IDEALEM: Statistical Similarity Based Data Reduction

Dongeun Lee[1] Alex Sim[2] Jaesik Choi[3] Kesheng Wu[2]

[1]Texas A&M University-Commerce [2]Lawrence Berkeley National Laboratory [3]Ulsan National Institute of Science and Technology (UNIST), Korea

dongeun.lee@tamuc.edu asim@lbl.gov jaesik@unist.ac.kr kwu@lbl.gov

How to compress high-entropy data?

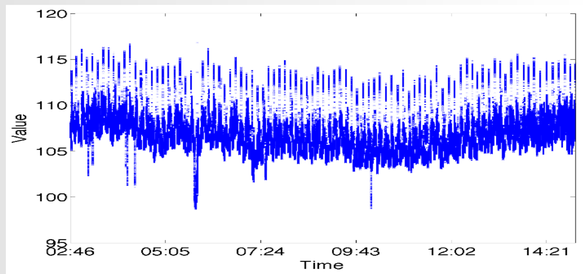
- High-entropy data is hard to compress!
- But, high-entropy data may be generated from a few simple generators

Is it possible to capture these generators while also capture "interesting" features?

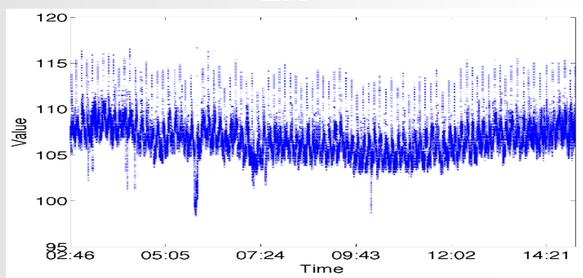
Compression Quality Revisited

- Popular distance measure - MSE, SNR
- Works fine for many applications (image, etc.) where data varies smoothly

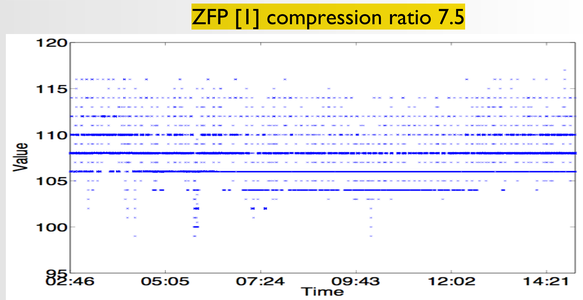
- Scientific data, sensor data? - random fluctuation (ex, power grid monitoring data)



Compression with state-of-art method ZFP



ZFP [1] compression ratio 7.5



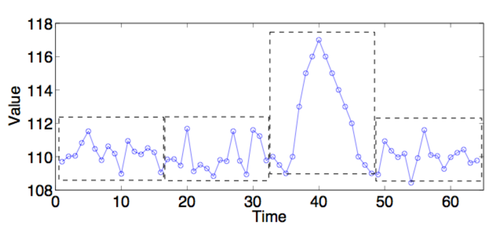
ZFP compression ratio 9.14

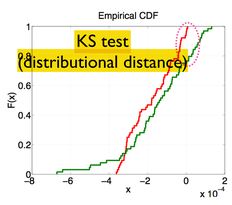
Trying hard to reproduce every peak and valley leads to unsatisfactory performance

Statistical Similarity Based Data Reduction

- IDEALEM (Implementation of Dynamic Extensible Adaptive Locally Exchangeable Measures) [2]
- Preserve key statistical properties
  - **Reproducing the probability distribution of original data, instead of every peak and valley**
- Which statistical similarity? - Kolmogorov-Smirnov test (KS test) in current implementation

How IDEALEM Works





1st block stored

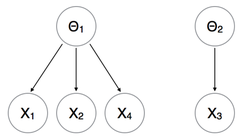
2nd block similar

3rd block not similar

4th block similar

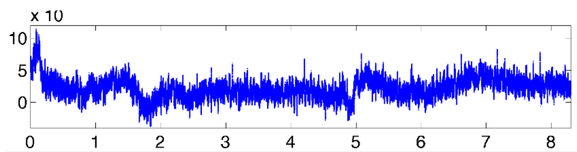
compressed stream

1st block | 3rd block |

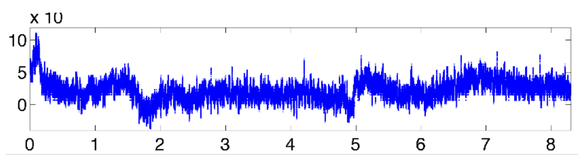


graphical representation

IDEALEM Achieves High Compression Performance While Capturing Interesting Events



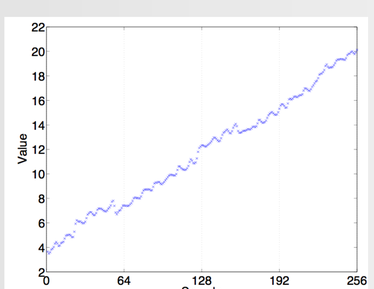
Original brain (EEG) data



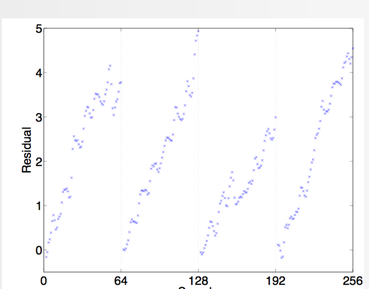
IDEALEM compressed data (compression ratio 106.6)

Extending IDEALEM for Non-Stationary Data

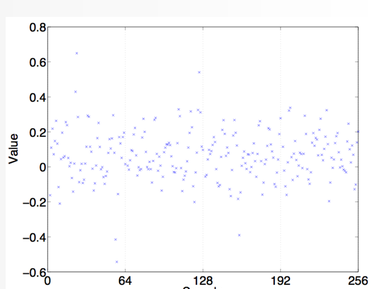
- Initial IDEALEM was effective for stationary distributions, but real-world data are often non-stationary
- Designed two methods for transforming non-stationary time series into locally stationary block
  - Residual transformation / delta transformation
  - Capture long-range trends in data and allow local variations to be compared through KS test
- Values in bounded ranges are also considered (angles between 0° and 360°, etc.)



phase angle of electricity

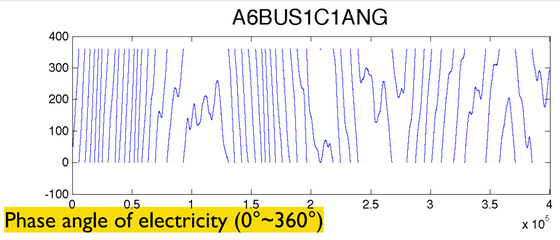


residual transformation

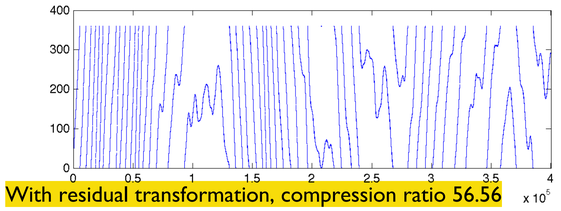


delta transformation

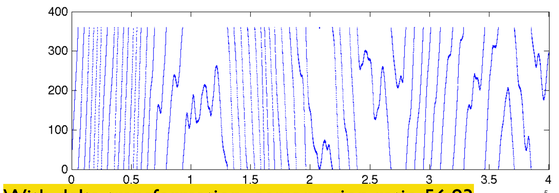
Reconstruction Quality of New IDEALEM



Phase angle of electricity (0°~360°)

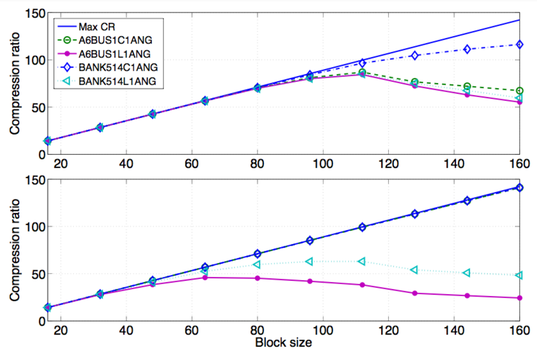


With residual transformation, compression ratio 56.56



With delta transformation, compression ratio 56.83

Compression Ratio of New IDEALEM



Conclusions

- **Yes, IDEALEM captures the "generators" as well as "interesting" features**
- IDEALEM compression ratios can be > 100
- IDEALEM shows better performance than best floating-point compression methods such as ZFP and SZ [3]

References

[1] P. Lindstrom. Fixed-rate compressed floating-point arrays. IEEE Trans Vis. Comput. Graphics, 20(12):2674–2683, Dec. 2014.

[2] D. Lee, A. Sim, J. Choi, and K. Wu. Novel data reduction based on statistical similarity. In Proc. SSDBM, 21:1–21:12, 2016. code available at http://datagrid.lbl.gov/idealem

[3] S. Di and F. Cappello. Fast error-bounded lossy HPC data compression with SZ. In Proc. IPDPS, 730–739, 2016.

This work was supported by the Office of Science of the U.S. Dept. of Energy under Contract No. DE- AC02-05CH11231, and by Basic Science Research Program through the National Research Foundation of Korea under NRF- 2014R1A1A1002662 and NRF-2014M2A8A2074096