Delays in representation of positional information in the EC-hippocampus circuit

Vikrant Jaltare

February 6, 2025

Overview

1. Overview of Skaggs Information
    - Information per second and per spike from a single neuron
2. Highlights from Dubius et al. *PNAS* (2013)
    - Expression of *Hb* gap gene in *Drosophila* embryos
3. Understanding Bias in computing the entropy
4. Mathematical Framework for analyzing hippocampal data
    - Mathematical Framework
    - Mutual information in the circuit
5. Revisiting Skaggs MI analysis

Information between spiking and location from one neuron

- The information for a given neuron in terms of its firing rate $\lambda(x)$ and probability density of the rat visiting a location x i.e. "occupancy probability" is given as:

$$I = \int_x \lambda(x) p(x) \log_2 \frac{\lambda(x)}{\bar{\lambda}} dx$$

  Here, I is in bits/s.

- Similarly, if I is normalized by the average firing rate $\bar{\lambda} = \int_x \lambda(x)p(x)dx$ then we get information in bits/spike. [1]

Delaying spiketrains: sign convention I

The spiketrain is assumed to be composed of a series of Dirac delta functions. Let **s(t)** denote the raw spiketrain from data. Then **s(t)** is a binary vector with elements $s_i^{(k)} \in \{0, 1\}$ where i stands for index of a neuron and k is the time bin.

$$\mathbf{s}_i(t) = \sum_{k=1}^{N} \delta(t - t_k)$$

where N is the length of spiketrain. In this analysis $N =$ task duration in seconds.
To see if the spiketrain is aligned with location-stimulus (will refer to as 'stimulus'), I performed the following operation

$$\mathbf{s}_i^* = \Theta(t - \tau)\mathbf{s}_i$$

$$\therefore \mathbf{s}_i^* = \Theta(t - \tau) \sum_{k=1}^{N} \delta(t - t_k)$$

Delaying spiketrains: sign convention II

Where $\Theta(.)$ is the Heaviside step function.
Thus, if:

1. $\tau < 0$, the spiketrain will be left shifted w.r.t. stimulus.
2. $\tau > 0$, the spiketrain will be right shifted w.r.t. stimulus.
3. $\tau = 0$, the spiketrain is aligned to the stimulus.

Shortcomings of this approach

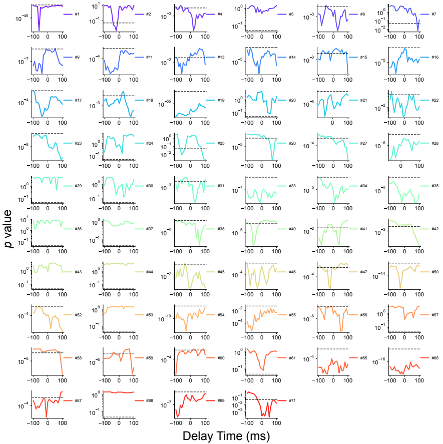


Figure: p values for each neuron. $p < 0.05$ indicates significant information

- There is a large variability in the information from individual neuron.
- Also, we do not have a handle on the systematic bias we introduce when inferring the underlying probability distributions using frequencies.

Shortcomings of this approach

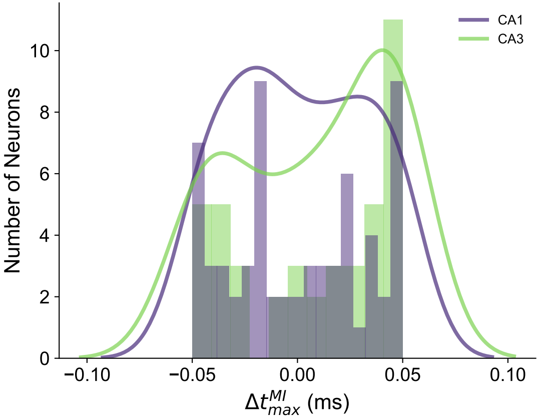


Figure: Caption

- The distribution of delay times does not decay towards the ends.

MI for CA3 neurons

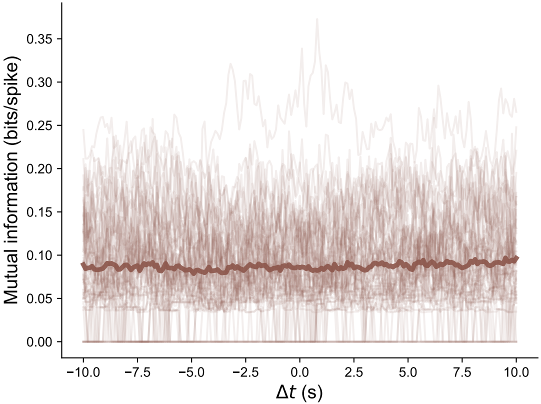
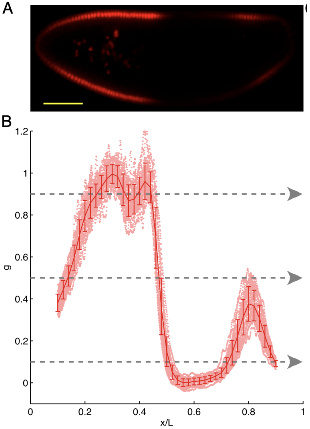


Figure: Mutual information for CA3 cells from *gor01*. The thick curve is the average Mutual information across all neurons.

Expression of *Hb* gap gene in *Drosophila* embryos



- We can construct the distributions $P(g)$ and $P(g, x)$ from this data.

Figure: Expression of Hb gene in drosophila embryo

Expression of *Hb* gap gene in *Drosophila* embryos

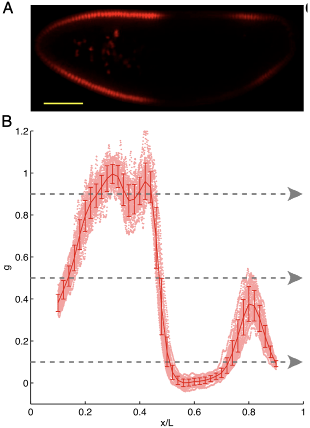


Figure: Expression of Hb gene in drosophila embryo

- We can construct the distributions $P(g)$ and $P(g, x)$ from this data.
- The mutual information between gene expression and location along the anterior-posterior axis is given as

$$I(g; x) = \int_x P(x)(S(P(g)) - S(P(g|x)))$$

Expression of *Hb* gap gene in *Drosophila* embryos

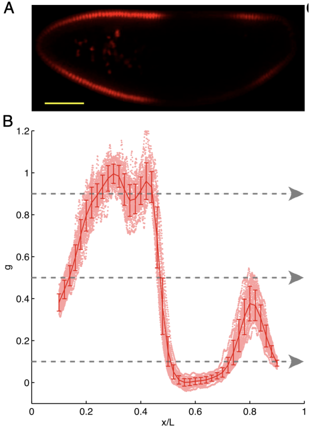


Figure: Expression of Hb gene in drosophila embryo

- We can construct the distributions $P(g)$ and $P(g, x)$ from this data.
- The mutual information between gene expression and location along the anterior-posterior axis is given as

$$I(g;x) = \int_x P(x)(S(P(g)) - S(P(g|x)))$$

- The authors argue that the intermediate expression levels, as opposed to gene being ON/OFF, provides more information – nearly 2 bits for a group of gap genes – about location. For ON/OFF genes the maximal information would be 1 bit.

Information given by multiple genes

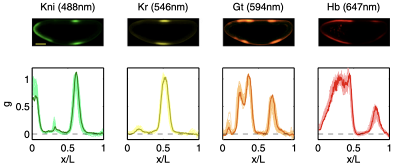


Figure: Expression of different gap genes in drosophila embryo

For considering many genes together, the authors performed immunofluorescence staining of multiple genes as shown. This data is good enough to construct $P(g_i)$. They then approximate $P(g_i|x)$ as gaussian to yield the following

$$P(\vec{g}|x) = \frac{1}{\sqrt{(2\pi)^{\dim(\vec{g})}}} \exp\left(F(\vec{g})\right)$$

Where,

$$F(\vec{g}) = \frac{-1}{2}(\vec{g} - \vec{\bar{g}})^T C_{ij}^{-1}(\vec{g} - \vec{\bar{g}})$$

Understanding Bias in computing the entropy

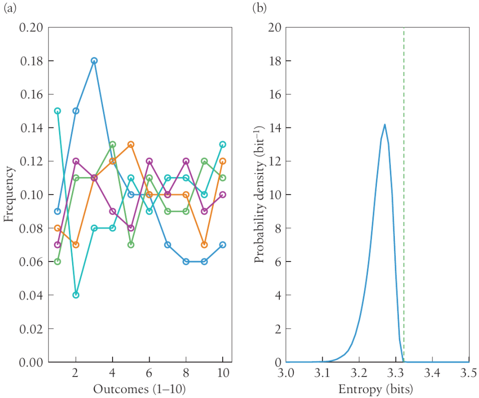


Figure: Bias in determining entropy. From Biophysics: Searching for principles, Bialek.

- The left figure shows $N = 100$ random samples drawn from a uniform distribution over $0 - 10$. There are $K = 10$ bins.
- The distribution does not look flat due to significant fluctuations (order of $\frac{1}{\sqrt{10}}$).
- On the right, the green line is the true entropy of this distribution ($-0.1 \log_2 0.1$). And the blue curve is the estimated entropy for different probability densities (is this $-\log_2 P_i$?)
- The fluctuations add to zero as we add more data, but as long as we estimate probabilities using frequency of occurrence, we'll introduce bias in calculating entropy.

Quantifying bias as a function of data and number of bins I

Let S_{naive} be the entropy estimate from calculation using frequencies f_i. If we draw N samples with probabilities p_i, and if n_i samples have outcome i, then $< n_i >= Np_i$

$$S_{naive} = -\sum_{i=1}^{K} f_i \log_2 f_i$$

$$\therefore S_{naive} = -\sum_{i=1}^{K} (p_i + \delta f_i)(p_i + \delta f_i)$$

Using Taylor expansion,

$$S_{naive} = -\sum_{i=1}^{K} p_i \log_2 p_i - \sum_{i=1}^{K} \left(\log_2 p_i + \frac{1}{\ln 2} \right) \delta f_i - \frac{1}{2} \sum_{i=1}^{K} \left(\frac{1}{p_i \ln 2} \right) (\delta f_i)^2 - \dots$$

Quantifying bias as a function of data and number of bins II

After simplification, we get

$$\langle S_{naive} \rangle = S_{true} - \frac{K}{(2\ln 2)N} - \dots$$

Where K is number of bins (number of accessible states) and N is number of sample data points.

Mathematical Framework

Assume that all the regions (CA3, CA1 and EC) have rate code. Let, $r_i(x)$ be the firing rate of neuron i at a given location x. Note that the firing rate $r_i(x)$ is normalized by the maximum firing rate for a given neuron. We construct the following distributions

- $P_k(\{r_i\})$: probability density of firing rate for k out of N neurons. Note that, for CA3, we have about 70 pyramidal cell clusters and for CA1 we have about 60 pyramidal cell clusters.
- $P_k(\{r_i\}, x)$: joint probability density of firing rate and location for k out of N neurons.
- $P(x)$: probability density of the rat/animal being at a location x on the track. Note that x is circularized.
- These distributions can be easily computed from the datasets we have.

Finding information between $r(x)$ and x

We can then find information by simply using the KL-divergence definition as

$$I_k(r;x) = \sum_{i \in x} \sum_{j \in r} P_k(r=j, x=i) \log_2 \frac{P_k(r=j, x=i)}{P(x=i) P_k(r=j)}$$

With addition of more neurons, we get more datapoints to construct the histogram, we should get increasingly reliable reading for information as we keep adding more data.

Withholding data to quantify the bias in finding mutual information I

Sampling bias and information

Entropy scales proportionally to accessible states (K) inversely as the number of samples (N). Thus, in turn information too can be written as

$$I_N(r;x) = I_\infty(\Delta x) + \frac{A(\Delta(x))}{N} + \frac{B(\Delta(x))}{N^2} + \cdots$$

- I withheld different fractions of total number of pyramidal neurons from CA3 and CA1 for this analysis, from 80% to 96%.
- For each of the fractions, I took random samples of neurons, found their respective distributions (rate, location and joint of rate and location) and computed mutual information.
- This gives a trend. I then fitted the information to a line (note that x axis is $\frac{1}{N}$ i.e. inverse of number of neurons).

Withholding data to quantify the bias in finding mutual information II

- The y - intercept of this line (I_∞) gives the information estimate when theoretically, we have infinite number of neurons (like the population of neurons).
- I_∞ should be ideally free from sampling and binning bias.

Place Cells from HC3 Dataset

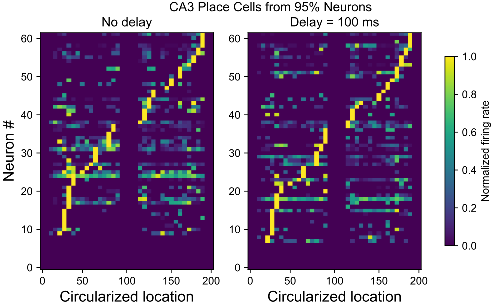


Figure: CA3 Place Cells for 62 neurons from *gor01*. This is 95% of the data.

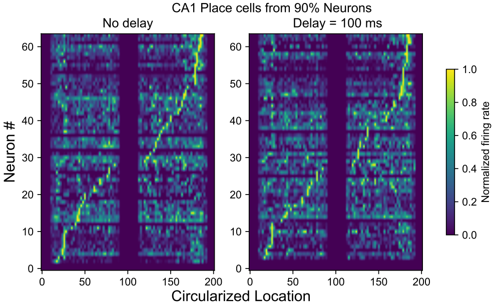
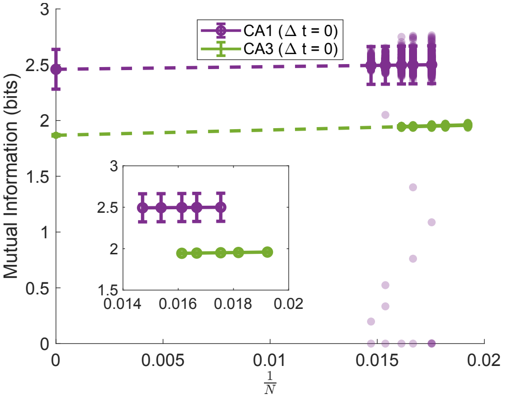


Figure: CA1 Place Cells for 64 neurons from *ec014*. This is 90% of the data.

Mutual information in CA3 and CA1 neurons

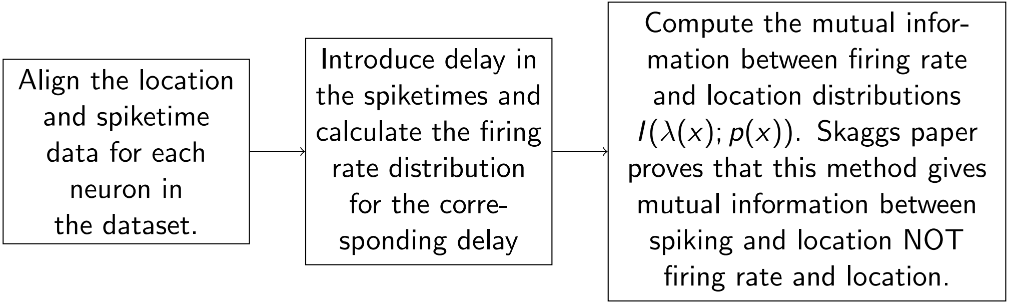


Drawbacks of this approach

- The Strong and Bialek [2] approach requires defining a neural word.
- This approach groups all the firing rates as an encoding of location. However, location is encoded in the sequence of neural activation.
- So this approach will only work if the code is a vector indicating state of neurons for a given location, potentially over time.
- The above analysis does not address this question.

Revisiting Skaggs Information-based analysis

The workflow used in Skaggs analysis is as follows:

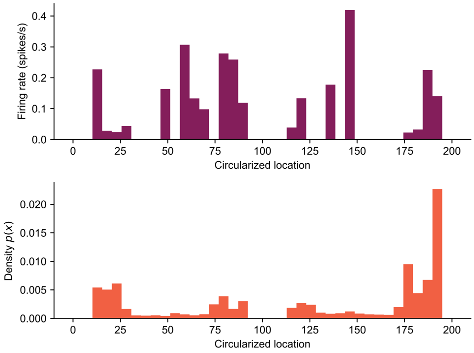


Align the location and spiketime data for each neuron in the dataset.

Introduce delay in the spiketimes and calculate the firing rate distribution for the corresponding delay

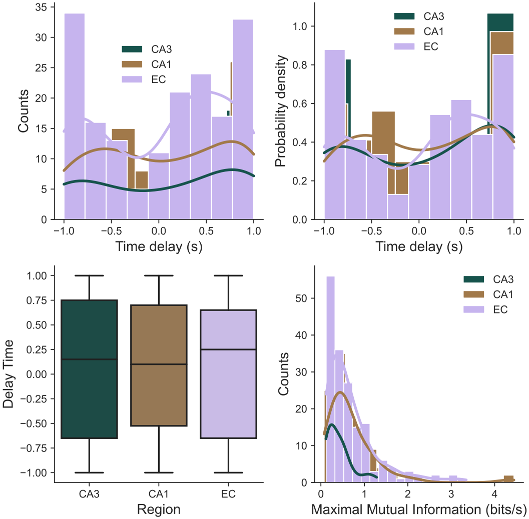
Compute the mutual information between firing rate and location distributions $I(\lambda(x); p(x))$. Skaggs paper proves that this method gives mutual information between spiking and location NOT firing rate and location.

Firing-rate and location plots: Bimodal



- The firing rate (max) for most of the neurons in the dataset is usually below 1 Hz. As such, the code filers out these neurons.
- The distribution $p(x)$ usually has peaks around the extremeties of the track. Note that I considered only central 80% of the track to ignore the endpoint effects. The speed threshold is 4 cm/s.
- As we'll see further there is surprising correlation between $p(x)$ and the distribution of delay times.
- I'm not yet sure, why this correlation exists.

Delay time distribution



- The delay time distributions also display a bimodal distribution – there is a curious correlation with $p(x)$.
- Surprisingly after aggregating the results it seems that the delays are positive, i.e. neurons firing ahead of the stimulus (?)

Summary statistics

- The analysis was done for top 10 sessions with most number of cells in each of CA3, CA1 and EC.
- Cells with peak firing rate below 1 Hz were removed from the analysis.

Region	Median Delay (s)
EC	+0.25
CA3	+0.15
CA1	+0.1

Region	No. viable cells
EC	174
CA3	59
CA1	107

References

1. Skaggs, W. E., McNaughton, B. L., Gothard, K. M. & Markus, E. J. *An information-theoretic approach to deciphering the hippocampal code.* in *Proceedings of the 5th International Conference on Neural Information Processing Systems* (Morgan Kaufmann Publishers Inc., Denver, Colorado, Nov. 1992), 1030–1037.
2. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **80,** 197–200 (Jan. 1998).