PaSTRI: A Novel Data Compression Algorithm for Two-Electron Integrals in Quantum Chemistry*

Extended Abstract

Ali Murat Gok
Northwestern University
Argonne National Labs
IL, USA
amg@u.northwestern.edu

Dingwen Tao
University of California
Riverside, CA, USA
ustc.dingwentao@gmail.com

Sheng Di
Argonne National Labs
IL, USA
sdi1@anl.gov

Vladimir Mironov
Lomonosov Moscow State University
Russia

Yuri Alexeev
Leadership Computing Facility,
Argonne National Labs
IL, USA

Franck Cappello
Argonne National Labs
IL, USA

ABSTRACT

Integral computations for two-electron repulsion energies are very frequently used applications in quantum chemistry. Computational complexity, energy consumption and the size of the output data generated by these computations scales with O(N4), where N is the number of atoms simulated in the system. In many applications, the same integrals are required to be calculated multiple times. Storing these values and reusing them requires impractical amounts of storage space; whereas recalculating them requires a lot of computations. On the other hand, generated data typically requires much less precision than the built-in floating point data types. We propose PaSTRI (Pattern Scaling for Two-electron Repulsion Integrals), a fast novel compression algorithm which makes it possible to calculate these integrals only once, store them, and reuse them at much smaller computational cost then recalculation. PaSTRI is lossy compared to floating point numbers, but still maintains the precision level required by the integral computations. PaSTRI is an extension to SZ[1] compressor package as a part of ECP-EZ. We have evaluated our compressor using GAMESS[4] dataset, and achieved 17.5:1 compression ratio whereas compression ratios for original SZ was 8.0:1 and ZFP[3] was 7.1:1.

KEYWORDS

Floating point lossy compression, integer compression, two-electron repulsion energy

*Acknowledgments This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This work was supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research Program, under Contract DE-AC02-06CH11357.

SC'17, November 2017, Denver, CO, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 EXASCALE COMPUTING PROJECT FOR DATA COMPRESSION (ECP-EZ)

The ECP EZ project aims at significantly improving the compression factors and the usability of SZ. Our goal is to produce a production-quality lossy compressor responding to the needs of exascale application users. PaSTRI is a part of ECP-EZ project, implemented as one of the compression algorithms in the SZ compressor.

2 INTRODUCTION

All properties of the quantum mechanical system are determined by the wave functions which are obtained by solving the Schrodinger Equation. These wave functions define the molecular orbitals, and they can be represented as finite sets of Linear Combination of Atomic Orbitals (LCAO approximation). Solving the Schrodinger Equation requires constructing and solving Fock matrix, where the most time-consuming step is computation of two-electron repulsion integrals. Number of integrals scale with N^4, where N is the number of atomic orbitals(basis functions). Typically, the integrals cannot fit into memory and iterative solution of Schrodinger equation requires these integrals to be recalculated for every iteration, which is typically 20-30 times. A much better strategy is to compute electron repulsion integrals (ERIs) once, compress and store them in memory, and read ERIs from memory every time we need them. Consequently, we can potentially speed up calculations by 20-30 times.

3 TWO-ELECTRON INTEGRALS

Two-electron repulsion integral calculation involve two couples of atoms, where each couple is sharing one electron. Each one of the four atoms are represented by an index, leading each output data point to include four 16-bit integers as indexes, and a 64-bit double precision floating point number for the resulting integral value. Although electrons are shared by two atoms, they are represented by a single Gaussian function during calculations. This results in periodic behavior in the output data, where the periodicity is defined by the shell types in atom couples. There are different block types ((sp|df), (df|pd), (dd|ff), (ff|ff), etc.) which are represented by orbital shell types of each atom.

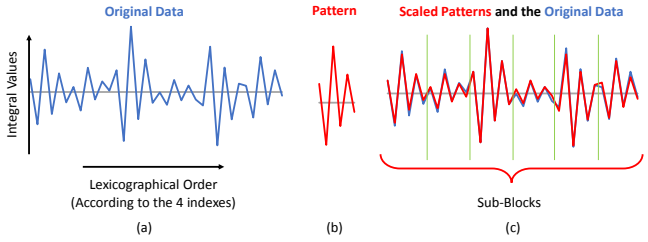

Figure 1: Original data has a pattern that repeats itself at different scales, with small deviations

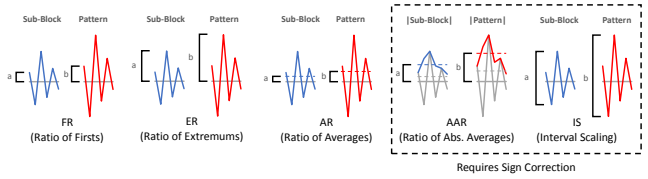

Figure 2: Different scale calculation methods. Scale is calculated as a/b (Sign correction may also be applied)

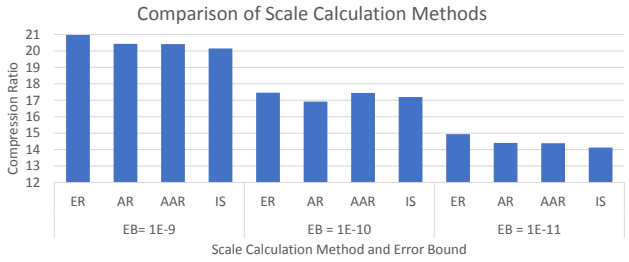

Figure 3: Compression ratios for different scale calculation methods

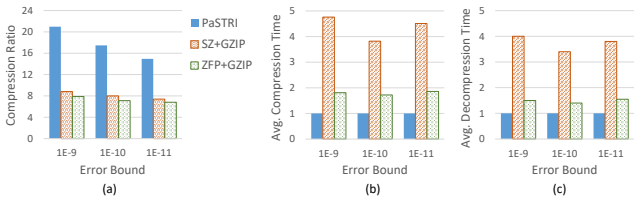

Figure 4: PaSTRI in comparison with other compressors

It should also be noted that integral values typically require less precision than the double precision floats, leading the original data to be unnecessarily large. Maximum tolerable error is called Error Bound(EB), which is 10^{-10} by default in GAMESS. PaSTRI always satisfy the specified error bound in our decompressed values.

4 ALGORITHM

In order to observe the periodicity, we have a non-sparse representation of the integral values in a single block (Fig1a). All possible index combinations are included in the lexicographical order, and if the original block has sparse representation the skipped values can simply be replaced by zeros. In this representation, we observe a periodic behavior within each block. Although there are some relatively small deviations, the general behavior is observed as there is a Pattern(P)(Fig1b) that repeats itself in different scales. We name each one of these repetitions as a Sub-Block(SB) (Fig1c). SB length is fixed, and determined by the multiplication of the number of the last 2 orbital's shells. We choose one of these sub-blocks to be the pattern.

For each SB, one scale value is calculated. We evaluated several methods to calculate the scales (Fig2), Where Sign Correction is 1 or -1. AAR and IS methods require Sign Correction because they can only produce positive values without it, whereas the scale values can be positive or negative. Reliable sign correction methods increase computational costs.

A data point whose values is not within EB distance to the scaled pattern is called an Outlier. We only store the difference between the value of each outlier point and the scaled pattern, and name it Error Correction(EC). This difference is typically much smaller than the outlier values, consequently contributing to a higher compression ratio. EC representation also employs quantization, always using bin size of 2*EB, which guarantees all decompressed values to be within EB distance to the original data values. Outliers are encoded by using a variable length encoding where the values 0, 1 and -1 are encoded by using fewer bits than other values since they constitute the vast majority of cases.

During encoding phase, firstly the quantized pattern and scale values are written to the output file. Then our algorithm chooses the output format with the smallest size of the following: Uncompressed Sparse, Compressed Non-Sparse, Compressed Sparse.

5 EVALUATION

Due to their computational complexity and output data size, we have focused on d and f orbitals during our evaluations. We present compression ratios for (dd|dd), (dd|ff), (ff|dd), (ff|ff) configurations using different scale calculation methods in Fig3. We choose ER(Ratios of Extremums) in the implementation of PaSTRI because it has slightly better compression ratios, and also requiring the smallest amount of computations. We then compare PaSTRI with vanilla SZ and ZFP compressors in Fig4, where integer indexes are compressed using GZIP[2] in the latter two.

PaSTRI achieves 2.5x better compression ratio while maintaining the lowest execution time.

6 CONCLUSION

We propose PaSTRI, a fast, novel compression algorithm for two-electron repulsion integrals which is a frequent application in quantum chemistry. Our algorithm consist of determining the pattern, calculating the scales, calculating error corrections and then encoding the whole block. All the steps have O(N) complexity, where N is the points per block. PaSTRI achieves 2.5 times better compression ratios at lower computational costs. We have evaluated our results on GAMESS dataset, but also expect to have similar results for other ERI applications because the way integrals are calculated would be similar.

REFERENCES

[1] Z. Chen D. Tao, S. Di and F. Cappello. 2017. Significantly Improving Lossy Compression for Scientific Data Sets Based on Multidimensional Prediction and Error-Controlled Quantization. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 1129–1139.

[2] Marc Adler Jean-loup Gailly. [n. d.]. Gzip compression. ([n. d.]). http://www.gzip.org

[3] P. Lindstromo. 2014. Fixed-Rate Compressed Floating-Point Arrays. In *IEEE Transactions on Visualization and Computer Graphics*. 2674–2683.

[4] J. A. Boatz M. W. Schmidt, K. K. Baldridge et al. 1993. General atomic and molecular electronic structure system. In *J. Comput. Chem*, Vol. 14. 1347–1363.