor-structured methods is a substitution of the 3D numerical integration by multilinear algebraic procedures like the scalar, Hadamard, and convolution products, with linear 1D complexity, $O(n)$. On the one hand, this weak dependence on the grid-size is the ultimate payoff for generality, in the sense that rather general approximating basis sets may be equally used instead of analytically integrable Gaussians. On the other hand, the approach also serves for structural simplicity of implementation, since the topology of the molecule is caught without any physical insight, only by the algebraically determined rank parameters of the fully grid-based numerical scheme.

Due to $O(n \log n)$ complexity of the algorithms, there are rather weak practical restrictions on the grid-size n allowing calculations on really large $n \times n \times n$ $3D$ Cartesian grids in the range $n \sim 10^3 \div 10^5$, avoiding the grid refinement. The latter allows high resolution of the order of the size of atomic nuclei. For storage consuming operations, the numerical expense can be reduced to logarithmic level, $O(\log n)$, by QTT representation of the discretized $3D$ basis functions and their convolutions.

We summarize that the rank-$O(N_b)$ Cholesky decomposition of B, combined with the canonical-QTT data compression of long vectors, allows to reduce the asymptotic complexity of grid-based tensor calculations in HF and some post-HF models. Notice that in the recent years the grid-based numerical methods became attractive in electronic and molecular structure calculations since they allow, in principle, the efficient approximation to the physical entities of interest with a controllable precision [14, 3, 23]. Alternative approaches to optimization of the HF, MPx, CCSD and other post-HF calculations can be based on using physical insight to sparsify the TEI tensor $\mathbf{B}$ by zero-out all "small" elements [41, 34, 2].

The tensor numerical methods are getting established for the solution of multidimensional differential equations, since they allow to avoid the so-called "curse of dimensionality". Some algebraic tensor algorithms for the low-rank approximation of multivariate data have been originally worked out for the problems of chemometrics and signal processing, see review [27] and references therein. In the recent years, the main

theoretical concepts of the rank-structured representation to operators and functions as well as the novel tensor-structured numerical methods designed for solving multi-dimensional PDEs have been developed and successfully applied [26, 7, 23, 20, 21, 9].

The rest of the paper is organized as follows. Section 2 outlines the basic tensor formats and some tensor-structured bilinear operations. Section 3 presents two approaches on the rank-structured representation to the TEI matrix B. In §3.1, we discuss the Galerkin-type representation to this matrix in the complete initial product-basis, and in §3.2 we provide the theoretical rank bounds on the QTT-approximation applied to the large grid-supported data arrays. This reduces the representation cost to the $O(\log n)$ scaling. §3.3 describes the new redundancy-free factorization to the matrix B in the optimized product-basis reducing the number of 3D convolutions by the order of magnitude. The subsequent Cholesky decomposition applies to both factorizations of a matrix B as above. Section 3.5 analyses numerically the effect of QTT compression of the Cholesky factor in B, finding out that the average QTT-rank is approximately a multiple of 3 and the number of orbitals in a molecule, $\sim 3N_{orb}$. Hence the representation complexity can be reduced to $O(N_b N_{orb}^2)$. Section 4 discusses the motivating applications of TEI tensor to electronic structure calculations including the Hartree-Fock equations and MP2 perturbation scheme, as well as the higher dimensional FCI-DMRG models. Given the Cholesky factor of the matrix B, the Coulomb and exchange parts in the Fock matrix are obtained by contracted products with the density matrix D according to (4.5), see §4.2. Calculation of the Coulomb and exchange matrices based on the ε-truncated Cholesky factorization is illustrated on examples of some moderate size compact molecules demonstrating the asymptotic performance of the presented schemes.

All algorithms are implemented in Matlab on a SUN station using 8 Opteron Dual-Core/2600 processors.

2 Basic tensor formats

In this section, we sketch some tensor formats used in this paper. Tensors of order d are defined as the elements of finite dimensional tensor-product Hilbert space $\mathbb{W}_{\mathbf{n}} \equiv \mathbb{W}_{\mathbf{n},d}$ of the d-fold, $n_1 \times ... \times n_d$ real/complex-valued arrays, and equipped with the Euclidean scalar product $\langle \cdot, \cdot \rangle : \mathbb{W}_{\mathbf{n}} \times \mathbb{W}_{\mathbf{n}} \to \mathbb{R}$. Each tensor in $\mathbb{W}_{\mathbf{n}}$ can be represented component-wise,

$$\mathbf{A} = [a(i_1, ..., i_d)] \quad \text{with} \quad i_\ell \in I_\ell := \{1, ..., n_\ell\}, \quad \text{and} \quad \mathbf{n} = (n_1, ..., n_d),$$

where for the ease of presentation, we mainly consider the equal-size tensors, i.e., $I_\ell = I = \{1, ..., n\}$ $(\ell = 1, ..., d)$. We call the elements of $\mathbb{W}_{\mathbf{n}} = \mathbb{R}^{\mathcal{I}}$ with $\mathcal{I} = I_1 \times ... \times I_d$, as $n^{\otimes d}$ tensors. The dimension of the tensor-product Hilbert space $\mathbb{W}_{\mathbf{n}}$ scales exponentially in d, $\dim \mathbb{W}_{\mathbf{n},d} = n^d$, implying exponential storage cost for a tensor ("curse of dimensionality").

To get rid of the exponential scaling in the dimension, the rank-structured tensor formats can be applied. The basic formats include the so-called canonical and Tucker tensor representations, see [27] and references therein. The R-term canonical representation of a tensor is defined as

$$\mathbf{A} = \sum_{k=1}^{R} c_k u_k^{(1)} \otimes \ldots \otimes u_k^{(d)}, \quad c_k \in \mathbb{R}, \tag{2.1}$$

where $u_k^{(\ell)} \in \mathbb{R}^{I_\ell}$ are normalized vectors (also called as skeleton vectors). In tensor numerical methods the R-term sum (2.1) usually approximates the initial data array up to certain tolerance $\varepsilon > 0$ providing the upper bound on the canonical ε-rank of a tensor, $rank_\varepsilon(\mathbf{A}) \leq R$. In a similar way, ε-rank can be estimated for any tensor format to be considered so far in this paper (Tucker, TT, QTT, canonical-QTT, etc.).

Given the d-tuple of rank parameters, $\mathbf{r} = (r_1, ..., r_d)$, the Tucker representation of a tensor $\mathbf{A}$ is defined by

$$\mathbf{A} = \sum\nolimits_{\nu_1=1}^{r_1} \dots \sum\nolimits_{\nu_d=1}^{r_d} \beta_{\nu_1,\dots,\nu_d}\, v_{\nu_1}^{(1)} \otimes \dots \otimes v_{\nu_d}^{(d)},$$

where $v_{\nu_\ell}^{(\ell)} \in \mathbb{R}^{I_\ell}$ $(1 \leq \nu_\ell \leq r_\ell)$ are the orthonormal vectors. The parameter $r = \max\limits_{\ell}\{r_\ell\}$ is called the Tucker rank, and the coefficients tensor $\boldsymbol{\beta} = [\beta_{\nu_1,\dots,\nu_d}]$ is called the core tensor (usually, for function related tensors, $r \ll n$).

Rank-structured tensor representations allow efficient reduction of storage and fast multilinear algebra with linear scaling in the dimension d. We illustrate how the standard multilinear algebra operations on tensors $\mathbf{A}_1$, $\mathbf{A}_2$, represented in the canonical format (2.1),

$$\mathbf{A}_1 = \sum_{k=1}^{R_1} c_k u_k^{(1)} \otimes \dots \otimes u_k^{(d)}, \quad \mathbf{A}_2 = \sum_{m=1}^{R_2} b_m v_m^{(1)} \otimes \dots \otimes v_m^{(d)}, \tag{2.2}$$

are reduced to independent treatment of the univariate canonical vectors. For simplicity of notation, we assume that $n_\ell = n$. For given canonical tensors $\mathbf{A}_1$, $\mathbf{A}_2$, the Euclidean scalar product can be computed by

$$\langle \mathbf{A}_1, \mathbf{A}_2 \rangle := \sum_{k=1}^{R_1} \sum_{m=1}^{R_2} c_k b_m \prod_{\ell=1}^{d} \left\langle u_k^{(\ell)}, v_m^{(\ell)} \right\rangle, \tag{2.3}$$

at the expense $O(dnR_1R_2)$. The Hadamard product of tensors $\mathbf{A}_1, \mathbf{A}_2$ given in the canonical format (2.2) is calculated in $O(dnR_1R_2)$ operations by

$$\mathbf{A}_1 \odot \mathbf{A}_2 := \sum_{k=1}^{R_1} \sum_{m=1}^{R_2} c_k b_m \left(u_k^{(1)} \odot v_m^{(1)} \right) \otimes \dots \otimes \left(u_k^{(d)} \odot v_m^{(d)} \right). \tag{2.4}$$

In electronic structure calculations, the three-dimensional convolution transform with the Newton convolving kernel, $\frac{1}{\|x-y\|}$, is one of the most computationally expensive operations. We employ the tensor-structured computation of this transform over large $n \times n \times n$ Cartesian grid with $O(n \log n)$ complexity introduced in [25].

The convolution product of the canonical tensors $\mathbf{A}_1$, $\mathbf{A}_2$, is represented by the double sum

$$\mathbf{A}_1 * \mathbf{A}_2 = \sum_{k=1}^{R_1} \sum_{m=1}^{R_2} c_k b_m \left(u_k^{(1)} * v_m^{(1)} \right) \otimes \dots \otimes \left(u_k^{(d)} * v_m^{(d)} \right), \tag{2.5}$$

where $u_k^{(\ell)} * v_m^{(\ell)}$ denotes the convolution product of vectors. The complexity is estimated by $O(dR_1R_2n \log n)$. In our applications the tensor product convolution considerably outperforms the conventional 3D FFT-based algorithm having the complexity $O(n^3 \log n)$, see numerics in [22].

When using the rank-structured representations of functions and operators in the Hartree-Fock equation, the 3D and 6D integrations are replaced by multilinear algebra operations such as the scalar and Hadamard products, the 3D convolution transforms which are implemented with an almost $O(n)$-complexity [22, 19]. However, the rank-structured operations mandatory lead to increase of tensor ranks. For tensor rank reduction one can apply the robust algorithm based on the canonical-to-Tucker and Tucker-to-canonical transforms introduced in [22], which is also of linear complexity with respect to parameters of the target 3D canonical tensor, $O(Rn)$.

The matrix product states (MPS) tensor factorization [42, 39, 38] was proved to be efficient in high-dimensional electronic/molecular structure calculations, in quantum computing and in multiparticle dynamics. In the mathematical literature the MPS-type tensor decompositions were recently recognized and further developed as the so-called tensor train (TT) format [30, 32]. For a given rank parameter $\mathbf{r} = (r_0, ..., r_d)$, and the respective index sets $J_\ell = \{1, ..., r_\ell\}$ $(\ell = 0, 1, ..., d)$, with the constraints $r_0 = r_d = 1$ (MPS with open boundary conditions), the rank-$\mathbf{r}$ tensor train format contains all elements $\mathbf{A} = [a(\mathbf{i})]$, $\mathbf{i} = (i_1, ..., i_d)$, in $\mathbb{W}_{\mathbf{n}} = \mathbb{R}^{\mathcal{I}}$ that can be represented as the chain of contracted products of 3-tensors over the d-fold product index set $J := \times_{\ell=1}^{d} J_\ell$,

$$a(\mathbf{i}) = \sum_{\alpha_1 \in J_1} \cdots \sum_{\alpha_d \in J_d} G^{(1)}(i_1, \alpha_1) G^{(2)}(\alpha_1, i_2, \alpha_2) \cdots G^{(d-1)}(\alpha_{d-2}, i_{d-1}, \alpha_{d-1}) G^{(d)}(\alpha_{d-1}, i_d).$$

In the matrix form we have the entry-wise MPS-type factorization

$$a(i_1, i_2, \ldots, i_d) = A^{(1)}_{i_1} A^{(2)}_{i_2} \ldots A^{(d)}_{i_d}, \tag{2.6}$$

where each $A^{(k)}_{i_k}$ is $r_{k-1} \times r_k$ matrix.

The TT representation reduces the storage complexity of $n^{\otimes d}$ tensor to $O(dr^2 n)$, $r = \max r_k$ (same for more general MPS-type formats). The important multilinear algebraic operations with TT tensors can be implemented with linear complexity scaling in d and n. For example, the scalar product of tensors $\langle \mathbf{A}, \mathbf{B} \rangle$ can be computed at the cost $O(dr^3 n)$.

The novel *quantics-TT* (QTT) approximation method was invented by one of the authors in 2009[1] (see [24]) as a fascinating tool to compress function related vectors or $n^{\otimes d}$-tensors to the logarithmic amount of data, $O(d \log_2 n)$. The "quantized" approximation is proven to provide noticeable separation properties for the wide class of functions [24] including discrete exponential, trigonometric, piecewise polynomial, and many other types of analytic functional n-vectors that ensure the low TT-rank of $2 \times ... \times 2$-quantized images. This reduces the storage cost $O(n)$ to the logarithmic scale,

$$2k^2 L \ll 2^L, \quad n = 2^L,$$

where k is the (small) QTT-rank, thus yielding the redundancy-free tensor representation of long functional vectors. For example, a discretized exponential vector (say, signal) on a large grid of size $n = 2^{50}$ ($n \approx 10^{15}$), can be stored by only $2 \cdot 1^2 \cdot 50 = 10^2$ numbers parametrizing the quantized image of this vector.

[1] B. Khoromskij, Preprint 55/2009, Max-Planck Institute for Mathematics in the Sciences, Leipzig 2009. http://www.mis.mpg.de/publications/preprints/2009/prepr2009-55.html

It is worth to note that the effect of low-rank numerical TT-representation to the dyadically reshaped $2^L \times 2^L$-discretization of the 1D-Laplacian was, first, observed in numerical tests [31]. This numerics and related techniques in signal processing based on folding of vectors to 2D-arrays [36] motivated the invention of QTT-approximation theory for function related vectors/tensors and its further generalizations [24]. We refer to survey [26] for the state-of-the-art in the QTT method for functions and operators.

To sketch the construction, we suppose that $n = 2^L$ with some $L = 1, 2, ...$ The quantization of $n^{\otimes d}$ tensor into the element of auxiliary D-dimensional tensor space with $D = d \log_2 n$ is performed by the dyadic reshaping of indexes. The respective binary folding of degree L,

$$\mathcal{Q}_{d,L} : \mathbb{W}_{\mathbf{n},d} \to \mathbb{W}_{\mathbf{m},dL}, \quad \mathbf{m} = (\mathbf{m}_1, ..., \mathbf{m}_d), \quad \mathbf{m}_\ell = (m_{\ell,1}, ..., m_{\ell,L}),$$

with $m_{\ell,\nu} \in \{1, 2\}$ for $\nu = 1, ..., L$, $(\ell = 1, ..., d)$, that reshapes the initial $n_1 \times ... \times n_d$ tensor in $\mathbb{W}_{\mathbf{n},d}$ to the elements of *quantized tensor space*,

$$\mathbb{W}_{\mathbf{m},dL} = \bigotimes_{\ell=1}^{d} \mathbb{K}^{n_\ell} = \bigotimes_{\ell=1}^{d} \bigotimes_{\nu=1}^{L} \mathbb{K}^2, \quad \mathbb{K} \in \{\mathbb{R}, \mathbb{C}\},$$

is defined for $d = 1$ as follows: a vector $\mathbf{X} = [X(i)]_{i \in I} \in \mathbb{W}_{n,1}$, is reshaped to the element of $\mathbb{W}_{2,L}$ by

$$\mathcal{Q}_{1,L} : \mathbf{X} \to \mathbf{Y} = [Y(\mathbf{j})] := [X(i)], \quad \mathbf{j} = \{j_1, ..., j_L\},$$

with $j_\nu \in \{1, 2\}$ for $\nu = 1, ..., L$. For fixed i, $j_\nu = j_\nu(i)$ is defined by $j_\nu - 1 = C_{-1+\nu}$, where the $C_{-1+\nu}$ are found from the binary representation (binary coding) of $i - 1$,

$$i - 1 = C_0 + C_1 2^1 + \cdots + C_{L-1} 2^{L-1} \equiv \sum_{\nu=1}^{L} (j_\nu - 1) 2^{\nu-1}.$$

For $d > 1$, the construction is similar, see [24].

Now the main idea of the QTT approximation method is to solve the initial computational problem in the quantized tensor space $\mathbb{W}_{\mathbf{m},dL}$ of higher dimension, where the functions and operators allow good separation properties.

It should be noted that the simple reshaping transforms of vectors to tensors (tensorization), or tensors to vectors or matrices (vectorization, matricization) do not hold any size-reduction properties. They can be interpreted just as the dual isometry specified by the commonly used schemes of reordering a multivariate index set, as, for example, in the standard "reshape" command in Matlab[2]. The core of the *QTT approximation method* [24] applied to discrete functions and operators is the efficient low-rank TT-tensor representation (approximation) to their quantized images justified by the sound QTT-approximation theory, where the 'indivisible' mode size '2' represents a *quant* of information (compare with quantum bits, i.e. qubits in quantum computations). This concept enables us to introduce the new generation of tensor numerical methods with logarithmic complexity scaling for solving multi-dimensional PDEs, by their approximation on the QTT-manifolds.

[2]In some subsequent papers the quantized approximation method was renamed as "tensorization" of tensors, which, obviously, does not reflect the main idea of the approach, but increases the redundancy in use of the key-word 'tensor'.

Remark 2.1 *Every $2^{\otimes dL}$ tensor in the quantized tensor space $\mathbb{W}_{2,dL}$ can be represented (approximated) in the TT format leading to the QTT representation of high order tensors. Assuming that $r_k \le r$, $k = 1, ..., dL$, the complexity of QTT representation can be estimated by $O(dr^2 \log n)$, providing* log*-volume asymptotic in the size of initial tensor, n^d. QTT representation of skeleton vectors in the canonical format leads to the so-called canonical-QTT (C-QTT) format to be applied below. This format is characterized by the canonical rank R and the QTT mode ranks, $\mathbf{r}^{(\ell)}$. Further promising generalizations can be based on a combination of the Tucker, TT and QTT formats [26, 11].*

Similar to the canonical and Tucker formats, the basic multilinear algebraic operations with TT and QTT tensors can be implemented with linear complexity scaling in n and d. Moreover, the important classes of linear operators can be efficiently treated in the QTT and C-QTT formats (see survey [26] and references therein). The manifold of rank-$\mathbf{r}$ TT (and QTT) tensors is known to be closed in the Frobenius norm ensuring the stability of approximation.

In our application the two-electron integrals $[b_{\mu\nu\kappa\lambda}]$ constitute the $N_b^{\otimes 4}$ tensor of order 4, requiring N_b^4 storage size. Each entry is calculated as 3D tensor convolution over large tensor grid. Hence, in the case of a large number of basis functions, N_b, the direct calculations become prohibitive already for N_b of order few hundreds. In this way, the canonical, Tucker and TT tensor formats can be considered only for small basis-size N_b since they rely on the full format target TEI tensor. Another bottleneck is due to bad scaling of the rank parameters, at least as $O(N_b)$, that makes the representations non-tractable for N_b on the hundred scale.

In the following we focus on the combined canonical and QTT representations which appear as the main building block in the new tensor factorization and Cholesky decomposition of the TEI matrix.

3 TEI tensor in combined rank-structured formats

In sections 3.1 - 3.3 we describe the Galerkin-type factorizations to the full tensor TEI $\mathbf{B}$ and the TEI matrix B, respectively. Sections 3.4 - 3.5 considers the approximate Cholesky decomposition of the matrix B based only on the computation of the selected set of columns in B.

3.1 Galerkin factorization of TEI tensor in the full product basis

In this section we describe a factorization scheme for the efficient representation of the TEI tensor $\mathbf{B}$ and the respective unfolding matrix $B = mat(\mathbf{B})$ in the full product basis.

Suppose that all basis functions $\{g_\mu\}_{1\le\mu\le N_b}$, are supported by a finite box $\Omega = [-b, b]^3 \subset \mathbb{R}^3$, and assume, for ease of presentation, that $rank(g_\mu) = 1$ (see Remark 3.3). The size of the computational box is chosen in such a way that the truncated part of the most slowly decaying basis function does not exceed the given tolerance $\varepsilon > 0$. Since the exponential decay in molecular orbitals, the parameter $b > 0$ is chosen to be only few times larger than the molecular size.

Introducing the $n \times n \times n$ Cartesian grid over Ω (also denoted by $n^{\otimes 3}$-grid) and using the standard collocation discretization in the volume by piecewise constant basis functions, we get a grid tensor representation of the initial basis set $g_\mu(x)$, $x \in \mathbb{R}^3$, via rank-1 tensors,

$$g_\mu(x) = g_\mu^{(1)}(x_1) g_\mu^{(2)}(x_2) g_\mu^{(3)}(x_3) \approx \mathbf{G}_\mu = G_\mu^{(1)} \otimes G_\mu^{(2)} \otimes G_\mu^{(3)} \in \mathbb{R}^{n \times n \times n}.$$

Then the entries of $\mathbf{B}$ can be written by using the tensor scalar product over "grid" indices,

$$b_{\mu\nu\kappa\lambda} = \langle \mathbf{G}_{\mu\nu}, \mathbf{H}_{\kappa\lambda} \rangle_{n^{\otimes 3}}, \tag{3.1}$$

where

$$\mathbf{G}_{\mu\nu} = \mathbf{G}_\mu \odot \mathbf{G}_\nu \in \mathbb{R}^{n^{\otimes 3}}, \quad \mathbf{H}_{\kappa\lambda} = \mathbf{P}_{\mathcal{N}} * \mathbf{G}_{\kappa\lambda} \in \mathbb{R}^{n^{\otimes 3}}, \quad \mu, \nu, \kappa, \lambda \in \{1, ..., N_b\}, \tag{3.2}$$

with the rank-$R_{\mathcal{N}}$ canonical tensor $\mathbf{P}_{\mathcal{N}} \in \mathbb{R}^{n^{\otimes 3}}$ approximating the Newton potential $\frac{1}{\|x\|}$ (see [6, 22] for more details). Here and in the following $*$ stands for the 3D tensor convolution (2.5), and $\odot$ denotes the 3D Hadamard product (2.4).

The element-wise accuracy of the tensor representation (3.1) is estimated by $O(h^2)$, where $h = 2b/n$ is the step-size of the Cartesian grid [25]. The Richardson extrapolation reduces the error to $O(h^3)$.

Remark 3.1 *It is worth to emphasize that in our scheme the $n^{\otimes 3}$ tensor Cartesian grid does not depend on the positions of nuclei in a molecule. Consequently, the simultaneous rotation and translation of the nuclei positions does not effect the approximation error on the level of $O(h^2)$. The more detailed analysis of numerical effects due to the change of coordinates will be considered elsewhere.*

Remark 3.2 *The TEI tensor* $\mathbf{B}$ *has multiple symmetries,*

$$b_{\mu\nu\kappa\lambda} = b_{\nu\mu\kappa\lambda} = b_{\mu\nu\lambda\kappa} = b_{\kappa\lambda\mu\nu}, \quad \mu, \nu, \kappa, \lambda \in \{1, ..., N_b\}.$$

The result is a direct consequence of definition (1.1) (see also (3.1)) and symmetry of the convolution product. The above symmetry relation allows to reduce the number of precomputed entries in the full TEI tensor to $N_b^4/8$.

Let us introduce the 5th order tensors

$$\mathbf{G} = [\mathbf{G}_{\mu\nu}] \in \mathbb{R}^{N_b \times N_b \times n^{\otimes 3}} \quad \text{and} \quad \mathbf{H} = [\mathbf{H}_{\kappa\lambda}] \in \mathbb{R}^{N_b \times N_b \times n^{\otimes 3}}.$$

Then (3.1) is equivalent to the contracted-product representation over $n^{\otimes 3}$-grid indexes,

$$\mathbf{B} = \mathbf{G} \times_{n^{\otimes 3}} (\mathbf{P}_{\mathcal{N}} *_{n^{\otimes 3}} \mathbf{G}) = \langle \mathbf{G}, \mathbf{P}_{\mathcal{N}} *_{n^{\otimes 3}} \mathbf{G} \rangle_{n^{\otimes 3}} = \langle \mathbf{G}, \mathbf{H} \rangle_{n^{\otimes 3}}, \tag{3.3}$$

where the right-hand part is recognized as the discrete counterpart of the Galerkin representation (1.1) in the full product basis. When using the full grid calculations, the total storage cost for the $n \times n \times n$ product-basis tensor $\mathbf{G}$ and its convolution $\mathbf{H}$ amounts to $3\frac{N_b(N_b+1)}{2}n$ and $3R_{\mathcal{N}}\frac{N_b(N_b+1)}{2}n$, respectively. The numerical cost of N_b^2 tensor-product convolutions to compute $\mathbf{H}$ is estimated by $O(R_{\mathcal{N}} N_b^2 n \log n)$ [25]. Based on representation (3.3), each entry in the TEI tensor $\mathbf{B}$ can be calculated with the cost $O(R_{\mathcal{N}} n)$ which might be too expensive for the large grid-size n. In the next section we discuss the quantized representation of the canonical factors which reduces the cost of scalar products to the logarithmic level.

Remark 3.3 *If the separation rank of a basis set is larger than* 1 *then the complexity of scalar products in (3.3) will increase quadratically in this parameter, see (2.3). However, the use of basis functions with non-unit rank (say, Slater-type functions) can be motivated by the reduction of the basis size N_b, that has a fourth order effect on the complexity.*

3.2 QTT approximation of tensors G and H

Tensor factorization (3.3) means that as far as the tensors **G** and **H** are precomputed and stored efficiently in the canonical-QTT tensor format (can be an off-line step), further calculations of tensor entries in **B** can be performed by using only simple QTT-tensor scalar product multiplications with $O(\log n)$ complexity for each entry. In practical computations the payoff for grid calculus, $\log n$, becomes negligible and algorithms may perform, in general, as in the special cases with (mesh-less) analytically integrable GTO basis.

We point out, however, that the super-fast 1D QTT-convolution and QTT-FFT of complexity $O(\log^p n)$ proposed in [10, 18] outperform the full-grid 1D-FFT based tensor algorithm [25] only in asymptotical regime, i.e. for large enough vector-size n of order 10^6. Hence, in our case we apply the FFT-based canonical tensor convolution of complexity $O(n \log n)$.

At this point, we notice that large grids are mandatory for high resolution of the 3D GTO basis functions $\{g_k(x)\}$, which usually have strong singularities at the positions of the nuclei. It was demonstrated in [19, 22, 21] that the integral operators in the Hartree-Fock equation can be approximated with the satisfactory precision on $n^{\otimes 3}$ grids with n in the range $[10^4 \div 10^5]$.

To justify that in the case of large grids the canonical-QTT format allows the fast computation and efficient storage of tensors $\mathbf{G}_{\mu\nu}$ and $\mathbf{H}_{\mu\nu}$ ($\mu, \nu \in \{1, ..., N_b\}$) with logarithmic cost $O(\log n)$, we present the rigorous estimates on the rank parameters.

On the one hand, $rank(\mathbf{G}_{\mu\nu}) = 1$, and the ε-rank of $\mathbf{H}_{\mu\nu}$ is small, i.e.

$$rank(\mathbf{H}_{\mu\nu}) = rank(\mathbf{P}_{\mathcal{N}}) = O(|\log \varepsilon|), \tag{3.4}$$

see e.g. [12, 25, 6]. For example, $R_{\mathcal{N}} = rank(\mathbf{P}_{\mathcal{N}}) \approx 25$ for $\varepsilon = 10^{-6}$. It was proven that the approximation error of the tensor-product convolution with the discrete Newton kernel $\mathbf{P}_{\mathcal{N}}$ decays exponentially in the rank parameter $R_{\mathcal{N}}$ and quadratically in the mesh-size, $O(h^2)$, see [25]. Hence, this error can be effectively controlled by the choice of approximation parameters.

On the other hand, in the case of Gaussian-type AO basis (multiple of Gaussians and low-degree polynomials), and with fixed $\varepsilon > 0$, we are able to prove $rank_{QTT}(G_\mu^{(\ell)}) = const$, $\ell = 1, 2, 3$, with a small constant, and the same for the canonical vectors of $\mathbf{H}_{\mu\nu}$ as shown in the following Lemmata 3.4, and 3.5.

Lemma 3.4 *Given the rank-*1 *GTO basis, $\{g_k(x)\}$, and $\varepsilon > 0$, then the QTT ε-rank of the product basis functions is bounded by*

$$rank_{QTT,\varepsilon}(\mathbf{G}_{\mu\nu}) \le C\sqrt{|\log \varepsilon|}, \quad \mu, \nu = 1, ..., N_b,$$

uniformly in the grid-size n.

Proof. The proof is based on the observation that the product of two Gaussians is again a Gaussian but with a shifted center,

$$e^{-\lambda(x-a)2}e^{-\beta(x-b)2} = e^{\sigma}e^{-\gamma(x-c)2}, \quad \sigma = \frac{\lambda\beta(a-b)2}{\lambda+\beta},\ \gamma = \lambda+\beta,\ c = \frac{a\lambda+b\beta}{\lambda+\beta}.$$

In general, the canonical vectors generated by the resultant rank-1 function can be represented by vectors,

$$\{e^{-\lambda(x_i+\delta)^2}\} = C\{e^{-\lambda x_i^2}e^{-2\lambda\delta x_i}\}, \quad \delta \in [0,h],$$

where $\{x_i\}$ denotes the set of uniform sampling points. Since the QTT-rank of any uniformly sampled Gaussian vector is bounded by $C\sqrt{|\log\varepsilon|}$, see [9], and taking into account that the second factor in the right-hand side above is an exponential vector of rank-1, see [24], we arrive at the desired bound. In the presence of degree-p polynomial factors, their QTT rank is bounded by $p+1$ [24] and the result follows. ∎

Next let us estimate the canonical-QTT ranks of the tensor $\mathbf{H}_{\mu\nu}$ in (3.2).

Lemma 3.5 *Given the rank-1 GTO basis, $\{g_k(x)\}$, and $\varepsilon > 0$, then the ε-ranks in the canonical-QTT format are bounded by*

$$rank_C(\mathbf{H}_{\mu\nu}) \le C_0|\log\varepsilon|\log n, \quad rank_{C-QTT}(\mathbf{H}_{\mu\nu}) \le C_0\sqrt{|\log\varepsilon|}, \quad \mu,\nu = 1,...,N_b,$$

uniformly in the grid-size n.

Proof. Inspecting the constructive description of the canonical approximation $\mathbf{P}_{\mathcal{N}}$ to the Newton potential [6], we further consider this tensor as the rank-$R_{\mathcal{N}}$ sum of Gaussian tensors, each possessing QTT-rank of order $\sqrt{|\log\varepsilon|}$. Now we find for the canonical ranks $rank_C$,

$$rank(\mathbf{H}_{\mu\nu}) \le rank(\mathbf{P}_{\mathcal{N}})rank(\mathbf{G}_{\mu\nu}) = rank(\mathbf{P}_{\mathcal{N}}) = O(|\log\varepsilon|\log n),$$

while the QTT-rank of the canonical vectors in $\mathbf{H}_{\mu\nu}$ is bounded by the multiple of $rank(\mathbf{P}_{\mathcal{N}})$ and $\sqrt{|\log\varepsilon|}$. The latter is due to the rank-multiplicativity in the Hadamard product of tensors. Now the result follows. ∎

Figure 3.1 illustrates the QTT-rank distribution for the precomputed canonical tensors $\mathbf{G}_{\mu\nu}$ and $\mathbf{H}_{\mu\nu}$ (CH_4 molecule) with grid-size $n = 8192$, $N_b = 55$, and $N = 55^2$ (cf. Lemmata 3.4 and 3.5). One observes the uniform bound on QTT-rank for any combination of μ,ν. We also found the almost constant behavior of the QTT ranks along all virtual dimensions in quantized representation of the respective n-vectors.

Now we summarize the complexity issues in computation the Newton potential of the product basis, $\mathbf{H}_{\mu\nu}$, under assumption $rank(\mathbf{G}_{\mu\nu}) = 1$.

Remark 3.6 *For a fixed accuracy $\varepsilon > 0$, the set of tensors $\mathbf{H}_{\mu\nu}$ ($\mu,\nu = 1,...,N_b$) can be precomputed in the canonical tensor format at the expense $O(R_{\mathcal{N}}N_b^2 n\log n)$. Tensors $\mathbf{G}_{\mu\nu}$ and $\mathbf{H}_{\mu\nu}$ can be stored in C-QTT format by $O(N_b^2|\log\varepsilon|\log n)$ and $O(N_b^2|\log\varepsilon|^2\log^2 n)$ reals, respectively.*

Proof. Follows by (3.4) and Lemmas 3.4 and 3.5. ∎

Given tensors $\mathbf{G}$ and $\mathbf{H}$ in the canonical-QTT format, each entry in TEI tensor $\mathbf{B}$ can be calculated with $O(\log n)$ complexity. However, tensor-structured convolution step $\mathbf{H} =$

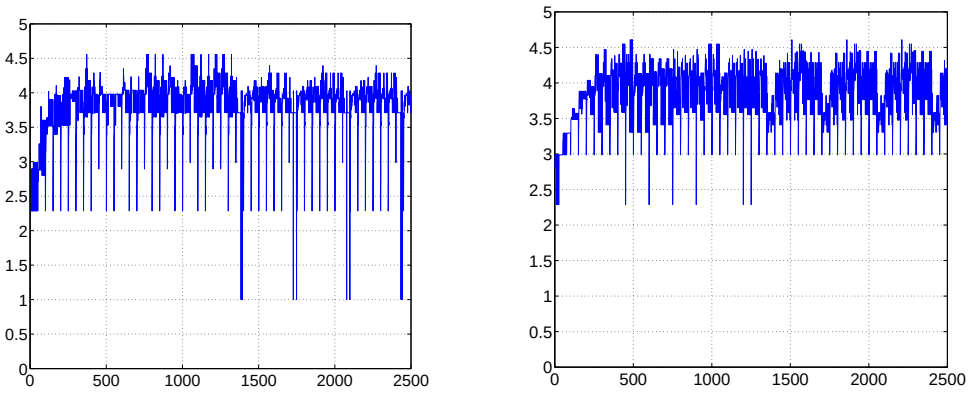

Figure 3.1: CH_4: average QTT ranks of product basis functions, $\mathbf{G}_{\mu\nu}$, (left) and their Newton potential, $\mathbf{H}_{\mu\nu}$, (right), $\varepsilon = 10^{-6}$, $\mu, \nu = 1, ..., N_b = 55$, $n = 8192$.

$\mathbf{P}_{\mathcal{N}} *_{n\otimes 3} \mathbf{G}$ is performed as in (2.5) based on 1D FFT, since in the considered range of grid-size n the 1D QTT convolution is less efficient than 1D FFT. Thus, this step amounts to N_b^2 tensor-product convolutions, each at the expense $O(R_{\mathcal{N}} n \log n)$, which can be rather large for the required n of the order 10^5, and N_b of the order of several hundred. Moreover, since the matrix representation of all scalar products in (3.3) in the QTT format remains an open question, the Matlab implementation of this scheme may be rather slow.

In the next section, we show that the number of convolutions in TEI calculations can be reduced to the order $O(N_b)$ since the product basis set appears to be highly redundant.

3.3 Redundancy-free factorization of the TEI matrix B

Here we introduce an alternative solution of the TEI problem by means of factorized representation in the redundancy free modified product basis. It serves to minimize the number of convolution products in (3.3) using the reduced HOSVD (RHOSVD) introduced in [22] for tensor-rank optimization in the canonical format. The RHOSVD-type factorization applied to the 3D tensor $\mathbf{G}$ allows us to represent it in a "squeezed", only entanglement-based form.

Again, without loss of generality, we assume that $rank(\mathbf{G}_k) = 1$, $(k = 1, ..., N_b)$ for the given basis functions, $\mathbf{G}_k = G_k^{(1)} \otimes G_k^{(2)} \otimes G_k^{(3)} \in \mathbb{R}^{n\times n\times n}$.

Letting $\mathbf{G}^{(\ell)} = \left[\mathbf{G}_\mu^{(\ell)} \odot \mathbf{G}_\nu^{(\ell)}\right]_{1\le\mu,\nu\le N_b} \in \mathbb{R}^{n\times N_b\times N_b}$ be the side tensor, define the respective unfolding matrices

$$G^{(\ell)} = \mathrm{mat}(\mathbf{G}^{(\ell)}) \in \mathbb{R}^{n\times N_b^2}, \quad \ell = 1, 2, 3.$$

We denote by $P^{(\ell)} \in \mathbb{R}^{n\times R_{\mathcal{N}}}$ the factor matrices in the rank-$R_{\mathcal{N}}$ canonical tensor $\mathbf{P}_{\mathcal{N}} \in \mathbb{R}^{n\times n\times n}$. Introduce the ε-truncated SVD-based left-orthogonal decomposition of $G^{(\ell)}$, $G^{(\ell)} \cong U^{(\ell)} V^{(\ell)T}$, $\ell = 1, 2, 3$, with $n\times R_\ell$ and $N\times R_\ell$ matrices $U^{(\ell)}$ (orthogonal) and $V^{(\ell)}$, respectively, where $N = N_b^2$, and where $U^{(\ell)}, V^{(\ell)}$ will be called left and right redundancy-free (RF) basis sets, and denote $R_G = \max R_\ell$.

Lemma 3.7 *Given $\varepsilon > 0$, the redundancy-free factorized ε-approximation to the matrix B,*

$$B \cong B_\varepsilon := \sum_{k=1}^{R_\mathcal{N}} \odot_{\ell=1}^{3} V^{(\ell)} M_k^{(\ell)} V^{(\ell)^T}, \tag{3.5}$$

where $V^{(\ell)}$ is the corresponding right RF basis and

$$M_k^{(\ell)} = U^{(\ell)^T}(P_k^{(\ell)} *_n U^{(\ell)}) \in \mathbb{R}^{R_\ell \times R_\ell}, \quad k = 1, ..., R_\mathcal{N}, \tag{3.6}$$

stands for the Galerkin convolution matrix on the left RF basis, $U^{(\ell)}$, $\ell = 1, 2, 3$, exhibits the following properties:

(A) The storage demand for factorizations (3.5) and (3.6) is estimated by $R_\mathcal{N} \sum_{\ell=1}^{3} R_\ell^2 + N_b^2 \sum_{\ell=1}^{3} R_\ell$ and $O((R_G + R_\mathcal{N})n)$, respectively. The numerical complexity of the ε-truncated representation (3.6) is bounded by $O(R_\mathcal{N} R_G^2 n + R_G R_\mathcal{N} n \log n)$.

(B) The ε-rank of the matrix B_ε admits the following upper bound,

$$rank(B_\varepsilon) \le \min\{N_b^2, R_\mathcal{N} \prod_{\ell=1}^{3} R_\ell\}. \tag{3.7}$$

(C) Let $A_\ell(k) = G^{(\ell)} P_k^{(\ell)} *_n G^{(\ell)}$, then the error estimate in the Frobenius norm holds,

$$\|B - B_\varepsilon\|_F \le 6\varepsilon \max_\ell \| G^{(\ell)} \|_F \sum_{k=1}^{R_\mathcal{N}} \max_\ell \|A_\ell(k)\|_F^2 \| P_k^{(\ell)} \|_F . \tag{3.8}$$

(D) Assume that QTT ranks of the column vectors in $P_k^{(\ell)} *_n U^{(\ell)}$ and $U^{(\ell)}$ are small. Then the QTT representation of tensor factors in (3.6) amounts to $O(R_G R_\mathcal{N} \log n)$ reals. The QTT-complexity to compute matrices $M_k^{(\ell)}$, $k = 1, ..., R_\mathcal{N}$, is estimated by $O(R_\mathcal{N} R_G^2 \log n)$.

Proof. (A) Using the Galerkin-type representation of the TEI tensor $\mathbf{B}$ as in (3.3), we obtain

$$B = \mathrm{mat}(\mathbf{B}) = \sum_{k=1}^{R_\mathcal{N}} \odot_{\ell=1}^{3} G^{(\ell)^T} \left[P_k^{(\ell)} *_n G^{(\ell)} \right],$$

where $\odot$ denotes the Hadamard product of matrices. Plugging the truncated SVD factorization of $G^{(\ell)}$ in the right-hand side above leads to the desired representation,

$$\begin{aligned} B_\varepsilon &= \sum_{k=1}^{R_\mathcal{N}} \odot_{\ell=1}^{3} V^{(\ell)} U^{(\ell)^T} \left[P_k^{(\ell)} *_n (U^{(\ell)} V^{(\ell)^T}) \right] \\ &= \sum_{k=1}^{R_\mathcal{N}} \odot_{\ell=1}^{3} V^{(\ell)} \left[U^{\ell)^T} (P_k^{(\ell)} *_n U^{(\ell)}) \right] V^{(\ell)^T} \\ &= \sum_{k=1}^{R_\mathcal{N}} \odot_{\ell=1}^{3} V^{(\ell)} M_k^{(\ell)} V^{(\ell)^T}. \end{aligned} \tag{3.9}$$

The storage cost for the above RHOSVD-type factorization (3.9) to the $N_b^2 \times N_b^2$ matrix B is bounded by $R_{\mathcal{N}} \sum_{\ell=1}^3 R_\ell^2 + N_b^2 \sum_{\ell=1}^3 R_\ell$ being independent on the grid-size n.

The computational complexity at this step is dominated by the cost of reduced Cholesky algorithm to compute truncated SVD of matrices $G^{(\ell)}$, that is $O(R_G(N_b^2 + n))$, and by the total cost of convolution products in (3.6), $O(R_{\mathcal{N}} R_G n \log n)$.

(B) Using the rank properties of Hadamard product of matrices, it is easy to see that (3.9) implies the direct ε-rank estimate for the matrix B_ε, as in (3.7), where the rank parameters R_ℓ characterize the entanglement of a molecule.

(C) The error bound can be derived along the line of [22], Theorem 2.5 (d), related to the RHOSVD error analysis. In fact, the approximation error can be represented by

$$B - B_\varepsilon = \sum_{k=1}^{R_{\mathcal{N}}} \left(\odot_{\ell=1}^3 G^{(\ell)} P_k^{(\ell)} *_n G^{(\ell)} - \odot_{\ell=1}^3 V^{(\ell)} U^{(\ell)^T} P_k^{(\ell)} *_n U^{(\ell)} V^{(\ell)^T} \right).$$

Denote $\widetilde{A}_\ell(k) = V^{(\ell)} U^{(\ell)^T} P_k^{(\ell)} *_n U^{(\ell)} V^{(\ell)^T}$, then for each fixed $k = 1, \dots R_{\mathcal{N}}$, we have

$$\|A_\ell - \widetilde{A}_\ell\| \le 2\varepsilon \|P_k^{(\ell)}\| \|G^{(\ell)}\| \tag{3.10}$$

since the stability in the Frobenius norm $\|U^{(\ell)} V^{(\ell)^T}\| \le \|G^{(\ell)}\|$ holds. Now, for fixed k, we obtain

$$\begin{aligned} A_1 \odot A_2 \odot A_3 - \widetilde{A}_1 \odot \widetilde{A}_2 \odot \widetilde{A}_3 = {} & A_1 \odot A_2 \odot A_3 - \widetilde{A}_1 \odot A_2 \odot A_3 \\ & + \widetilde{A}_1 \odot A_2 \odot A_3 - \widetilde{A}_1 \odot \widetilde{A}_2 \odot A_3 \\ & + \widetilde{A}_1 \odot \widetilde{A}_2 \odot A_3 - \widetilde{A}_1 \odot \widetilde{A}_2 \odot \widetilde{A}_3. \end{aligned}$$

Summing up the above representation over $k = 1, \dots R_{\mathcal{N}}$, and taking into account (3.10), we arrive at the bound

$$\|B - B_\varepsilon\|_F \le 6\varepsilon \max_\ell \|G^{(\ell)}\|_F \sum_{k=1}^{R_{\mathcal{N}}} \max_\ell \|A_\ell(k)\|_F^2 \|P_k^{(\ell)}\|_F, \tag{3.11}$$

which proves the result.

(D) The complexity bound is the direct consequence of assumptions on the QTT-ranks (see numerics below). ∎

Proof of Lemma 3.7 is constructive and outlines the way to an efficient implementation of (3.5), (3.6). Some numerical results on performance of the corresponding black-box algorithm are shown in Sections 3.4 and 4.2. Here the algebraically optimized separation ranks R_ℓ are only determined by the entanglement properties of a molecule, while the number $N - R_G$ indicates the measure of redundancy in the product basis set. In numerical experiments we observe $R_\ell \le N_b$ and $R_\ell \ll n$ for large n.

Figure 3.2, left, represents the ε-rank R_ℓ, $\ell = 1, 2, 3$, and R_B, computed on the examples of some compact molecules with $\varepsilon = 10^{-6}$. We observe that the Cholesky rank of B, R_B, is a multiple of N_b with a factor ~ 6, see also Fig. 3.3. Remarkably, the RHOSVD separation ranks $R_\ell \le N_b$ remain to be very weakly dependent on N_b, but primarily depend on the topology of a molecule.

Molecules	NH_3	H_2O_2	N_2H_4	C_2H_5OH
$N_b; N_{orb}$	48; 5	68; 9	82; 9	123; 13
Av. QTT rank of $U^{(1)}$	7.3	7.9	7.5	7.6
Av. QTT rank of $V^{(1)}$	15	21	24	37
(Av. QTT rank of $V^{(1)}$)/N_{orb}	3	2.3	2.6	2.85

Table 3.1: Average QTT ε-ranks of $U^{(1)}$ and $V^{(1)}$ in $G^{(1)}$-factorization, $\varepsilon = 10^{-6}$.

Figure 3.2, right, provides average QTT ranks of column-vectors in $U^{(1)} \in \mathbb{R}^{n\times R_1}$ for NH_3, H_2O_2, N_2H_4 and C_2H_5OH molecules. Again, surprisingly, the rank portraits appear to be nearly the same for different molecules, and the average rank over all indices $m = 1, ..., R_1$ is a small constant at about $r_0 \backsimeq 7$. The more detailed results are listed in Table 3.1.

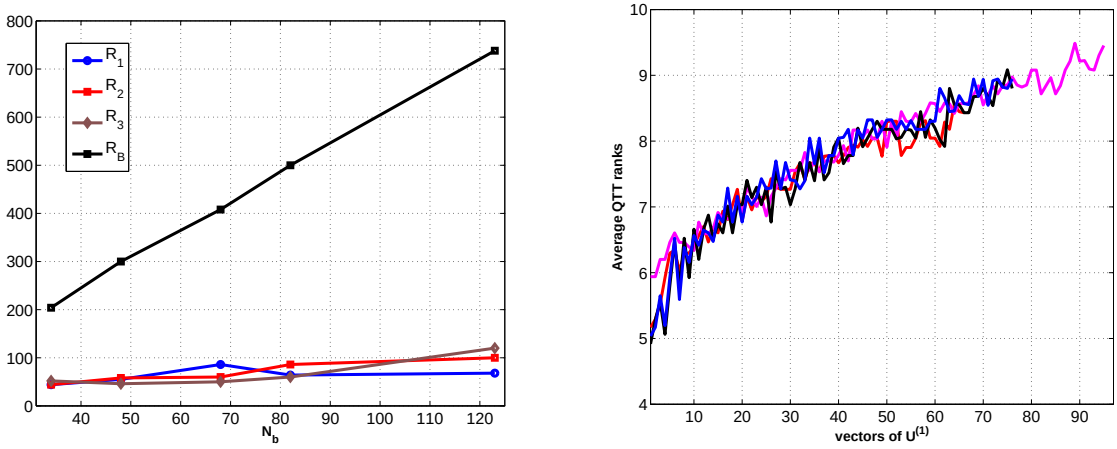

Figure 3.2: Left: ε-ranks R_ℓ and R_B for HF, NH_3, H_2O_2, N_2H_4 and C_2H_5OH molecules, with the number of basis functions $N_b = 34$, 48, 68, 82 and 123, respectively. Right: Average QTT ε-ranks of column-vectors in $U^{(1)} \in \mathbb{R}^{n\times R_\ell}$ for NH_3, H_2O_2, N_2H_4 and C_2H_5OH molecules, $\varepsilon = 10^{-6}$.

The a priori rank estimate (3.7) looks too pessimistic compared with the results of numerical experiments. However, in the case of flattened or extended molecules (some of the directional ranks are small) this estimate provides a much sharper bound.

The RHOSVD factorization (3.5), (3.6) is a reminiscent of the exact Galerkin representation (3.3) in the right RF basis, while matrices $M_k^{(\ell)}$ play the role of "directional" Galerkin projections of the Newton kernel onto the left RF basis. This factorization can be applied directly to fast calculation of the reduced Cholesky decomposition of the matrix B considered in the next section.

Finally, we point out that our RHOSVD-type factorization can be viewed as the algebraic tensor-structured counterpart of the *density fitting scheme* commonly used in quantum chemistry [1]. The robust error control in the proposed basis optimization approach is based on purely algebraic SVD-like procedure that allows to eliminate the redundancy in the product basis set up to given precision $\varepsilon > 0$.

3.4 Low-rank Cholesky decomposition of the TEI matrix B

In this section, we present the economical truncated Cholesky decomposition scheme of complexity $O(N_b^3)$ providing the $O(N_b)$-rank factorization of the TEI matrix B. Then we describe how the complexity can be reduced to $O(N_{orb}^2 N_b)$ using the quantized representation of the Cholesky vectors. The Cholesky scheme requires only $O(N_b)$ adaptively chosen columns in B calculated on-line using the off-line precomputed data either in the form of one-electron integral tensor $\mathbf{H}$, or the redundancy-free factorization (3.5).

We denote the long indexes in the $N \times N$ ($N = N_b^2$) matrix unfolding B by

$$i = vec(\mu, \nu) := (\mu - 1)N_b + \nu, \quad j = vec(\kappa, \lambda), \quad i, j \in I_N := \{1, ..., N\}.$$

Lemma 3.8 *The unfolding matrix B is symmetric and positive semidefinite.*

Proof. The symmetry is enforced by the definition (see Lemma 3.2). The positive semidefiniteness follows from the observation that the matrix B can be viewed as the Galerkin matrix $\langle -\Delta^{-1} u_i, u_j \rangle$, $i, j \in I_N$, in the finite product basis set $\{u_i\} = \{g_\mu g_\nu\}$, where Δ^{-1} is the inverse of the self-adjoint and positive definite in $H^1(\mathbb{R}^3)$ Laplacian operator subject to the homogeneous Dirichlet boundary conditions at $x \to \infty$. ∎

We consider the ε-truncated Cholesky factorization of $B \approx B_\varepsilon = LL^T$, where

$$\|B - LL^T\| \le C\varepsilon, \quad L \in \mathbb{R}^{N \times R_B}.$$

Based on the previous observation (see Introduction), we will postulate rather general ε-rank estimate (in electronic structure calculations this conventional fact traces back to [5]), see numerics on Fig. 3.3.

Conjecture 3.9 *Fixed truncation error $\varepsilon > 0$, for the Gaussian-type AO basis functions there holds, $R_B = rank(LL^T) \le CN_b$, where the constant $C > 0$ is independent of N_b.*

Clearly, the fastest version of the numerical Cholesky decomposition is possible in the case of given full TEI tensor $\mathbf{B}$. In this case the CPU time for the Cholesky decomposition becomes negligible compared with those to compute the TEI tensor $\mathbf{B}$. However, the practical use of algorithm is limited to the small basis sets because of the large storage requirements, N_b^4.

In what follows, we describe the two approaches to compute the truncated Cholesky decomposition with reduced storage demands based on different types of precomputed input:

(A) one-electron integrals tensor $\mathbf{H}$ (see §3.1) or

(B) the redundancy-free factorization of B in form (3.5).

In case (A), we propose the optimized two-step approximation method to compute the ε-truncated Cholesky factorization that operates on the input of the "one-electron" tensors $\mathbf{G}_{\mu\nu}$ and $\mathbf{H}_{\mu\nu}$, both represented in the mixed canonical-QTT data format.

The complexity optimization is based on the idea to recognize the finer data structure by quantization of long n-vectors in precomputing stage using merely the algebraic algorithms with controlled approximation error via the adaptive choice of separation rank parameters.

Suppose that $n = 2^L$, then the rank-1 canonical tensor $\mathbf{G}_{\mu\nu} = G^{(1)} \otimes G^{(2)} \otimes G^{(3)} \in \mathbb{R}^{n \times n \times n}$, with $G^{(\ell)} \in \mathbb{R}^n$, is mapped to its quantized image in dimension D,

$$\mathbf{G}_{\mu\nu} \mapsto \mathcal{Q}(\mathbf{G}_{\mu\nu}) = \mathcal{Q}(G^{(1)}) \otimes \mathcal{Q}(G^{(2)}) \otimes \mathcal{Q}(G^{(3)}) \in \bigotimes_{\nu=1}^{D} \mathbb{R}^2,$$

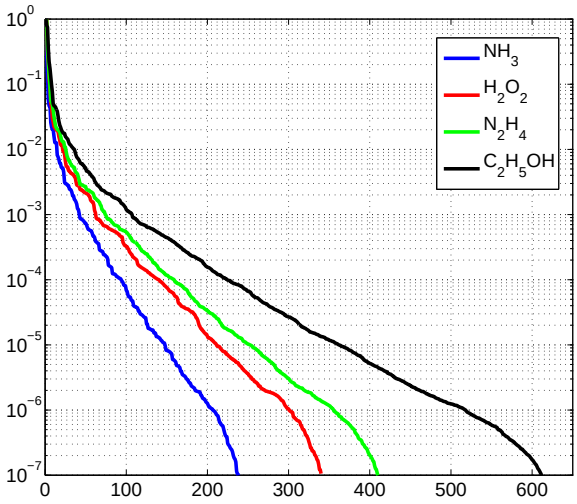

Figure 3.3: Singular values of LL^T for NH_3, H_2O_2, N_2H_4 and C_2H_5OH molecules, with the number N_b of basis functions 48, 68, 82 and 123, respectively.

with $D = 3L$ and maximal QTT-rank r_Q (for GTO basis we have $r_Q \le 4$). Similar quantization transform applies to each rank-1 term representing rank-$R_{\mathcal{N}}$ tensor $\mathbf{H}_{\mu\nu}$.

Quantization of the column N-vectors in the Cholesky factor L applies to their zero extension to the size $N_Q = 2^M \ge N$, that is nearest close to $N = N_b^2$.

Algorithm (A). Given tensors $\mathbf{G}_{\mu\nu}$ and $\mathbf{H}_{\mu\nu}$ ($\mu, \nu \in \{1, ..., N_b\}$) stored in the canonical-QTT format. Compute the optimized ε-truncated Cholesky decomposition in two steps:

(A.I) The Cholesky factorization calculates $R_B = O(N_b)$ required column vectors of size $O(N)$, to be marked by indices $j^* = vec(\kappa^*, \lambda^*)$, by using scalar products in the QTT-canonical format

$$b_{\mu\nu\kappa^*\lambda^*} = \langle \mathbf{G}_{\mu\nu}, \mathbf{H}_{\kappa^*\lambda^*} \rangle \in \mathbb{R}^N.$$

Numerical cost for each N-vector amounts to $3 \cdot 2r_Q^3 \log_2 n N_b^2$ operations.

(A.II) Apply the rank-R_L QTT approximation of dimension $M \approx \log_2 N$ to the column vectors (skeletons) of Cholesky factorization generating the mixed canonical-QTT tensor decomposition of the matrix B.

Fixed $\varepsilon > 0$, our numerical tests indicate the following asymptotic behavior (see Table 3.3),

$$R_L \backsimeq 3N_{orb} \le \frac{3}{10} N_b,$$

that is to be postulated in the following. Hence, the total numerical cost to represent the LL^T Cholesky decomposition is estimated by $O(N_b N_{orb}^2 \log_2 N_b)$.

Table 3.2 provides the CPU times to calculate the convolution tensor $\mathbf{H}$ and then full TEI tensor $\mathbf{B}$ on the example of the NH_3 molecule ($N_b = 48$), $\varepsilon = 10^{-6}$. It can be seen that in the case of small molecules the precomputing step dominates over the direct evaluation of full tensor $\mathbf{B}$ and may quickly become a bottleneck for larger compounds. The CPU-time for ε-Cholesky decomposition is about 1 sec., $\varepsilon = 10^{-5}$.

n^3	2048^3	4096^3	8192^3	16384^3
3D-Conv, **H** (sec)	8	41	112	182
B (sec)	16	19	23	28

Table 3.2: NH_3 molecule: CPU times for **H** and **B** vs. grid-size n^3.

Case (B). In the situation with the precomputed RHOSVD-type factorization (3.5) one can compute the truncated Cholesky decomposition using the column and diagonal elements of matrix B calculated on the fly as in the following

Algorithm (B). Given the redundancy-free factorization (3.5).

(B.I) Implement the Cholesky factorization where the column and diagonal elements in the TEI matrix B are represented by the following tensor operations

$$B(:, j^*) = \sum_{k=1}^{R_{\mathcal{N}}} \odot_{\ell=1}^{3} V^{(\ell)} M_k^{(\ell)} V^{(\ell)}(:, j^*)^T,$$

and

$$B(i, i) = \sum_{k=1}^{R_{\mathcal{N}}} \odot_{\ell=1}^{3} V^{(\ell)}(i, :) M_k^{(\ell)} V^{(\ell)}(:, i)^T,$$

respectively.

(B.II) Repeat Step (A.II) in Algorithm (A).

The factorization (3.5) applied at Step (B.I) essentially reduces the amount of work on the "preprocessing" stage in the limit of large N_b (see Lemma 3.7) since the number of convolutions is now estimated by CN_b (usually, $C \leq 1$) instead of N_b^2.

The cost to compute the column vector at Step (BI) is estimated by $3N_b^2 R_G$. Taking into account that $r_Q \approx 4$ and $R_G \leq N_b$ we conclude that in the range of grid-size $n \leq 2^{14}$, Step (A.I) in Algorithm (A) outperforms Algorithm (B) if $2 \cdot \log_2 n \cdot r_Q^3 = 2 \cdot 14 \cdot 4^3 \leq N_b$, i.e. for large enough $N_b \gtrsim 2000$. The total numerical complexity of Step (BI) in Algorithm (B) is dominated by $O(R_G N_b^3)$.

3.5 Analyzing QTT compression to the Cholesky factor L

This section collects the important observations obtained in numerical experiments. In the QTT analysis of the TEI matrix B for several moderate size compact molecules, we revealed that, with fixed approximation error $\varepsilon > 0$, the average QTT ranks of the Cholesky vectors have the following behavior $r_{QTT} \sim k_{chol} N_{orb}$, with $k_{chol} \leq 3$. From this numerics we make a conclusion that factor $k_{chol} = 3$ is due to the spatial dimensionality of the considered molecular system (or problem) observed for compact compounds and it becomes closer to 2 for more stretched molecules, see Table 3.3 below.

Based on this numerical experiments we formulate our hypothesize:

Hypothesize 3.10 *The structural complexity of the Cholesky factor L of matrix B in the QTT representation is characterized by the rank parameter*

$$R_L \cong 3N_{orb}.$$

Molecule	HF	H_2O	NH_3	H_2O_2	N_2H_4	C_2H_5OH
N_{orb}	5	5	5	9	9	13
r_{QTT}	12	13.6	15	21	24	37
$k_{chol} = r_{QTT}/N_{orb}$	2.4	2.7	3	2.3	2.6	2.85

Table 3.3: Average QTT ranks of the Cholesky vectors vs. N_{orb} for some molecules.

The effective representation complexity of the Cholesky factor $L \in \mathbb{R}^{N\times R_B}$ is estimated by

$$9R_B N_{orb}^2 \ll R_B N_b^2.$$

Assuming that the conventional relation $N_b \approx 10\, N_{orb}$ is fulfilled, we conclude that the reduction factor in the storage size with QTT representation of L is about 10^{-1}.

Similar rank characterizations have been observed by the QTT analysis of $U^{(\ell)}$ and $V^{(\ell)}$ factors in the rank factorization of the initial product bases tensors $\mathbf{G}^{(\ell)}$, $\ell = 1, 2, 3$ (see Table 3.1).

It is interesting to note, that the average QTT ranks of the reduced higher order SVD factors $V^{(\ell)} \in \mathbb{R}^{N_b^2\times R_\ell}$ in the rank factorization of the initial product bases tensors $\mathbf{G}^{(\ell)}$, $\ell = 1, 2, 3$, have almost the same rank scaling, $r_Q(V^{(\ell)}) \le 3N_{orb}$, as a factor $k_{chol} \approx 3$ in the Cholesky decomposition of the matrix B (see Table 3.1). Hence, the QTT representation complexity for the factor $V^{(\ell)}$ in (3.5) can be reduced to

$$10N_{orb}^2 R_G \approx \frac{1}{10} N_b^2 R_G.$$

4 Applications to electronic structure calculations

4.1 TEI tensor in the Hartree-Fock calculations

The numerical treatment of the two-electron integrals (TEI) is the main bottleneck in the fast solution of the Hartree-Fock equation and in DFT calculations for large molecules. The $2N$-electrons Hartree-Fock equation for pairwise L^2-orthogonal electronic orbitals, $\psi_i : \mathbb{R}^3 \to \mathbb{R}$, $\psi_i \in H^1(\mathbb{R}^3)$, reads as

$$\mathcal{F}\psi_i(x) = \lambda_i\, \psi_i(x), \quad \int_{\mathbb{R}^3} \psi_i\psi_j dx = \delta_{ij},\ i, j = 1, ..., N_{orb} \tag{4.1}$$

with $\mathcal{F}$ being the nonlinear Fock operator

$$\mathcal{F} := -\frac{1}{2}\Delta + V_c + V_H + \mathcal{K}.$$

Here the nuclear potential takes the form

$$V_c(x) = -\sum_{\nu=1}^{M} \frac{Z_\nu}{\|x - a_\nu\|}, \quad Z_\nu > 0,\ \ a_\nu \in \mathbb{R}^3,$$

while the Hartree potential $V_H(x)$ and the nonlocal exchange operator $\mathcal{K}$ read as

$$V_H(x) := \rho \star \frac{1}{\|\cdot\|} = \int_{\mathbb{R}^3} \frac{\rho(y)}{\|x-y\|}\, dy, \quad x \in \mathbb{R}^3, \tag{4.2}$$

and

$$(\mathcal{K}\psi)(x) := -\sum_{i=1}^{N_{orb}} \left(\psi\, \psi_i \star \frac{1}{\|\cdot\|}\right) \psi_i(x) = -\frac{1}{2} \int_{\mathbb{R}^3} \frac{\tau(x,y)}{\|x-y\|}\, \psi(y) dy, \tag{4.3}$$

respectively. Conventionally, we use the definitions

$$\tau(x,y) := 2\sum_{i=1}^{N_{orb}} \psi_i(x)\psi_i(y), \quad \rho(x) := \tau(x,x),$$

for the density matrix $\tau(x,y)$, and electron density $\rho(x)$.

Usually, the Hartree-Fock equation is approximated by the standard Galerkin projection of the initial problem (4.1) posed in $H^1(\mathbb{R}^3)$ (see [29] for more details). For a given finite Galerkin basis set $\{g_\mu\}_{1\le\mu\le N_b}$, $g_\mu \in H^1(\mathbb{R}^3)$, the occupied molecular orbitals ψ_i are represented (approximately) as

$$\psi_i = \sum_{\mu=1}^{N_b} C_{\mu i} g_\mu, \quad i = 1, ..., N_{orb}. \tag{4.4}$$

To derive an equation for the unknown coefficients matrix $C = \{C_{\mu i}\} \in \mathbb{R}^{N_b \times N_{orb}}$, first, we introduce the mass (overlap) matrix $S = \{S_{\mu\nu}\}_{1\le\mu,\nu\le N_b}$, given by

$$S_{\mu\nu} = \int_{\mathbb{R}^3} g_\mu g_\nu dx,$$

and the stiffness matrix $H = \{h_{\mu\nu}\}$ of the core Hamiltonian $\mathcal{H} = -\frac{1}{2}\Delta + V_c$ (the single-electron integrals),

$$h_{\mu\nu} = \frac{1}{2}\int_{\mathbb{R}^3} \nabla g_\mu \cdot \nabla g_\nu dx + \int_{\mathbb{R}^3} V_c(x) g_\mu g_\nu dx, \quad 1 \le \mu, \nu \le N_b.$$

The core Hamiltonian matrix H can be precomputed in $O(N_b^2)$ operations, see [21] for the detailed description of the grid-based approach.

In computational quantum chemistry the nonlinear terms representing the Galerkin approximation to the Hartree and exchange operators are calculated traditionally by using the two-electron integrals tensor $\mathbf{B} = [b_{\mu\nu\kappa\lambda}]$ as defined in (1.1), that initially has the computational and storage complexity of order $O(N_b^4)$.

Introducing the $N_b \times N_b$ matrices $J(D)$ and $K(D)$,

$$J(D)_{\mu\nu} = \sum_{\kappa,\lambda=1}^{N_b} b_{\mu\nu,\kappa\lambda} D_{\kappa\lambda}, \quad K(D)_{\mu\nu} = -\frac{1}{2}\sum_{\kappa,\lambda=1}^{N_b} b_{\mu\lambda,\nu\kappa} D_{\kappa\lambda}, \tag{4.5}$$

where $D = 2CC^T \in \mathbb{R}^{N_b \times N_b}$ is the low-rank symmetric density matrix, such that

$$rank(D) = N_{orb} \ll N_b,$$

one then represents the complete Fock matrix F by

$$F(D) = H + J(D) + K(D). \tag{4.6}$$

The resultant Galerkin system of nonlinear equations for the coefficients matrix $C \in \mathbb{R}^{N_b \times N_{orb}}$, and the respective eigenvalues Λ, reads as

$$\begin{aligned} F(D)C &= SC\Lambda, \quad \Lambda = diag(\lambda_1, ..., \lambda_{N_b}), \\ C^T SC &= I_N, \end{aligned} \tag{4.7}$$

where the second equation represents the orthogonality constraints $\int_{\mathbb{R}^3} \psi_i \psi_j dx = \delta_{ij}$. Here I_N denotes the $N_b \times N_b$ identity matrix.

In the standard quantum chemical implementations based on the analytically precomputed set of two-electron 3D convolution integrals, the numerically confirmed rank bound $rank(B) \le C_B N_b$ ($C_B \sim 10$), allows to reduce the complexity of building up the Fock matrix F to $O(N_b^3)$, which is by far dominated by computational cost for the exchange term $K(D)$. In the following we will show how this cost can be reduced further using certain low-rank structures in matrices B and D.

4.2 The Hartree-Fock calculus using tensor structure in B

In this Section we consider in more detail the multilinear algebraic calculation of the Coulomb and exchange matrices in the Fock operator. One can use the canonical-QTT structure of the target tensor and low-rank structure of the matrix D to evaluate and represent matrices $J(D)$ and $K(D)$ efficiently.

Remark 4.1 *(The Coulomb matrix). Precomputed tensors $\mathbf{G}_{\mu\nu}, \mathbf{H}_{\kappa\lambda}$, in view of (3.1), we have*

$$J(D)_{\mu\nu} = \sum_{\kappa,\lambda=1}^{N_b} b_{\mu\nu,\kappa\lambda} D_{\kappa\lambda} = \sum_{\kappa,\lambda=1}^{N_b} \langle \mathbf{G}_{\mu\nu}, \mathbf{H}_{\kappa\lambda} \rangle D_{\kappa\lambda}. \tag{4.8}$$

Vectorizing matrices $\overline{J} = vec(J), \overline{D} = vec(D)$, we arrive at the simple matrix representation,

$$\overline{J} = B\overline{D} \approx L(L^T \overline{D}), \tag{4.9}$$

which can be easily evaluated taking into account the rank structure of B as well as the QTT-structure in vectors $\overline{D}$ and in the column vectors of L.

The straightforward calculation by (4.9) amounts to $O(R_B N_b^2)$ operations where R_B is the ε-rank of B. Our analysis indicates that imposing the QTT-structure of the matrix L may reduce this cost to $O(R_B N_{orb})$.

Remark 4.2 *(The HF exchange). Tensor evaluation of the exchange matrix $K(D)$ is much more involved since in this case it reduces to a summation over permuted indices,*

$$K(D)_{\mu\nu} = -\frac{1}{2} \sum_{\kappa,\lambda=1}^{N_b} b_{\mu\lambda,\nu\kappa} D_{\kappa\lambda} = -\frac{1}{2} \sum_{\kappa,\lambda=1}^{N_b} \langle \mathbf{G}_{\mu\lambda}, \mathbf{H}_{\nu\kappa} \rangle D_{\kappa\lambda}. \tag{4.10}$$

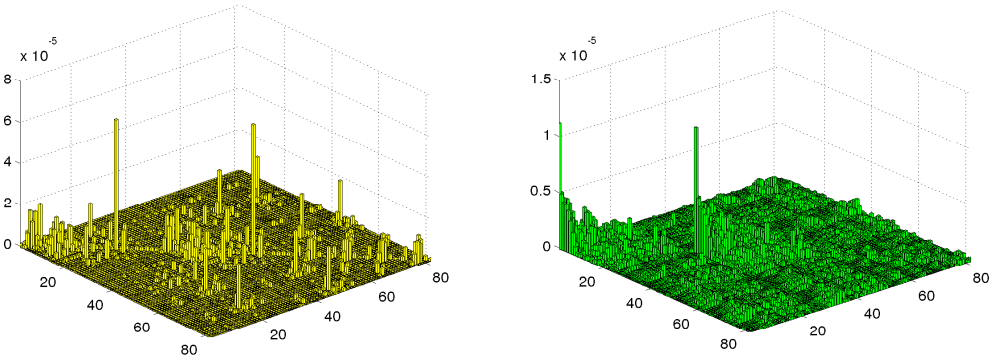

Figure 4.1: N_2H_4 molecule: error in the Coulomb (left) and exchange (right) matrices, using the Richardson extrapolation on $n^{\otimes 3}$ grids with $n = 8192$ and $n = 16384$.

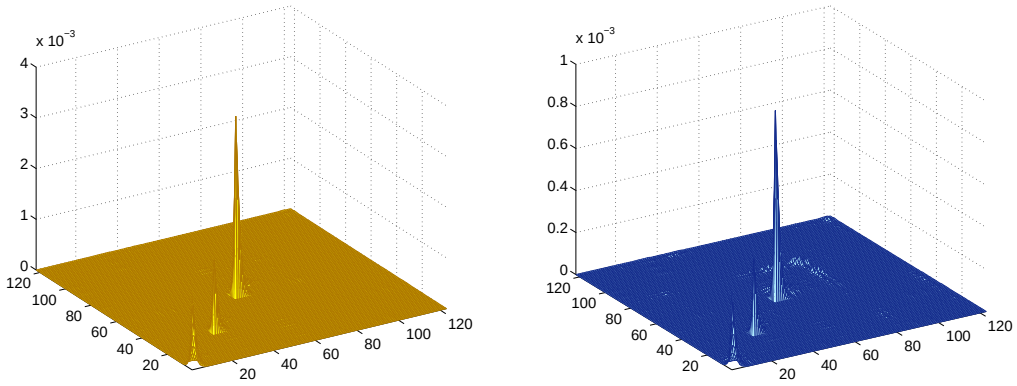

Figure 4.2: C_2H_5OH molecule: accuracy of the exchange matrices on $n^{\otimes 3}$ grids with $n = 8192$ and $n = 16384$ (the decay ratio $1 : 4$ is well suited for the Richardson extrapolation).

Introducing the permuted tensor $\widetilde{\mathbf{B}} = permute(\mathbf{B}, [2,3,1,4])$*, and the respective accompanying matrix* $\widetilde{B} = mat(\widetilde{\mathbf{B}})$*, we then obtain*

$$vec(K) = \overline{K} = \widetilde{B}\overline{D}. \tag{4.11}$$

The calculation by (4.11) amounts to $O(R_B N_b^3)$ *operations. However, using the rank-*N_{orb} *decomposition of* $D = 2CC^T$ *allows to reduce the cost to* $O(R_B N_{orb} N_b^2)$*, by the representation,*

$$K(D)_{\mu\nu} = -\sum_{i=1}^{N_{orb}} (\sum L_{\mu\lambda} C_{\lambda i})(\sum L_{\kappa\nu} C_{\kappa i})^T,$$

where $L_{\mu\nu} = reshape(L, [N_b, N_b, R_B]) \in \mathbb{R}^{N_b \times N_b \times R_B}$ *is the* $N_b \times N_b \times R_B$*-folding of the Cholesky factor* L*.*

In this work the numerical illustrations are performed using the grid representation of Gaussian AO bases merely to have means to compare the accuracy of the resulting Galerkin Fock matrix with a standard output from the MOLPRO package [40], where the entries of TEIs are computed analytically.

Figure 4.1 represents the error portrait corresponding to the Coulomb and exchange matrices computed via rank-R_B LL^T approximation to B as above, in the case of N_2N_4 molecule. Figure 4.2 shows the absolute error in the exchange matrix of C_2H_5OH molecule calculated on $n^{\otimes 3}$ grids with $n = 8192$ and $n = 16384$. The numerical error scales quadratically in the grid size, $O(h^2)$, and can be improved to $O(h^3)$ by the Richardson extrapolation. The observed decay ratio $1:4$ indicates the applicability of Richardson extrapolation to the results on a pair of diadically refined grids.

4.3 MP2 perturbation theory

The various degrees Møller-Plesset perturbation theory (in particular, second-order MP2 model) significantly improves the HF correlation energy and other molecular characterizations in the case of large basis sets [41]. However, the numerical payoff scales as N_b^5. Our tensor-structured representations of the matrix B allows to reduce this cost by some orders of magnitude. Here we sketch the main idea.

First, one has to transform the TEI tensor $\mathbf{B} = [b_{\mu\nu\kappa\lambda}]$, from the initial AO basis set to the MO-basis represented by (4.4),

$$\mathbf{V} = [v_{iajb}]: \quad v_{iajb} = \sum_{\mu,\nu,\lambda,\sigma=1}^{N_b} C_{\mu i} C_{\nu a} C_{\lambda j} C_{\sigma b} b_{\mu\nu\lambda\sigma}, \quad i,a,j,b \in \{1,...,N_b\},$$

that makes the dominating impact to the overall numerical cost of order $O(N_b^5)$. Given the tensor $\mathbf{V} = [v_{iajb}]$, then the second order MP2 perturbation to the HF energy is calculated by

$$E_{MP2} = -\sum_{a,b \in I_{vir}} \sum_{i,j \in I_{occ}} \frac{v_{iajb}(2v_{iajb} - v_{ibja})}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j},$$

where $I_{occ} := \{1,...,N_{orb}\}$, $I_{vir} := \{N_{orb}+1,...,N_b\}$, and ε_k, $k = 1,...,N_b$, represent the HF eigenvalues.

It can be shown that the rank-$O(N_b)$ approximation to the TEI matrix B and to the matrix unfolding of $\mathbf{V}$, $V = [v_{ia;jb}]$, combined with the low-rank approximation to the unfolding matrix E of a tensor

$$\mathbf{E} = [\frac{1}{\varepsilon_a + \varepsilon_b - \varepsilon_i - \varepsilon_j}], \quad i, a, j, b \in \{1, ..., N_b\}, \tag{4.12}$$

reduces the total cost by the order of magnitude (we suppose that the so-called *homo-lumo* gap is larger than some fixed constant $\delta > 0$). In this aspect, the rank-$|\log \varepsilon \log \delta|$ approximation to the unfolding matrix $E = [e_{k\ell,mn}]$ follows by the well-known sinc-approximation to the Hilbert matrix $\{\frac{1}{i+j}\}$, [12].

Further reduction of the numerical complexity can be based on more specific properties of the matrix unfolding V when using a physical insight to the problem.

4.4 Multidimensional Hamiltonian in FCI models

In the formulation of second quantization, the Hamiltonian operator for the high-dimensional electronic Schrödinger equation

$$\mathcal{H}\Psi = E\Psi$$

is given by

$$\mathcal{H} = \sum_{pq} h_{pq} a_q^+ a_p + \frac{1}{2} \sum_{pq,rs} v_{qspr} a_r^+ a_s^+ a_p^- a_q^-,$$

where the single-electron integrals correspond to the core Hamiltonian in HF operator (see §4.1), and v_{qspr} represents the TEI in the MO basis (the so-called canonical molecular orbitals). The symbols a^+ and a^- denote creation and annihilation operators of second quantization. The details on electronic structure calculations via FCI models, particularly based on DMRG optimization over MPS-type tensor manifolds, can be found in the seminal paper [42].

5 Conclusions

In this paper, we present a new grid-based tensor-structured method for the efficient calculation of the TEI tensor in a general low-rank basis. The approach is based on the tensor-product numerical integration, redundancy-free matrix factorizations on the optimized product-basis, and, finally, the pivoted Cholesky decomposition of the TEI matrix B accomplished by the QTT-recompression of the column vectors in the Cholesky factor.

Due to the $O(n \log n)$ complexity scaling of the tensor-structured numerical convolution over $n \times n \times n$ Cartesian grids, and, hence, the possibility to imply fine spatial meshes, we achieve high accuracy in the Coulomb and exchange matrices on examples of several organic molecules. The computational error is of the order $O(h^2)$, where h is the mesh-size, and it can be reduced to $O(h^3)$ by using the Richardson extrapolation. Making use of the quantized representation to grid functions allows to reduce the storage complexity to the logarithmic scaling in n, $O(\log n)$.

The main *theoretical results* of the paper are concerned with the QTT rank bounds proven for the GTO basis tensor $\mathbf{G}$ and for its Newtonian convolution $\mathbf{H}$, see Lemmas 3.4

and 3.5, and summarized in Remark 3.6 in the form of rigorous complexity bounds. These results justify the logarithmic complexity scaling in the mesh parameter n of the important precomputing step as in §3.2. We also prove the complexity and error estimates for the redundancy-free factorization of the TEI matrix, presented in Lemma 3.3.

We revealed in *numerical tests* that the separation ranks of the redundancy-free factorization to the matrix B remain to be bounded by N_b, while the ε-rank of the TEI matrix B itself has clear linear scaling in N_b. The structural complexity of the quantized representation to the Cholesky vectors, measured by the QTT-ranks, is close to $3N_{orb}$. The important numerical observation in the paper is that the storage complexity of Cholesky decomposition scales as $O(3^2 N_{orb}^2 N_b)$, reducing at about a factor of 10^{-1} the traditional scaling $O(N_b^3)$, since, conventionally, $N_b \backsimeq 10 N_{orb}$.

The presented method has a good potential for the post-HF models, since the higher dimensionality of tensor data makes the effect of multilinear algebra even more essential.

Acknowledgements. The authors are grateful Dr. habil. Dirk Andre (FU Berlin) and to Dr. Heinz-Juergen Flad (TU Munich) for valuable discussions. We are thankful to Dr. D. Savostianov (Uni. Southampton, UK) for the assistance with implementation of the incomplete Cholesky decomposition.

References

[1] J. Almlöf. *Direct methods in electronic structure theory.* In D. R. Yarkony: Modern Electronic Structure Theory. vol II, World Scientific, Singapore, 1995 pp. 110-151.

[2] P. Y. Ayala and G. E. Scuseria. *Linear scaling second-order Møller-Plesset theory in the atomic orbital basis for large molecular systems.* J. Chem. Phys. v. 110 No. 8, 1999, pp.3660-3671.

[3] M. Barrault, E. Cancès, W. Hager and C. Le Bris. *Multilevel domain decomposition for electronic structure calculations.* J. Comput. Phys. **222**, 2007, 86-109.

[4] M. Bebendorf and S. Rjasanow. *Adaptive low-rank approximation of collocation matrices.* Computing, 70(1):1-24, 2003.

[5] N.H.F. Beebe and J. Linderberg. *Simplifications in the generation and transformation of two-electron integrals in molecular calculations.* Int. Quantum Chem., v 12 7:683-705, 1977.

[6] C. Bertoglio, and B.N. Khoromskij. *Low-rank quadrature-based tensor approximation of the Galerkin projected Newton/Yukawa kernels.* Comp. Phys. Communications, v. 183(4), 904-912 (2012).

[7] I.P. Gavrilyuk, W. Hackbusch, and B.N. Khoromskij. *Tensor-product approximation to elliptic and parabolic solution operators in higher dimensions.* Computing **74** (2005), 131-157.

[8] S. A. Goreinov, E. E. Tyrtyshnikov and N. L. Zamarashkin. *A theory of pseudoskeleton approximations.* Linear Alebra Appl., 261, 1997, 1-21.

[9] S.V. Dolgov, B.N. Khoromskij, and I. Oseledets. *Fast solution of multi-dimensional parabolic problems in the TT/QTT formats with initial application to the Fokker-Planck equation.* SIAM J. on Sci. Comp., 2012 (accepted). Preprint 80/2011, MPI MiS, Leipzig 2011.

[10] S.V. Dolgov, B.N. Khoromskij, and D. Savostianov. *Superfast Fourier transform using QTT approximation.* J. Fourier Anal. Appl., 2012, vol.18, 5, 915-953.

[11] S.V. Dolgov, and B.N. Khoromskij. *Two-level Tucker-TT-QTT format for optimized tensor calculus.* Preprint 19/2012, MPI MiS, Leipzig 2012 (SIMAX, submitted).

[12] W. Hackbusch and B.N. Khoromskij. *Low-rank Kronecker product approximation to multi-dimensional nonlocal operators. Part I. Separable approximation of multi-variate functions;* Computing **76** (2006), 177-202.

[13] H. Harbrecht, M. Peters, and R. Schneider. *On the low-rank approximation by the pivoted Cholesky decomposition.* App. Numer. Math., 62(4), 2012, 428-440.

[14] R.J. Harrison, G.I. Fann, T. Yanai, Z. Gan, and G. Beylkin. *Multiresolution quantum chemistry: Basic theory and initial applications.* J. of Chemical Physics, 121 (23): 11587-11598, 2004.

[15] T. Helgaker, P. Jørgensen and J. Olsen. *Molecular Electronic-Structure Theory.* Wiley, New York, 1999.

[16] N. Higham. *Analysis of the Cholesky decomposition of a semi-definite matrix.* In M.G. Cox and S.J. Hammarling, eds. Reliable Numerical Computations, p. 161-185, Oxford University Press, Oxford, 1990.

[17] S. Holtz, T. Rohwedder, and R. Schneider. *On manifolds of tensors of fixed TT-rank,* Numer. Math., 120(12) pp.701-731, 2012.

[18] V. Kazeev, B.N. Khoromskij, and E.E. Tyrtyshnikov. *Multilevel Toeplitz matrices generated by QTT vectors and convolution with logarithmic complexity.* Preprint 36/2011, MPI MiS, Leipzig 2011, (SISC, submitted).

[19] V. Khoromskaia. *Computation of the Hartree-Fock exchange by the tensor-structured methods.* Comp. Meth. in Applied Math., Vol. 10 , No 2, 204-218, 2010.

[20] V. Khoromskaia, B.N. Khoromskij, and R. Schneider. *QTT representation of the Hartree and exchange operators in electronic structure calculations.* Comp. Meth. in Applied Math., v.11, No. 3, 327-341, 2011.

[21] V. Khoromskaia, D. Andrae, and B.N. Khoromskij. *Fast and accurate 3D tensor calculation of the Fock operator in a general basis.* Comp. Phys. Communications, 183 (2012) 2392-2404.

[22] B.N. Khoromskij and V. Khoromskaia. *Multigrid tensor approximation of function related multidimensional arrays.* SIAM J. on Sci. Comp., **31**(4), 3002-3026, 2009.

[23] B.N. Khoromskij, V. Khoromskaia, and H.-J. Flad. *Numerical solution of the Hartree-Fock equation in multilevel tensor-structured format.* SIAM J. on Sci. Comp., **33**(1), 45-65, 2011.

[24] B.N. Khoromskij. $O(d \log N)$*-Quantics approximation of* N*-*d *tensors in high-dimensional numerical modeling.* J. Constr. Approx., v.34(2), 2011, 257-280.

[25] B.N. Khoromskij. *Fast and accurate tensor approximation of a multivariate convolution with linear scaling in dimension.* J. of Comp. Appl. Math., **234** (2010) 3122-3139.

[26] B.N. Khoromskij. *Tensors-structured numerical methods in scientific computing: survey on recent advances.* Chemometrics and Intellingent Laboratory Systems, 110 (2012),1-19.

[27] T.G. Kolda and B.W. Bader. *Tensor decompositions and applications. SIAM Review*, **51/3**, 2009 455-500.

[28] H. Luo, W. Hackbusch and H.-J. Flad. *Quantum Monte Carlo study of Jastrow perturbation theory. I. Wavefunction optimization.* J. Chem. Phys., 131, 104106, 2009.

[29] C. Le Bris, *Computational chemistry from the perspective of numerical analysis.* Acta Numerica (2005), 363 - 444.

[30] I.V. Oseledets. *Tensor-train decomposition.* SIAM J. Sci. Comp., v. **33**(5), 2295-2317 (2011).

[31] I.V. Oseledets, *Approximation of* $2^d \times 2^d$ *matrices using tensor decomposition.* SIAM J. Matrix Anal. Appl., 31(4):2130-2145, 2010.

[32] I.V. Oseledets, and E.E. Tyrtyshnikov, *Breaking the curse of dimensionality, or how to use SVD in many dimensions.* SIAM J. Sci. Comp., 31 (2009), 3744-3759.

[33] P. Pulay. *Improved SCF convergence acceleration.* J. Comput. Chem. **3**, 556-560 (1982).

[34] G. Rauhut, P. Pulay, H.-J. Werner. *Integral transformation with low-order scaling for large local second-order Mollet-Plesset calculations.* J. Comp. Chem. 19, 1998, pp.1241-1254.

[35] D. V. Savostianov. *Fast revealing of mode ranks of tensor in canonical form.* Numer. Math. Theor. Meth. Appl. v.2 No.4 2009 pp439-444.

[36] N. D. Sidiropoulos. *Generalized Carath'eodory's Uniqueness of harmonic Parametrization to N Dimensions.* IEE Trans. Inform. Theory, 47 (2001) 1687-1690.

[37] E. E. Tyrtyshnikov. *Mosaic-skeleton approximation.* Calcolo, 33, 1996, 47-57.

[38] F. Verstraete, D. Porras, and J.I. Cirac, *DMRG and periodic boundary conditions: A quantum information perspective.* Phys. Rev. Lett., 93(22): 227205, Nov. 2004.

[39] G. Vidal. *Efficient classical simulation of slightly entangled quantum computations.* Phys. Rev. Lett. 91(14), 2003, 147902-1 147902-4.

[40] H.-J. Werner, P.J. Knowles, et al. *MOLPRO, Version 2002.10, A package of ab initio programs for electronic structure calculations.*

[41] H.-J. Werner, F.R. Manby, and P.J. Knowles. *Fast linear scaling second order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations.* J. Chem. Phys., 118:8149-8160, 2003.

[42] S.R. White. *Density-matrix algorithms for quantum renormalization groups.* Phys. Rev. B, v. 48(14), 1993, 10345-10356.

[43] S. Wilson. *Universal basis sets and Cholesky decomposition of the two-electron integral matrix.* Comput. Phys. Commun., 58:71-81, 1990.