**SEQUENTIAL STRUCTURE IN NEURAL DATA: BIG DATA CHALLENGES**

**Introduction**

*Encoding position—Place cells*

Populations of neurons in the entorhinal cortex (EC) have been shown to accurately encode information about the animal’s location in the environment, so called “place cells.”

*Sequential activity: replay*

Place cells fire sequentially as animals move around, and these sequences sometimes repeat during sharp wave ripples (SWRs) in the EEG. So what is replay? Motivated, time-compressed neural activity where the firing order is preserved.

*Quality of replay*

Which particular replay event is the “best”? Which ones are “good enough”?

Problem statement

1. How can we identify and understand these replay sequences?
2. How can we learn our models an iteratively of data?

Some considerations:

1. Subsets and all subsets of participating cells unknown
2. Not every subset participates in each sequence
3. Cells are shared (and non-exclusively) to another
4. Ground truth is difficult/impossible to establish
5. We want a quantitative handle on a qualitative phenomenon

Multiple models for different contexts

Many events are simply classified as “noise”—need more models/ ensembles.

**Experiment**

Linear track data from online replay.org

**Graphical abstract**

*Model learning*

- Training models without behavioral correlates
- Template matching and Bayesian decoding need behavioral correlates

*Model validation*

- Latent model states correspond to particular behaviors—improvement of places fields
- Sequence score effectively discriminates between real and shuffled data

*Scoring sequences*

- Scoring sequences difficult to compare two sequences directly using log probability
- Sequence normalization improves probability of long sequences becoming vanishingly small

**Preliminary results**

*Context classification*

Log probability can effectively discriminate between sequences from different environments

*Length normalization*

- Single sequence
- Single sequence
- All sequences

- Log probability

**Conclusion**

We are interested in unsupervised sequence and context identification—to identify replay along with its various contexts.

Hidden Markov models seem like an attractive and powerful approach, although the hierarchical and non-parametric Bayesian variants which may allow for dynamic context discovery and more efficient learning

However, working with big data sets, little to no ground truth, makes such an undertaking especially challenging, and other approaches may be considered as well.

References


Funding: This work was funded by an NSF CAREER award (1752106) and an NEH RESEACH GRANT (1539915).