Compression of Two-Electron Integrals

Jonathan Fajen

December 2023

Introduction

Modern quantum chemists are increasingly interested in the accurate and routine simulation of nuclear and electronic dynamics extending to the regime of hundreds to thousands of atoms, and timescales up to nanoseconds and beyond. This efficient predictive capability can then be leveraged toward the design of functional materials relevant for, for example, energy storage or photocatalysis. To achieve these simulations, the solution and propagation of the nuclear-electronic Schrodinger equation must be performed many times. The typical timestep used in high-level simulations, in order to achieve energy conservation and time-reversal symmetry, is on the order of half a femtosecond (1 fs = 10^{-15} s). This means that a simulation of 1 nanonsecond of dynamics requires one million timesteps, or one million solutions of the nuclear-electronic Schrodinger equation and its gradients. The solution of the nuclear-electronic Schrodinger equation, even at a gross level of approximation, can be very expensive. Normally, the dynamics of the nuclei and electrons can be decoupled, due to the much larger size and slower speed of nuclei relative to electrons. Under this approximation, the nuclear dynamics can be propagated essentially classically, on a potential energy surface defined by the electrons. This approach results in an exponential improvement in computational cost and complexity. However, the time-independent electronic Schrodinger equation, along with its gradient, must still be solved at each timestep.

Even the most approximate treatment of the electronic structure problem (Hartree-Fock theory) scales as $O(N^4)$, due to the presence of the Two-Electron Integrals (TEIs). HF is a self-consistent field problem where tens of iterations involving the digestion of the TEIs are generally required. Post-HF corrections have even worse scaling and more complicated manipulations of the TEIs are needed. Therefore, to improve the efficiency of electronic structure approaches, we must focus on the TEIs.

In a given one-electron basis, with basis functions $\{\phi_p\}$, the TEI tensor takes the form

$$(pq|rs) \tag{1}$$

And scales with the fourth power of the system size.[1] This also means that for large systems with hundreds or thousands of atoms, even the storage of this tensor can become an issue. Therefore, efforts to compress the size of the

TEI tensor have been widespread in the electronic structure community. The first idea was to simply throw away small values without ever computing them. This is intuitive as we might think that electrons that are very far away from one another in a molecule would not interact very strongly, and so integrals over basis functions that are well-separated can be neglected. This strategy is especially useful at reducing the number of integrals that need to be calculated in very large systems, but fails to totally solve the problem.

One of the earliest and still most popular methods for TEI compression is the Cholesky Decomposition (CD),[2] which asserts a factorization of the integrals as

$$(pq|rs) = \sum_K L_{pq}^K L_{rs}^K \tag{2}$$

Where K is the Cholesky index. Essentially, this routine amounts to finding a set of vectors (called auxilliary basis functions or Cholesky vectors) that span the set of TEIs up to some user-supplied threshold. If the threshold is taken to be zero, then the CD resolves to the full array. An appealing property of the CD is that is can be built up iteratively. That is, you add vectors one at a time until you reach convergence to the desired threshold. This means that large storage requirements can be elided.

Density fitting (DF) is a closely related approach to CD, but uses a tailored auxiliary basis to fit the TEIs, rather than building the auxiliary basis on the fly, as in CD.[3] In fact, a CD is often used to construct an initial auxiliary basis for DF methods.

Project Outline

In this work, we are mainly interested in making connections between the standard approaches in data compression and those popular in the context of electronic structure theory. In particular, we have focused on evaluating the performance of a few lossy compressors on the task of TEI compression. We have also been interested in the similarity between transform coding and the CD/DF approaches.

The key procedure in lossy compression is the quantization of the input data. Given some set of input data X, which is sourced from some continuous distribution, we must choose a finite codebook, C, that will be used to map each symbol X_i from the input data to an element of the codebook, C_j, with the goal of minimizing the distortion between the reconstructed message, $\hat{X}$ and the original data, X. The most naive quantization technique is to simply uniformly discretize the space of input data, without taking account additional structure of the data. This is known as scalar quantization. Vector quantization improves upon this approach by quantizing multiple input symbols at a time, thereby allowing for the incorporation of some correlation in the data. Unfortunately, it is not usually straightforward to determine the optimal codebook and mapping of the data for vector quantization, and we must resort to iterative optimization. In this work, we perform this iterative optimization using the k-means algorithm. Here, we initialize some codebook and assign each element of the input data to an element in the codebook (partitions). Then, we get a new codebook by taking

System	DF Max Integral Diff.	DF Avg. MSE	VQ Max. Integral Diff.	VQ Avg. MSE
H_2O	0.0268	7.12e-8	4.44	0.00039
CH_4	3.41e-5	1.87e-12	3.31	0.00010
C_6H_6	3.41e-5	3.12e-14	3.50	7.48e-6

Table 1: Table presenting the average MSE distortion to the full TEI array and maximum difference integral for both DF and VQ approaches across the set of test systems. Note that the VQ approaches here use a rate of 1 bps, so the compression ratio is much better than DF achieves.

the mean of each partition. We repartition the data with this new codebook, and check for convergence.

Transform coding allows us to exploit correlation in the input data in a more intelligent way. Usually, we perform some sort of linear transformation of the data, using either a set of data-defined basis vectors, such as the eigenvectors of the covariance matrix, or a pre-defined set of basis vectors as in the Discrete Cosine Transform (DCT). One can view the CD as a form of data-defined transform coding, and the DF method as more similar to the DCT approach.

To explore these questions, we selected a small set of toy systems, namely, H_2O, CH_4, and C_6H_6, with a fairly small one-electron basis set (cc-pVDZ). Most of the calculations are performed on water (H_2O) due to its manageable computational cost on a laptop.

First, we compared the performance in terms of reconstructing the full TEI tensor with DF, CD, and vector quantization. We note that we can achieve very large compression ratios with vector quantization, even for small systems, however, the distortion we get at even much larger compression rates is not competitive with DF or CD.

Next, we explored the distortion in terms of MSE of the reconstructed integrals with increasing rate, and with increasing quantization dimension at a given rate. We also performed this study for the DF integrals. In some sense, this is like doing transform coding on the original TEI tensor, particularly when larger rates are used in the final quantization. We also evaluated the distortion using the HF electronic energy. All DF calculations and electronic structure calculations were performed using *pyscf*,[5] an open-source, python-based code for quantum chemistry calculations.

Results

First, we looked at the performance in terms of maximum integral difference and average MSE between the original TEI array and the reconstructed TEI array, using either DF or standard VQ as described above, for the set of three test systems. These results are presented in the Table below. We observe that DF performs far better than VQ, and that compression performance improves with system size. This makes sense, as we recall that many more integrals will be essentially zero as the system gets larger.

For all of the remaining performance tests, we focus on water in the interest of low computational cost. We present the performance in the MSE between the original TEI array and the reconstructed TEI using VQ with increasing

quantization dimension in the first Figure. We observe that we can indeed improve the distortion performance by increasing the quantization dimension, incorporating some correlation between the source data. However, we are not competitive with the standard DF approach.

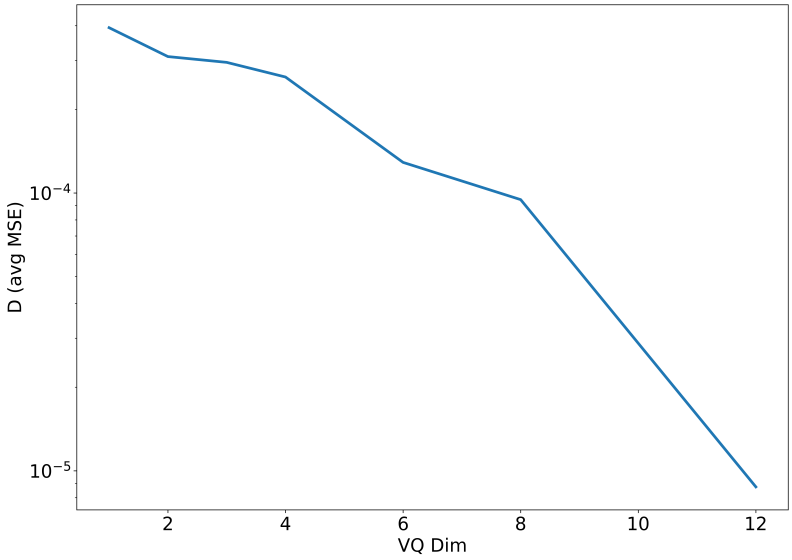

Figure 1: Average MSE between original and reconstructed TEI array using VQ with increasing dimension.

In analogy with the theory of transform coding, we examine the performance of VQ on top of DF. We can think of DF as performing some linear transform into the set of auxilliary basis vectors and then we are further compressing in this theoretically decorrelated space. We present the performance of this approach alongside that of simply Vector Quantizing the raw TEI tensor in the next two Figures. First, we simply evaluate the distortion as the MSE between the original and reconstructed TEI arrays as before. We observe that the DF-then-VQ approach far outperforms standard VQ, achieving MSE distortions that are much closer to the original DF values. This indicates that we are indeed better accounting for source correlation with this approach.

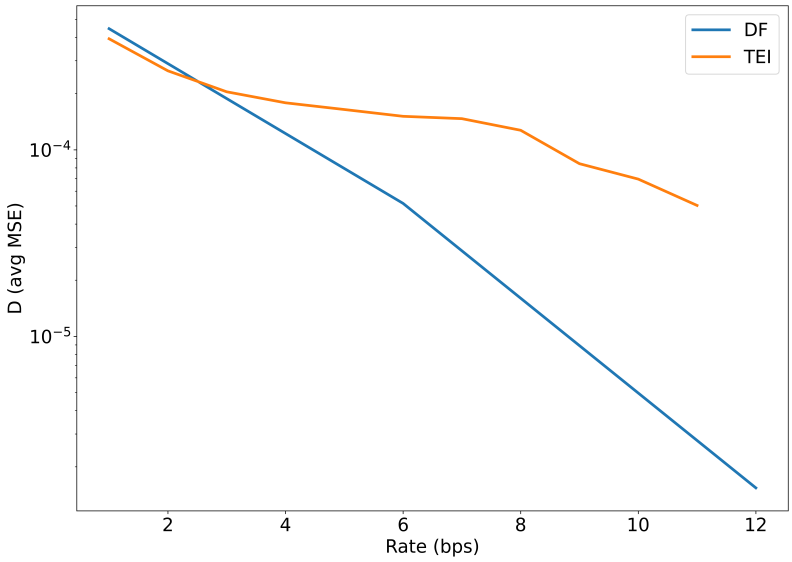

Figure 2: Average MSE in reconstructed full TEI array using VQ on the raw TEIs and using VQ on the DF vectors, L^A_{ia}, each with increasing rate in terms of bits per symbol.

The final performance metric we show is simply calculating the HF energy of water with the reconstructed integrals, both with raw VQ of the TEI array, and with the DF-then-VQ approach. We see that neither is able to get the electronic energy to within 10 Hartree (the atomic energy unit), but that DF-then-VQ always outperforms standard VQ.

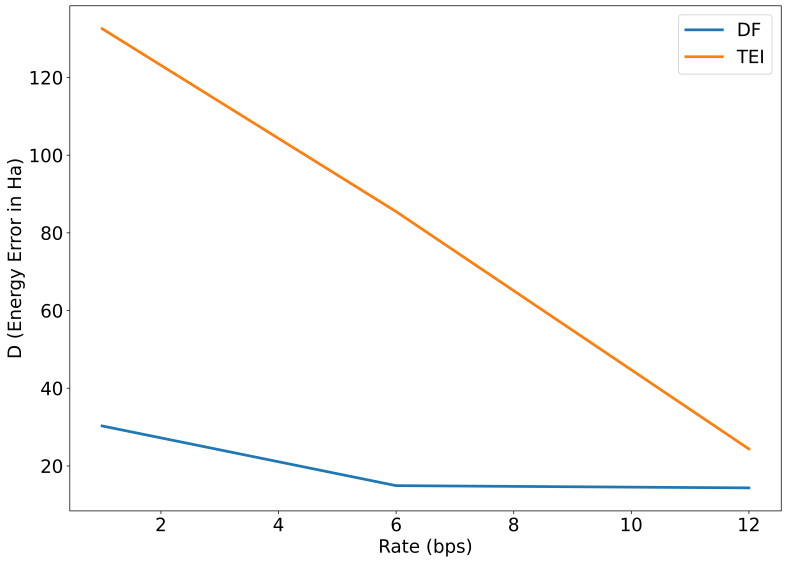

Figure 3: Distortion in terms of electronic energy at the HF level with both VQ on the TEI array and VQ on the DF vectors.

Conclusions and Future Work

The main takeaway of this work is perhaps a very obvious one. Naively using standard approaches from data compression does not perform competitively with the tailored methods from electronic structure theory for compression of the TEI tensor. This finding is of course expected, as the quantizers and encoders we have used are essentially blind to the correlation within the TEI data. For instance, we have not explicitly incorporated any of the symmetries that we ought to expect the TEI tensor to possess. There is an example of a bespoke lossy compressor for TEIs in the literature, which undertook extensive study of the proper quantization scheme and was ultimately able to achieve better compression rates with competitive performance to standard integral compression approaches.[5] The performance improvement demonstrated by our transform-coding like approach of DF-then-VQ provides further support for the idea that smarter quantization can easily improve performance. Combining the standard approaches of electronic structure theory with ideas from data compression could then be an interesting and potentially useful area.

References

1. Szabo, A. and Ostlund, N. "Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory," Dover Publications, New York, 1996.

2. Beebe, N. and Linderberg, J. *Int. Journal of Quantum Chem.* **1977** *12*

683-705.

3. DPedersen, T. B.; Lehtola, S.; Galván, I. F.; Lindh, R. The versatility of Cholesky decomposition of electron repulsion integrals. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-579sk-v4 .

4. Sun, Q. *et. al. WIREs Comput Mol Sci* **2018**, 8:e1340. doi: 10.1002/wcms.1340

5. Gok, A.; Di, S.; Alexeev, Y.; Tao, D.; Mironov, V.; Liang, X.; Capello, F. *IEEE Int. Conf. on Cluster Computing* **2018**