EE 274 Final Report: Compression and Neural Data

Alice Tor, Yuxin Wu

15 December, 2023

1 Introduction

The complexity of neural data changes as the brain process information during events. Compression ratio may be used as a convenient estimator for the complexity of a given signal. In particular, previous work has shown that compression algorithms may be used to detect seizure from neural array recordings in epileptic patients, historically done with tedious manual inspection [1,2]. Previous work has also demonstrated that lossy compression algorithms, such as those used in video compression, perform better than lossless compressors like GZIP [1, 3]. However, it remains unknown how these findings generalize to other events and neural data modalities.

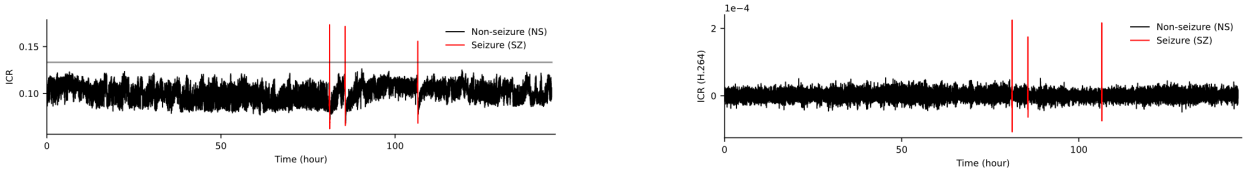
In this project, we investigated the use of the inverse compression ratio (ICR) over three distinct datasets spanning three different species and measurement modalities in order to analyze how compression algorithms may be used to track the underlying structure of neural signals recorded with different modalities and tracked behaviors. We also review previous work on detecting seizure in epilepsy patients. This project calculates lossless compression ratios using the ZSTD and GZIP lossless compressors, and calculates lossy compression ratios using the H264 lossless compressor. While our data is not video-based, lossy video compressors (like H264) have been used in the literature due to their ease in use with time-series data [1].

Through our survey of different datasets, we conclude that we are indeed able to track neural events across different measurement modalities and species. In particular, the inverse compression ratio may be used as a reliable and convenient way to detect events that perturb the temporal structure of neural activity.

2 Previous Work

2.1 Human Intracortical Array Data

Previous work has been done to characterize the performance of the inverse compression ratio in epilepsy, using high-resolution intracranial electroencephalography (iEEG) [1]. To summarize findings, ICR was successfully shown to produce threshold-detectable spikes that correlated tightly to seizure onset. Furthermore, lossy ICR was also computed using H264, and also successfully produced threshold-detectable spikes that correlated tightly to seizures. Notably, H264 demonstrated an improvement in detection accuracy over lossless compressors [1].



(a) Lossless ICR

(b) Lossy ICR

Figure 1: ICR plotted over 150 hours, calculated with both lossless (left) and lossy (right) compressors. Seizure times are labeled in red. [1]

3 Methods

3.1 Data

Access information for all data used in this project are provided in the **Code and Data** section below.

The first dataset explored in this project contains intracortical single-unit spike recordings over a series of days from one non-human primate (NHP). The arrays (96-electrode Utah array) were implanted in the primary motor cortex (M1) and dorsal premotor cortex. Due to the recording quality, in this project we only used the signals recorded

from the M1. Days -3 to -1 recorded the NHP performing a standard radial reaching task [4]. On Day 0, a small electrolytic lesion was administered, effectively removing part of the brain. Recordings were also collected for an additional 9 days.

The second dataset contains intracortical probe (NeuroPixel) recordings from the medial entorhinal cortex (MEC) of five different mice performing a navigation task. In this task, mice traversed a virtual reality linear track. During the first 100 trials, the mice explored the track without external intervention. Each mouse was then given a ketamine injection and recorded from for an additional 200 trials.

The final dataset contains time-varying intensity signals of individual cells recorded by longitudinal calcium imaging in ablation experiments in the zebrafish spinal cord. The recordings were made before, and after an ablation event, where part of the spinal cord was transversely lesioned.

3.2 Compressors

In this work, we use two lossless compressors, GZIP and ZSTD, as well as one lossy compressor, H264. GZIP is based on DEFLATE, a lossless data compression algorithm that is based on Huffman coding and the LZ77 algorithm. On the other hand, ZSTD is based on a combination of Huffman coding, finite state entropy coding, and the LZ77 algorithm [6]. The LZ77 algorithm works by using a sliding window to read and encode repetitive "phrases" of data, allowing it to take advantage of repeated sections of data. This allows it to perform well when data is not random and low in noise. Finite state entropy coding is a version of tabled asymmetrical numerical systems (tANS), discussed in class. Unlike GZIP, ZSTD is designed to best take advantage of modern computing resources [6]. GZIP and ZSTD are able to achieve similar compression rates, but ZSTD is often more efficient. However, GZIP is still widely used in most applications, so we chose to explore both compressors in this project. We used the GZIP and ZSTD packages in Python 3 to generate compression ratios in this project.

H264 is a lossy video compression algorithm. Like most video compression algorithms, it works by taking advantage of repeating details and visual features between frames so that video frames do not necessarily need to be individually saved in order to recreate a full-length video. The lossiness of H264 may be tuned. We used the SciKit-Video package (which itself uses FFmpeg) in Python 3 to generate compression ratios for this project.

3.3 The Inverse Compression Ratio

The compression ratio is defined as the size before compression divided by the size after compression. Thus the ICR can be computed as the size after compression over the size before compression; generally we always expect the ICR to be some value between 0 and 1. Using the aforementioned compressors, ICR is an easily-computed and ready-to-use metric.

In our work, we first divided our work into one of two conditions, depending on the dataset. We calculated lossless ICR values by simply applying GZIP and ZSTD to the data. To calculate lossy ICR values, we first created a set of "frames" by arranging the data according to spatial information provided by dataset metadata. We then generated a lossy video of these frames with H264. ICR was calculated using the size of the raw frames and the size of the outputted file.

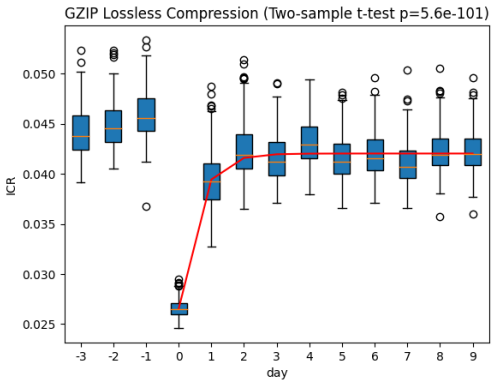
4 Results

4.1 NHP Intracortical Single-unit Spike Data

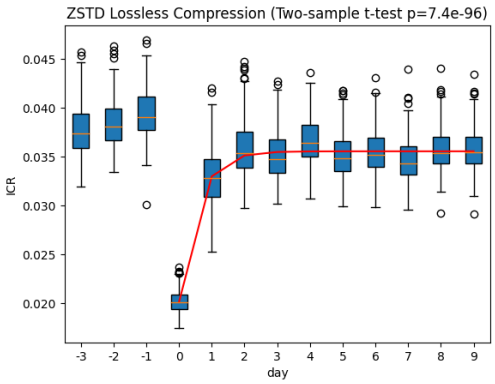
In this session, we applied the compressors to the NHP dataset to test if ICR is a good indicator of a lesion in the primary motor cortex.

Figure 2 shows the ICR of the two lossless compressors on trials across different days. The boxplots show the medians and quartiles. A two-sample t-test between pre-lesion days and post-lesion days suggests that the compression rates have significant changes after lesion (p=5.6e-101 for the GZIP compressor and p=7.4e-96 for the ZSTD compressor), which indicates that the compression rate can be used to measure brain damage. Moreover, the red line shows the regression between post-lesion compression rates and recovery days (the regression is performed using exponential functions), which suggests that the compression rates can also indicate the restoration of brain function.

Figure 3 shows the ICR of the H264 lossy compressor on trials across different days. Because H264 is a video compressor, we first sorted the 96-channel single-unit spike signal at each time stamp into a 10*10 black and white

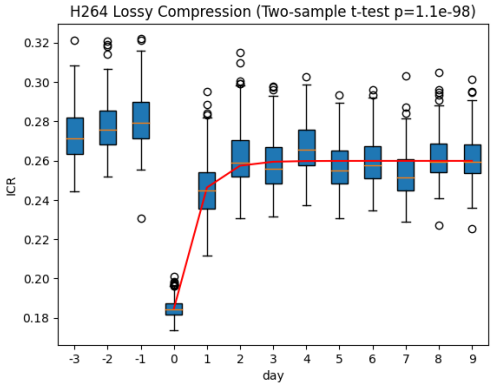


(a) GZIP Lossless ICR

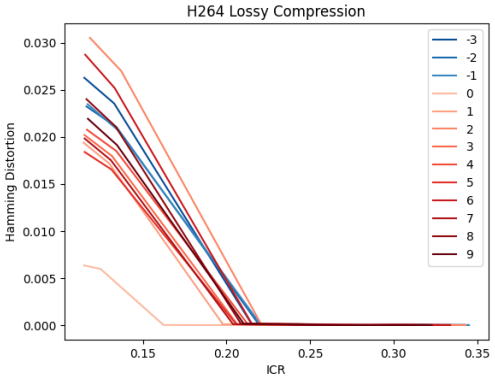


(b) ZSTD Lossless ICR

Figure 2: Lossless ICR plotted across 13 consecutive days. Small electrolytic lesions were performed on Day 0. Days with negative indices are pre-lesion days, and days with positive indices are post-lesion days. The boxplots show the medians and quartiles. The red lines denote the regression functions during the recovery.



(a) H264 Lossy ICR at Compression Level 3



(b) H264 ICR-Distortion Curve

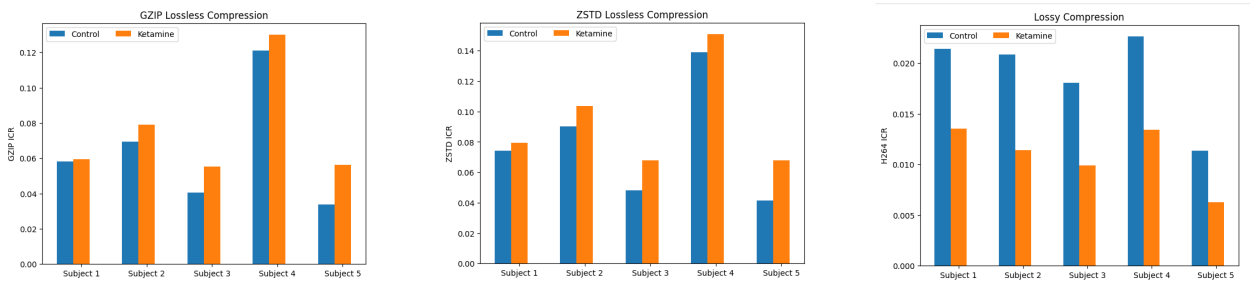
Figure 3: Lossy ICR plotted across 13 consecutive days. (a) The same caption as Figure 2. (b) A systematic sweep across different compression levels. Blue lines denote pre-lesion days, and red lines denote post-lesion days. The figure only shows the mean values of ICR and Hamming distortion at given compression levels and given recording days.

grid chart as one frame, and then converted the frames into videos using H264 coding at different compression levels (Level 1 corresponds to the level with the lowest distortion; Level 6 corresponds to the level with the highest distortion). Examples of pre-lesion spiking video and post-lesion spiking video can be found at Day -1 video and Day 0 video.

Comparing Figure 3a and Figure 2, we found that the results of the lossy compressor and the lossless compressors are similar and consistent. The systematic sweep (Figure 3b) suggests that post-lesion days usually have lower ICR values compared to the pre-lesion days at the same level of distortion, which is consistent with the previous results.

4.2 Mouse Intracortical Probe Data

The mouse data was split into two categories: control, 100 trials before the animal was given ketamine, and ketamine, 100 trials after the animal was administered ketamine. For the lossy compressor, data was organized into a 1D array using depth metadata for each cell detected by the intracortical probe. As shown in Figure 4, we see that the ICR of ketamine trials is slightly higher than control trials in both lossless compressors. Unexpectedly, we see these results flipped in the lossy compressor, where the ICR of control trials are much higher than ketamine trials. A Welch's t-test between the two populations for each compressor give p-values of 0.602 for GZIP, 0.531 for ZSTD, and 0.013 for H264, indicating that the lossless ICR differences are not significant, but the ICR differences of the lossy compressor are significant.



(a) GZIP Lossless ICR (b) ZSTD Lossless ICR (c) Lossy ICR

Figure 4: ICR plotted for five subjects. Control trials are in blue, and ketamine trials are in orange.

This flipping of results may be for a variety of factors, primarily including the action of ketamine on spatial mapping. Firstly, special neurons in the MEC known as grid cells encode a spatial mapping of a given environment. Ketamine disrupts the MEC's ability to form these spatial mappings [5]. This disruption likely removes the temporal coherence of the data (each 'frame' is less similar to the last one), leading to a lower ICR in ketamine trials in lossy compression. While both GZIP and ZSTD rely on repetition in the data, they do not look for temporal coherence the way that a video compressor would, so ketamine trials may not have a strongly lowered ICR in lossless compression.

A systematic sweep of possible lossiness values, controlled by the CRF variable for the H264 compressor, reveals that a CRF value of 20 gives us the most statistically significant differences between ketamine and control trials (p-value = 0.012). The default CRF value in H264 is 23, with 0 being lossless and 51 being lossiest. A value of 20 may therefore represent the ideal trade-off between removing unimportant features and maintaining the temporal structure of the data.

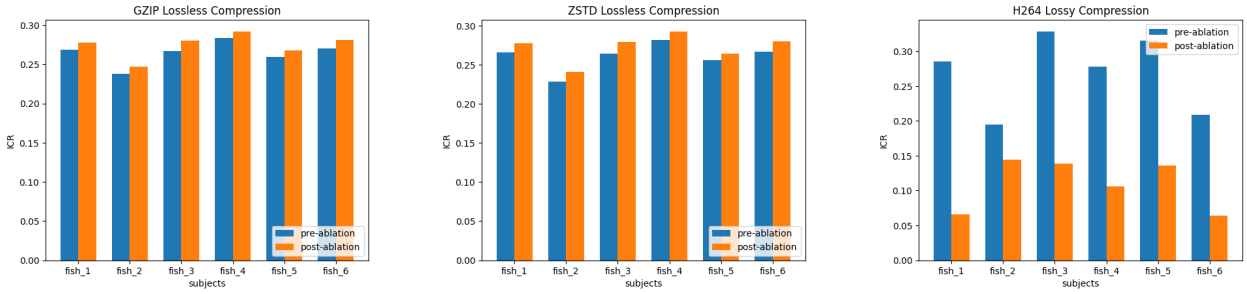
4.3 Zebrafish Calcium Imaging Data

In this session, we performed a similar analysis as the previous section on a zebrafish calcium imaging dataset. The data was divided into two categories: pre-ablation and post-ablation.

As shown in Figure 5, we see similar results to what we observe in Section 4.2: the differences in ICR between post-ablation recordings and pre-ablation recordings with lossless compressors are less significant compared to that of the lossy compressor, and the sign of the differences is also flipped for the two types of compressors.

We attribute this flipping to the similar reasons we discuss in Section 4.2. After the ablation in the spinal cord, the coherence of the neural activities is disrupted. The changes in this kind of property may be better captured by video compressors like H264 than by lossless compressors like GZIP and ZSTD.

Besides the coherence in neural activities, the noise in the nervous systems may also play an important role in contributing to the changes in the performance of the compressors, which we will further discuss in the next section.



(a) GZIP Lossless ICR (b) ZSTD Lossless ICR (c) Lossy ICR

Figure 5: ICR plotted for six subjects. Pre-ablation recordings are in blue, and post-ablation recordings are in orange.

5 Discussion

In this work, we demonstrate the utility of using compression, namely the inverse compression ratio, to detect neural events.

We performed three compression algorithms (two lossless compressors and one lossy compressor) on three datasets across different species and with different recording modalities. In the NHP intracortical single-unit dataset recorded for lesion experiments, we saw consistent drops in ICR with different compressors after the introduction of lesions. In the mouse intracortical probe dataset for ketamine experiments and the zebrafish calcium imaging dataset for ablation experiments, we saw flipped effects in ICR caused by the intervention with different types of compression methods (the lossless compressors show increases in ICR after the intervention, but the lossy compressor shows decreases in ICR after the intervention).

In previous sections, we have argued that the changes in the temporal coherence of the neural activities may contribute to these flipped effects, since H264 captures the temporal coherence of the data better than GZIP and ZSTD. However, we did not observe this flip in our single-unit dataset. Unlike the two other datasets, the single-unit dataset contains binary information on the spike events of the recorded neurons, e.g. 0 for no spike and 1 for firing a spike at a given time stamp. Thus, the single-unit dataset suffers from a lower level of noise compared to the other two datasets. When the noise level is low, the lossless compressors can capture the coherence of the neural activities as well as the lossy compressor, which explains the consistent results of different compression methods in the single-unit dataset. However, when the noise level grows higher, the lossless compressors will compress the increasing noise losslessly as well, which explains the flipped results of different compression methods in the other two datasets. Across different types of intervention (the M1 lesion, the ketamine injection, and the spinal cord ablation), the lossy compressor suggests consistent decreases in ICR after the intervention. Thus, we suspect the increases in ICR with the lossless compressors in the mouse dataset and zebrafish dataset are the results of increasing noise in the nervous systems after the ablation.

In conclusion, our results show that the ICR can be used as an indicator of the function of the brain or other nervous systems across different species, and across modalities. However, based on the nature of different datasets, the ICR of different kinds of compressors might need to be interpreted differently.

6 Future Work and Directions

Future work may choose to explore different compression algorithms. In particular, lossy linear dynamic systems-based compression may prove to be useful due to the established body of work on linear dynamic systems and neural dynamics. Future work may also choose to explore an even wider selection of species and measurement modalities. Finally, future work may choose to explore a single type of event across species/modalities, rather than general events that disrupt the temporal structure of neural activity.

7 Code and Data

Portions of the NHP stroke data is conditionally available upon request from the Brain Interfacing Lab at Stanford University: https://bil.stanford.edu/
The mouse ketamine task data was collected by Dr. Francis Masuda in the Giocomo Lab at Stanford University, and is freely available at this link: https://giocomolab.weebly.com/data-code-methods.html
The zebrafish ablation data was collected by the Keller lab at the Janelia Research Campus, and is available at this link: https://doi.org/10.25378/janelia.7607411.v1
For all code used in this project, the public GitHub repository is available at: https://github.com/Carolyxwu/ee274-final-project

8 Acknowledgements

We would like to thank Professor Paul Nuyujukian and Professor Tsachy Weissman for their guidance on this work. We would also like to thank Dr. Lisa Yamada for additional support and help developing the foundations of this work. Finally, we would like to thank Pulkit Tandon, Shubham Chandak, and Noah Huffman for their quality instruction in EE 274.

9 References

[1] Yamada, L., Nuyujukian, P. H., Nishimura, D. G., Weissman, T. A compression-enabled approach to analyze seizures for people with refractory epilepsy. Stanford University. 2023
[2] Higgins G, et al. The effects of lossy compression on diagnostically relevant seizure information in EEG signals. IEEE J Biomed Health Inform. 2013
[3] Nguyen, B., Ma, W., Tran, D. A study of combined lossy compression and seizure detection on epileptic EEG signals. Procedia Computer Science. 2018
[4] Bray, I., Clarke, S., Casey, K., Nuyujukian, P. Neuroelectrophysiology-Compatible Electrolytic Lesioning. bioRxiv 2022.11.10.516056; doi: https://doi.org/10.1101/2022.11.10.516056
[5] Masuda FK, Sun Y, Aery Jones EA, Giocomo LM. Ketamine evoked disruption of entorhinal and hippocampal spatial maps. bioRxiv [Preprint]. 2023 Feb 6:2023.02.05.527227. doi: 10.1101/2023.02.05.527227. Update in: Nat Commun. 2023 Oct 7;14(1):6285. PMID: 36798242; PMCID: PMC9934572.
[6] Collet, Y., Turner, C. Smaller and faster data compression with Zstandard. Engineering at Meta. 2018, June 28. https://engineering.fb.com/2016/08/31/core-infra/smaller-and-faster-data-compression-with-zstandard/