Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets

Julie Bessac[‡] , Robert Underwood [‡] , David Krasowska[*], Jon Calhoun[*], Sheng Di[‡] , and Franck Cappello[‡]
‡ Mathematics and Computer Science Division Argonne National Laboratory - Lemont, USA
∗ Holcombe Department of Electrical and Computing Engineering, Clemson University - Clemson, USA

Goals

In lossless compression, entropy provides theoretical limit on compressibility (associated with coding) of data
-> no equivalent for lossy compressors

1. Characterize statistics of the data that impact lossy compression, e.g. correlation structures of scientific datasets, patterns, range of values, …

2. Explore their relationships, through functional regression models, to compression ratios

-> These models form the first step towards evaluating theoretical limits of lossy compressibility
> how far are existing compressors to optimality
> help optimize compressors
> allow maximum efficiency for storing scientific datasets

Exploring Lossy Compressibility through Statistical Correlations of Scientific Datasets
D. Krasowska, J. Bessac, R. Underwood, S. Di, J. Calhoun, and F. Cappello.
7th International Workshop on Data Analysis and Reduction for Big Scientific Data in conjunction with SC '21, 2021.
https://arxiv.org/pdf/2111.13789.pdf

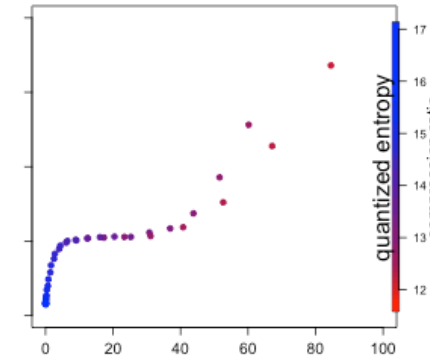
Procedure and quantities of interest

- Performed statistical and compression analysis on several datasets (analysis of 2D slices as a start)
  - > Synthetic 2D-Gaussian samples with controllable correlation structured
  - > Scientific datasets: Miranda, SCALE-LETKF available on SDRBench [1]

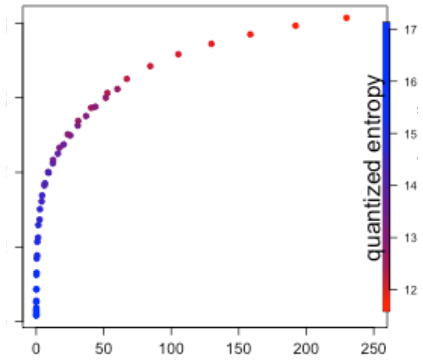
- Lossy compressors: SZ [2], ZFP [3], MGARD [4], BitGrooming, Digit Rounding
  - > compression ratios (impacted by error bound, compressor used, and structures within data)

- Statistics of interest: independent of the compressors
  - > correlation strength across grid-points
  - > variance: value range, variability
  - > quantized entropy: entropy $-\sum_{i=1}^{n} P(q(x_i))\log(P(q(x_i)))$ of quantized data

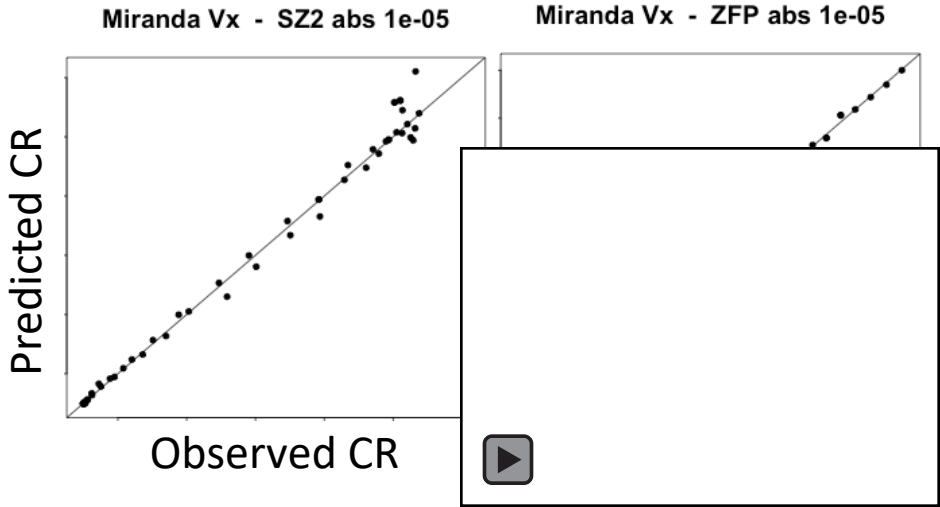
Miranda velocity x – SZ abs=1e-4



Miranda velocity x – ZFP abs=1e-4



Spline regression model to predict compression ratio



References

[1] K. Zhao, S. Di, X. Lian, S. Li, D. Tao, J. Bessac, Z. Chen, and F. Cappello, “SDRBench: Scientific data reduction benchmark for lossy compressors,” in 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020, pp. 2716–2724. [Online]. Available: https://sdrbench.github.io

[2] X. Liang, S. Di, D. Tao, S. Li, S. Li, H. Guo, Z. Chen, and F. Cappello, “Error-controlled lossy compression optimized for high compression ratios of scientific datasets,” in 2018 IEEE International Conference on Big Data (Big Data). IEEE, Dec. 2018. [Online]. Available: https://doi.org/10.1109/bigdata.2018.862252

[3] P. Lindstrom and M. Isenburg, “Fast and efficient compression of floating-point data,” IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 5, pp. 1245–1250, Sep. 2006. [Online]. Available: https://doi.org/10.1109/tvcg.2006.143

[4] M. Ainsworth, O. Tugluk, B. Whitney, and S. Klasky, “Multilevel techniques for compression and reduction of scientific data- the multivariate case,” SIAM Journal on Scientific Computing, vol. 41, no. 2, pp. A1278–A1303, Jan. 2019. [Online]. Available: https://doi.org/10.1137/18m1166651