Los Alamos
NATIONAL LABORATORY

Unbiased Reconstruction of Lossy-compressed Statistical Data

Boram Yoon (CCS-7)

ISTI Year-End Review
September 13, 2021

LA-UR-21-28904

Lossy Data Compression using Machine Learning

- Original data: $\{\boldsymbol{X}^{(k)} | \boldsymbol{X}^{(k)} \in \mathbb{R}^D, k = 1,2,3,\dots,N\}$
- Compressed data: $(\{\boldsymbol{a}^{(k)} | \boldsymbol{a}^{(k)} \in \mathbb{R}^L, k = 1,2,3,\dots,N\}, \Phi)$
  - Φ: Dictionary containing the information how to reconstruct $\mathbf{X}$ from $\boldsymbol{a}$; common for all k
  - $\boldsymbol{a}^{(k)}$: Vector in compressed (reduced dimension) space ($L \ll D$)
- Example: principal component analysis (PCA)
  - PCA finds orthogonal directions that maximize the variance as principal components
  - Compression by saving only the coefficients ($\boldsymbol{a}^{(k)}$) of the first few principal components
  - $\boldsymbol{X}^{(k)} \approx \Phi \boldsymbol{a}^{(k)}$
- **Problem:** Reconstruction is not exact, and could be biased.

How can we quantify the reconstruction error and correct the bias?

Bias Correction and Error Estimation of Lossy Reconstruction

- We are interested in an expectation value of a function of statistical data
  - $\langle f(\boldsymbol{X})\rangle = \frac{1}{N}\sum_k f\left(\boldsymbol{X}^{(k)}\right) \pm \frac{\sigma_{f(X)}}{\sqrt{N}}$
- Lossy reconstruction introduces error $\boldsymbol{X}^{(k)} \neq \boldsymbol{\Phi a}^{(k)} \equiv \boldsymbol{X}'^{(k)}$

  Simple average with the reconstruction is a biased estimator $\langle f(\boldsymbol{X})\rangle \neq \frac{1}{N}\sum_k f\left(\boldsymbol{X}'^{(k)}\right)$
- **Key idea:** Unbiased estimator of $\langle f(\boldsymbol{X})\rangle$ defined using small portion of original data

$$\bar{\boldsymbol{O}} = \frac{1}{N}\sum_{k=1}^{N} f\left(\boldsymbol{X}'^{(k)}\right) + \frac{1}{N_{bc}}\sum_{k=1}^{N_{bc}}\left(f\left(\boldsymbol{X}^{(k)}\right) - f\left(\boldsymbol{X}'^{(k)}\right)\right)$$

  - Bias correction is guaranteed: $\langle\bar{\boldsymbol{O}}\rangle = \langle f(\boldsymbol{X}')\rangle + \langle f(\boldsymbol{X}) - f(\boldsymbol{X}')\rangle = f(\boldsymbol{X})$
  - Risk is possibly large statistical error in real problems: The bias correction term (second term) increases the statistical error of $\bar{\mathrm{O}}$ accounting for the reconstruction quality

Statistical Error Increase due to Bias Correction

- For simplicity, consider a bias-corrected average of independent observables

$$\bar{\boldsymbol{O}} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{X}'^{(k)} + \frac{1}{N_{bc}}\sum_{k=1}^{N_{bc}} \left(\boldsymbol{X}^{(k)} - \boldsymbol{X}'^{(k)}\right)$$

- Variance of the *i*-th component of $\bar{\boldsymbol{O}}$ can be obtained as

$$\sigma_{\bar{O}_i}^2 \approx \frac{1}{N}\sigma_{X_i'}^2 + \frac{1}{N_{bc}}\sigma_{X_i - X_i'}^2 \approx \frac{\sigma_{X_i'}^2}{N}\left(1 + \frac{N}{N_{bc}} \frac{\sigma_{X_i - X_i'}^2}{\sigma_{X_i'}^2}\right)$$

- Quality of lossy-compression on statistical data: $\mathrm{Q}^2 \equiv \frac{1}{D}\sum_{i=1}^{D} \frac{\sigma_{X_i - X_i'}^2}{\sigma_{X_i}^2}$
  - ➢ Smaller Q^2 indicates the better compression

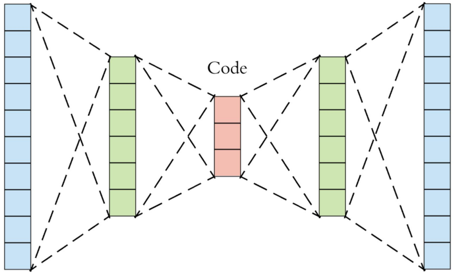
$\left(\quad N \quad\quad {}^{2} \right)$

Numerical Study with ML Compression Algorithms

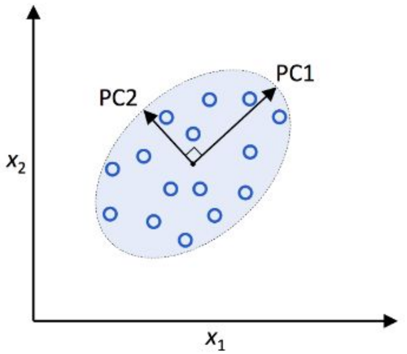
- **Binary compression using D-Wave**
  - Find a set of vectors (Φ) and their binary coefficients ($\boldsymbol{a}^{(k)}$) reconstructing $\boldsymbol{X}^{(k)}$

$$\min_{\Phi} \sum_{k} \min_{\boldsymbol{a}^{(k)}} \left[\sum_{i} \left(X_i^{(k)} - [\Phi \boldsymbol{a}^{(k)}]_i \right)^2 \right]$$

- **Bottle-neck Autoencoder (AE)**
  - Fully connected NN with ReLU
  - Encoder: (16, 128, 64, 32, N_z)
  - Decoder: (N_z, 32, 64, 128, 16)

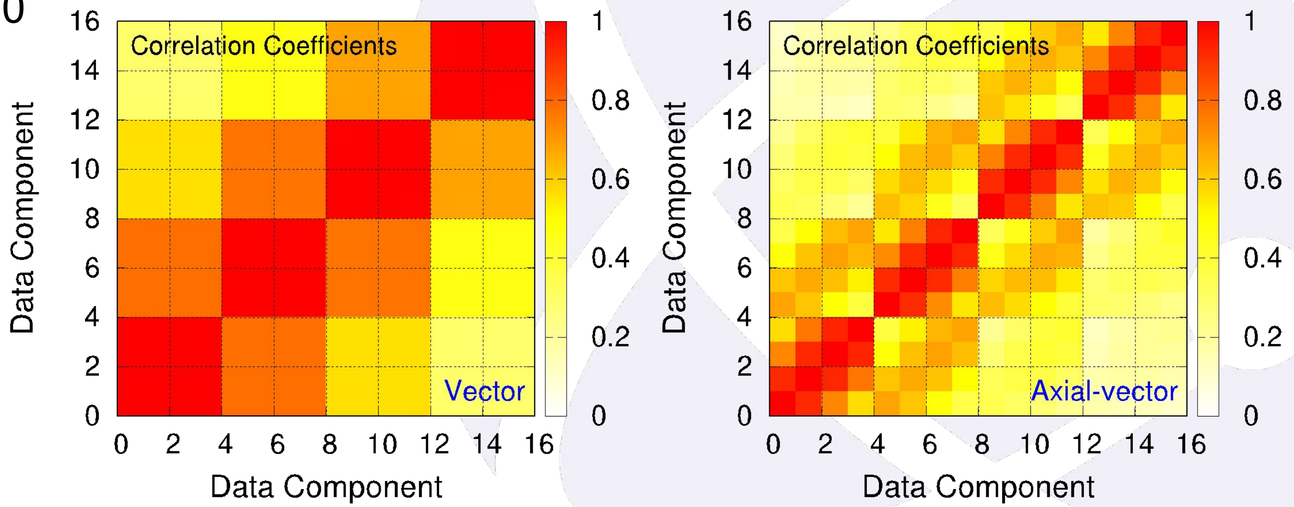


- **Principal Component Analysis (PCA)**
  - Compression by saving the first N_z coefficients of the principal components



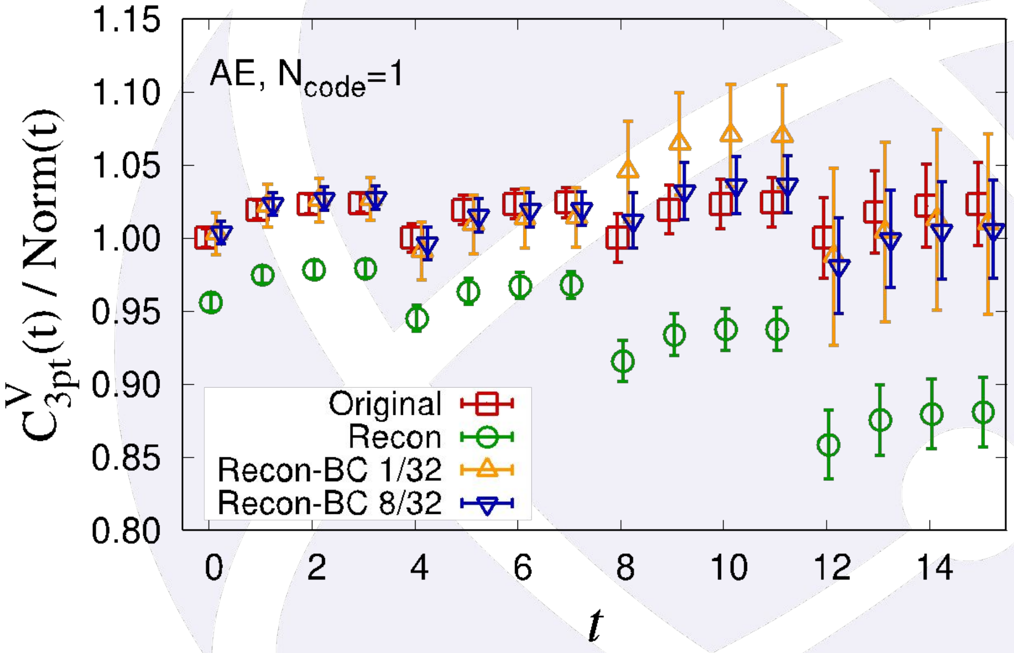
Numerical Study with ML Compression Algorithms

- Data
  - Two different sets of simulation results in lattice quantum chromodynamics (QCD)
  - Describing the interactions between a nucleon and external currents
  - Vector length: D = 16
  - Data size: N = 12800
  - Correlation pattern:



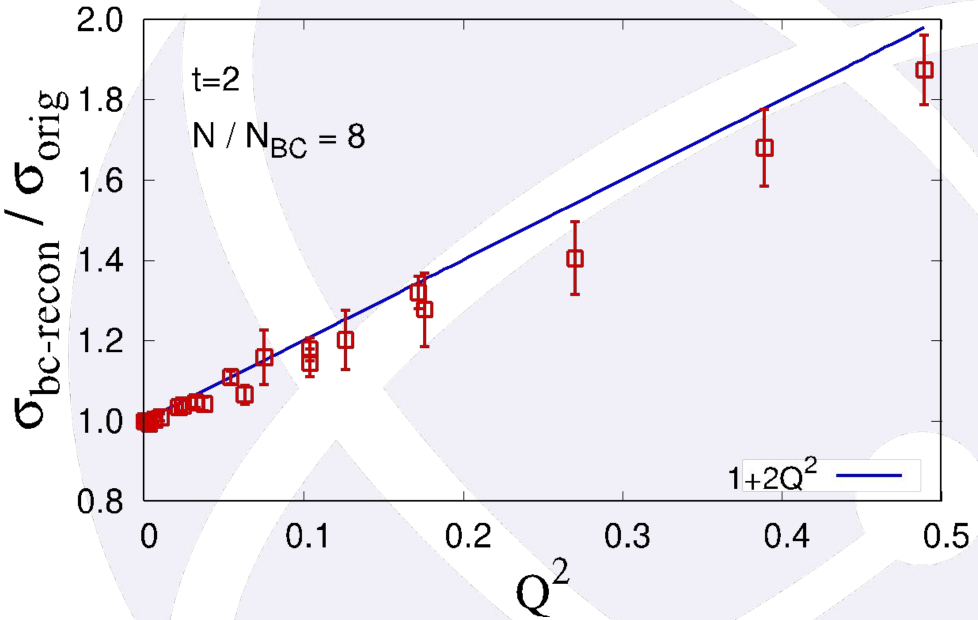
Effect of Bias Correction

- 16 components of the data compressed using Autoencoder with N_z =1
- Reconstruction without bias correction is biased
- Bias correction (yellow and blue) removes bias but increases statistical error
- The more bias correction data (N_{bc}) gives the smaller statistical error



Statistical Error Increase for Different Compression Qualities

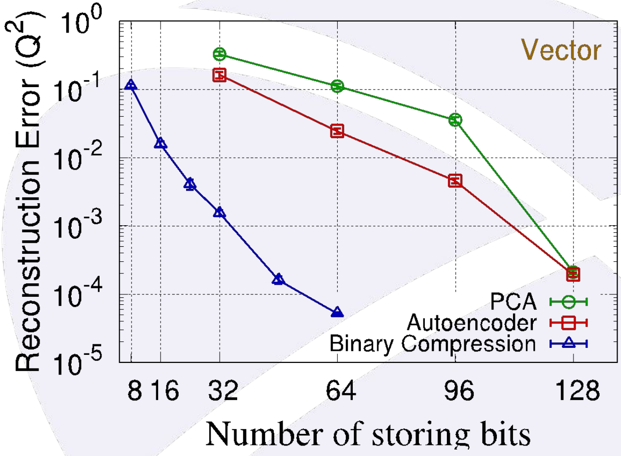
- The statistical error increase is proportional to $Q^2 \equiv \frac{1}{D}\sum_{i=1}^{D}\frac{\sigma^2_{X_i-X'_i}}{\sigma^2_{X_i}}$
- For *independent data*, the error increase ratio due to bias correction is expected to be $1+\frac{N}{2N_{bc}}Q^2$
- *Correlation between the data samples* makes it $1+\alpha\frac{N}{2N_{bc}}Q^2$ with $0<\alpha<1$

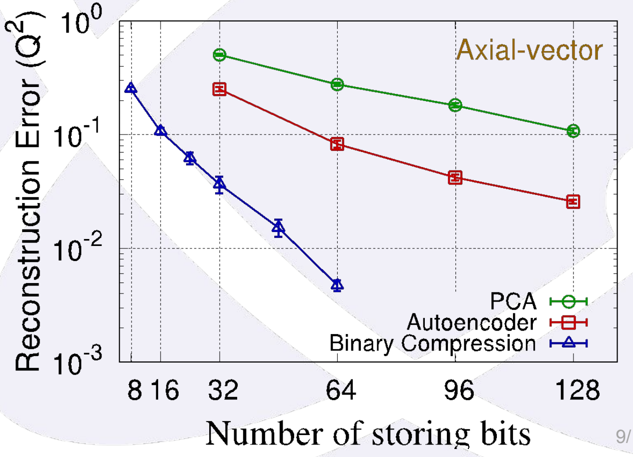


Expected Statistical Error Increase

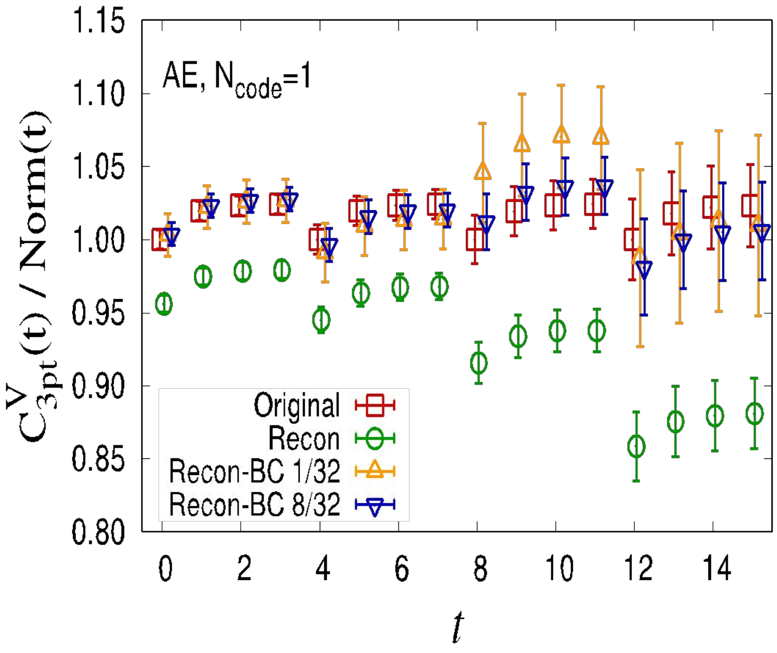
- $\frac{\sigma_{\mathrm{bc-recon}}}{\sigma_{\mathrm{orig}}} = 1 + \alpha \frac{\mathrm{N}}{2\mathrm{N}_{\mathrm{bc}}} \mathrm{Q}^2$ with $0 < \alpha < 1$
- With 10% of bias correction data ($\mathrm{N}/\mathrm{N}_{\mathrm{bc}} = 10$) and $\alpha = 0.5$, expected error increase is $1 + 2.5\mathrm{Q}^2$
- When $\mathrm{Q}^2 = 10^{-2}$, expected error increase is 2.5%
- When $\mathrm{Q}^2 = 10^{-3}$, expected error increase is 0.25%
- **Conclusion:**
  For good lossy compression algorithms, error increase due to bias correction is negligibly small

Figures: Compression of 16 floating-point numbers into $\mathrm{N}_{\mathrm{bit}}$ binary bits with three different compression algorithms





Unbiased Reconstruction of Lossy-compressed Statistical Data



Reconstruction of lattice QCD data from a lossy compression with and without bias correction. Bias correction (yellow and blue) removes bias but increases statistical error accounting for the reconstruction quality of the compression algorithm. The more bias correction data gives the smaller statistical error.

Project Description

Demonstration of a novel bias correction algorithm for the reconstruction of lossy-compressed statistical data

Project Outcomes

- Showed that the bias correction algorithm removes the bias in the lossy reconstruction
- Demonstrated that statistical error increase due to the bias correction is small for good compression algorithms
- Paper in preparation

PI: Boram Yoon (CCS-7)
Total Project Budget: $30K
ISTI Focus Area: Computational and Data Integrity

END