Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data

DAVID KRASOWSKA, Clemson University, USA

ROBERT UNDERWOOD (ADVISOR), Argonne National Laboratory, USA

JULIE BESSAC (ADVISOR), Argonne National Laboratory, USA

JON C. CALHOUN (ADVISOR), Clemson University, USA

SHENG DI (ADVISOR), Argonne National Laboratory, USA

FRANCK CAPPELLO (ADVISOR), Argonne National Laboratory, USA

In the fields of science and engineering, lossy compression plays a growing role in running scientific simulations, as output data is on the scale of terabytes. Using error bounded lossy compression reduces the amount of storage for each simulation; however, there is no known bound for the upper limit of lossy compressibility. Data correlation structures, compressors and error bounds are factors allowing larger compression ratios and improved quality metrics. This provides one direction towards quantifying lossy compressibility. Our previous work explored 2D statistical methods to characterize the data correlation structures and their relationships, through functional models, to compression ratios and quality metrics for 2D scientific data. In this poster, we explore the expansion of our statistical methods to 3D scientific data. The method was comparable to 2D. Our work is the next step towards evaluating the theoretical limits of lossy compressibility used to predict compression performance and optimally adapt compressors.

Additional Key Words and Phrases: compression, lossy compression, high performance computing, statistical correlation analysis

ACM Reference Format:

David Krasowska, Robert Underwood (advisor), Julie Bessac (advisor), Jon C. Calhoun (advisor), Sheng Di (advisor), and Franck Cappello (advisor). 2022. Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data. 1, 1 (October 2022), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: David Krasowska, dkrasow@clemson.edu, Clemson University, 433 Calhoun Dr., Clemson, South Carolina, USA, 29631; Robert Underwood (advisor), runderwood@anl.gov, Argonne National Laboratory, 9700 S Cass Ave., Lemont, Illinois, USA, 60439; Julie Bessac (advisor), jbessac@anl.gov, Argonne National Laboratory, 9700 S Cass Ave., Lemont, Illinois, USA, 60439; Jon C. Calhoun (advisor), jonccal@clemson.edu, Clemson University, 433 Calhoun Dr., Clemson, South Carolina, USA, 29631; Sheng Di (advisor), sdi1@anl.gov, Argonne National Laboratory, 9700 S Cass Ave., Lemont, Illinois, USA, 60439; Franck Cappello (advisor), cappello@mcs.anl.gov, Argonne National Laboratory, 9700 S Cass Ave., Lemont, Illinois, USA, 60439.

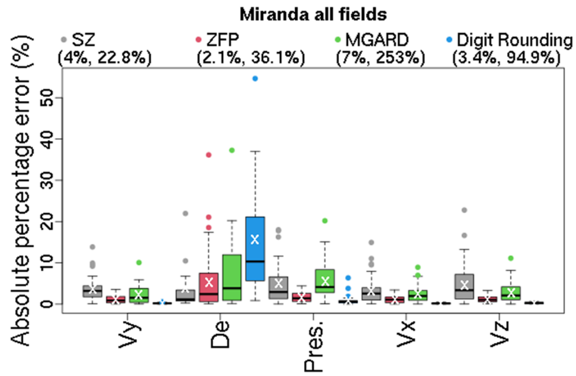

Fig. 1. Absolute percentage error for each buffer on the Miranda dataset from SDRBENCH [3]. Different lossy compresssors were studied as shown in Table 1. SZ, ZFP, MGARD, and Digit Rounding exhibit median absolute percentage error (MAPE) <= 7%.

1 INTRODUCTION AND MOTIVATION

Scientific research can benefit from error-bounded lossy compressors due to greater compression ratios (CRs) in relation to lossless compressors[4]. This greater performance allows applications to run with larger and more frequent datasets due to faster I/O times and smaller I/O volumes. Entropy[8] is the mathematical limit on *lossless* compressibility. "Compressibility" refers to the compression ratio (CR) possible. There is no known equivalent bound for *lossy* compression with a provided user error bound. Establishing this bound will allow researchers to anticipate compression performances and to see if compressors approach optimality.

One possible direction is using data correlation structures, compressor types and error bounds to study their impact on compressibility. In this poster, we focus on data correlations and their link to CRs. Many compressors implicitly or explicitly exploit correlation in the data. Analyzing the relationships between correlation structures and compression performances allows researchers to anticipate compression performances and adapt compressors. Ultimately, we seek to establish an entropy-like metric for lossy compression algorithms, which can guide the lossy compression community to optimal development and usage. Our previous work has established functional models to predict CRs within 2D data. The achieved goal of the research is to (1) explore relationships, through functional models, to CRs and quality metrics for 3D data and (2) be a next step towards the theoretical limit for lossy compressibility.

2 OUR PREVIOUS WORK

We test with the Miranda[3] hydrodynamics data for large turbulence simulations. These 3D samples are 256×384×384. The 3D datasets are split into separate 2D files based on equally spaced slices along the first dimension. We ran our training model and statistical predictors on the individual slice. SZ and ZFP exhibited a median percentage error (MAPE) <= 4% as shown in Figure 1. MAPE is the difference between predicted and observed CRs on the validation set. Generating 2D slices inhibits the ability to see correlation structures prevalent within the unsliced 3D dataset.

$$\log(\text{CR}) = a + b \times \log(\text{Qentropy}) + c \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) + d \times \log(\text{Qentropy}) \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) + \epsilon \quad (1)$$

3 METHODOLOGY

Software	Version	Purpose
SZ[5]	@2.1.12.2	lossy compressor
ZFP[6]	@0.5.5	lossy compressor
MGARD[1]	@1.0.0	lossy compressor
TTHRESH[2]	@0.0.5	lossy compressor
Bit Grooming[10]	@2.9.0	lossy compressor
Libpressio[9]	@0.83.1	compress and measure the data
LLVM	@12.0.1	optimize and produce low level code for Julia
Julia	@1.7.2	run HOSVD and Qentropy
CUDA	@11.7.0	run kernels on Nvidia GPUs
CUSOLVER	@11.3.5	linear systems solver on Nvidia GPUs

Table 1. Software Used For the Study

3.1 Statistical Methods and Compression Statistics

Both statisical predictors that were used in Section 2 need to be expanded for 3D. In this work, we focus on the Higher Order SVD (HOSVD) in combination with the quantized entropy and its effect on compressibility. The HOSVD provides a measure of correlation within the data. CR is an important statistic within lossy compression due to its positive correlation with storing and processing as much data as possible.

3.1.1 Compression Statistics. Compression ratio, the ratio of the uncompressed data size by the compressed data size, is used to compare differences between compressors and their efficiency. CR depends on: error bounds, compressor used, and correlation structures within data. The CR is easily comparable between different compressors and bounds.

3.1.2 Statistical Methods. The HOSVD is the expansion of the 2D SVD operation into higher dimensions. There are four main steps: (1) unfold a tensor along each dimension, (2) perform SVD on each unfolding to obtain singular values, (3) combine singular values using at least two from each unfolded dimension to reconstruct 99% of the data and (4) perform SVD truncation (ratio of singular values needed to reconstruct 99.9% of the data).

Quantized entropy is taking the entropy of quantized data. Quantizing is the mapping of original data into block intervals based off the user defined error bound. After the quantization of the data is done, a probability distribution function (PDF) of the different symbols is found. The entropy, $H(x) = -\sum P(xi) \log_2 P(xi)$, is calculated using the PDF.

4 RESULTS

4.1 Compressors and Software

Each lossy compressor in Table 1 is run with absolute error bounds: 1E-5, 1E-4, 1E-3, and 1E-2. All experiments are run on Clemson's Palmetto cluster using a DGX node consisting of an AMD EPYC 7742 64-Core Processor, 1 TB of RAM, and 4 Nvidia A100 GPUs. The OS is Rocky Linux 8.6 with compiler GCC 9.5.0.

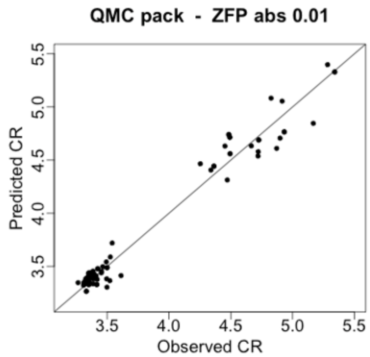

Fig. 2. ZFP 1e-2 absolute error bound observed vs predicted CR

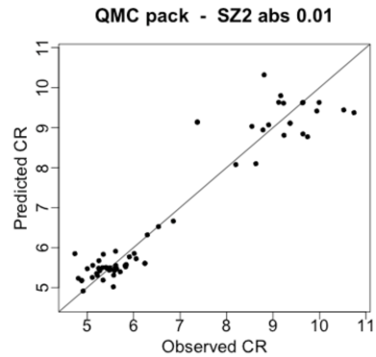

Fig. 3. SZ 1e-2 absolute error bound observed vs predicted CR

288 3D orbitals from QMCPack from SDRBENCH[3] containing structures of atoms, molecules, and solids were studied. Figure 2 and 3 both show the observed vs predicted CRs for QMCPack orbitals. Many of the points fall near the line of equality. As shown in Table 2, the predicted CR exhibits low MAPEs($< 7.5\%$) for SZ2, ZFP, MGARD, and Bit Grooming. However, TTHRESH produces a much higher error at 24.8%. The higher error is left to future exploration of more complex statistical predictors.

We determined that the regression model found in Eq. (1) worked with low MAPEs on 2D datasets. We determined there is a relationship between CR and overall correlation structure within 3D data which can be considered as next step toward developing an entropy-like metric for lossy compressors.

Compressor	MAPE (median percentage error)	10% Quantile	90% Quantile
SZ	4.5%	3.2%	5.7%
ZFP	1.7%	1.3%	3.5%
MGARD	0.6%	0.4%	1.3%
Bit Grooming	7.4%	5%	9.3%
TTHRESH	24.8%	15.7%	27.7%

Table 2. Results for 288 Orbitals of QMCPACK[3] using the 3D extension of the statistical methods

4.2 Performance

The HOSVD is a slow algorithm even in parallel on the CPU; therefore, an accelerated version is needed. We implemented a multi-GPU parallel version in CUDA[7]. Figure 4 shows, a noticeable speedup is achieved by adding a single GPU, with a maximum speedup of 57×. A slight increase was achieved for each GPU added after the first.

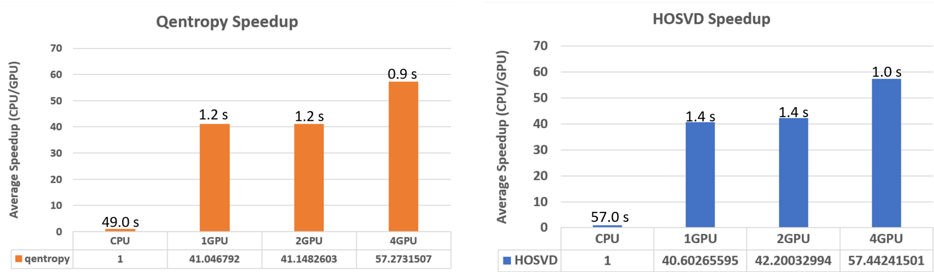

Fig. 4. Qentropy (left) and HOSVD (right) speedup (CPU/GPU runtime) of the average performance over 6 runs on the baryon_density buffer (512x512x512) from the NYX dataset[3]. The NVIDIA A100 GPU is scaled with the DGX node on the Clemson Palmetto Cluster.

5 CONCLUSION AND FUTURE WORK

The ability to accurately predict CRs in 3D is comparable to our 2D methods. Our method is flexible across compressors, error bounds, and datasets. The comparison between different compressors for minimum data size is enabled due to the reuse of the statistical predictors. This work leads to the next step towards theoretical quantification of lossy compressibility. Establishing this limit would allow for the maximum efficiency for storing large scientific datasets. In continuation of this research, there are a few goals for the future: i) further reduce computational costs by generating samples to estimate CR and ii) create a training free model of CR based on correlation metrics and error bound.

ACKNOWLEDGMENTS

Clemson University is acknowledged for generous allotment of compute time on the Palmetto cluster. This material is based upon work supported by the National Science Foundation under Grant No. SHF-1910197 and SHF-1943114.

REFERENCES

[1] Mark Ainsworth, Ozan Tugluk, Ben Whitney, and Scott Klasky. 2019. Multilevel Techniques for Compression and Reduction of Scientific Data—The Multivariate Case. *SIAM Journal on Scientific Computing* 41, 2 (Jan. 2019), A1278–A1303. https://doi.org/10.1137/18m1166651

[2] Rafael Ballester-Ripoll, Peter Lindstrom, and Renato Pajarola. 2019. TTHRESH: Tensor compression for multidimensional visual data. *IEEE transactions on visualization and computer graphics* 26, 9 (2019), 2891–2903.

[3] F Capello, K Zhao, S Di, D Tao, J Bessac, and Z Chen. 2018. https://sdrbench.github.io/

[4] Franck Cappello, Sheng Di, Sihuan Li, Xin Liang, Ali Murat Gok, Dingwen Tao, Chun Hong Yoon, Xin-Chuan Wu, Yuri Alexeev, and Frederic T Chong. 2019. Use cases of lossy compression for floating-point data in scientific data sets. *The International Journal of High Performance Computing Applications* 33, 6 (July 2019), 1201–1220. https://doi.org/10.1177/1094342019853336

[5] Xin Liang, Sheng Di, Dingwen Tao, Sihuan Li, Shaomeng Li, Hanqi Guo, Zizhong Chen, and Franck Cappello. 2018. Error-Controlled Lossy Compression Optimized for High Compression Ratios of Scientific Datasets. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. https://doi.org/10.1109/bigdata.2018.8622520

[6] Peter Lindstrom and Martin Isenburg. 2006. Fast and Efficient Compression of Floating-Point Data. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (Sept. 2006), 1245–1250. https://doi.org/10.1109/tvcg.2006.143

[7] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. 2022. CUDA, release: 11.7.0. https://developer.nvidia.com/cuda-toolkit

[8] C. E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3 (July 1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[9] Robert Underwood, Sheng Di, Jon C. Calhoun, and Franck Cappello. 2020. FRaZ: A Generic High-Fidelity Fixed-Ratio Lossy Compression Framework for Scientific Floating-point Data. *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2020). https://doi.org/10.1109/ipdps47924.2020.00065

[10] Charles S. Zender. 2016. Bit Grooming: Statistically Accurate Precision-Preserving Quantization with Compression, Evaluated in the netCDF Operators (NCO, v4.4.8+). *Geoscientific Model Development* 9, 9 (Sept. 2016), 3199–3211. https://doi.org/10.5194/gmd-9-3199-2016