Lecture 3: Long Timescales & Hierarchy In Behavior

Last Time: Went from Images → Time series

→ Low-d embedding

→ Notion of behavioral space, with peaks & valleys, with peaks corresponding to stereotyped behaviors, & ~1/2 of all time spent doing stereotyped movements

→ Show space & stereotypy plots

Today: Talk about how we can use this representation to learn things about behavior, particularly how the animal moves within this space & how long times emerge

First: Two types of measurements

1) Repertoire (time-independent)

2) Sequence (time-dependent)

Slides → Show male/female, D/V example

But the main topic today will be about how long time scale structures & patterns emerge in behavior & how we can characterize and/or model them.

Slides → Map → Transitions in Flies

→ Suggests a Markov Model: $T_{ij} = p(X_{t+1} = i \mid X_t = j)$

→ Very common in behavior literature (70s/80s) (not much data) → Variants still used commonly

---

Can we take this seriously?

Idea: Test for memory in the data

$T(\tau) \equiv T^\tau$ → transition Matrix between t & $t+\tau$

$T_{ij}(\tau) = p(X_{t+\tau} = i \mid X_t = j)$ → can measure from data

if Markov → $= \sum_k p(X_{t+\tau} = i \mid X_{t+\tau-1} = k)\, p(X_{t+\tau-1} = k \mid X_t = j)$

$\vdots$

matrix multiplication

$\sum_{\{k\}} p(X_{t+\tau} = i \mid X_{t+\tau_1} = k_1) \cdots\cdots (X_{t+1} = k_{\tau-1} \mid X_t = j)$

$= T(1)\, T(1) \cdots T(1) = T(1)^\tau$

(2)

Perron – Frobenius Theorem:

For a transition matrix that is fully-connected

$\lambda_1 = 1, 1 > |\lambda_2| \geq |\lambda_3| \geq |\lambda_4| \cdots\cdots \geq |\lambda_N|$ (eigenvalues of T)

→ corresponding eigenvector ($\hat{v}_1$) ~ steady-state distribution

→ $\hat{V}_2, \hat{V}_3, \ldots$ represent slow → fast decay directions

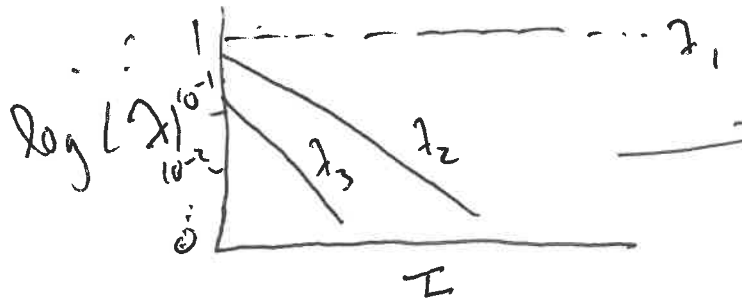
---

Markov: $T(1) = \sum_k \lambda_k U_k^{(1)} V_k^{T}(1)$ (left & right eigenvectors)

$T(1)\hat{V}_k = \lambda_k \hat{V}_k$

$T(\tau)\hat{V}_k = \lambda_k T(\tau-1)\hat{V}_k$

→ $T(\tau) = T(1)^{\tau} = \sum_k \lambda_k^{\tau} U_k(1) V_k^{T}(1)$



→ exponential decay w/ Markovian

In practice, because of finite data, use Shannon shuffling trick

1, 2, 5, 4, 1, 3, 2, 5, 4, 3, 1, ...

1) Pick random starting point 5
2) Find another random point in the data at this state
3) Take next & Add to list 5 4
4) Repeat 5 4 1 2 ....

Can we use the notion of predictability to make sense out of long timescale data?

↳ Information Bottleneck:

$$X(t) \longrightarrow Z(t) \longrightarrow X(t+\tau)$$

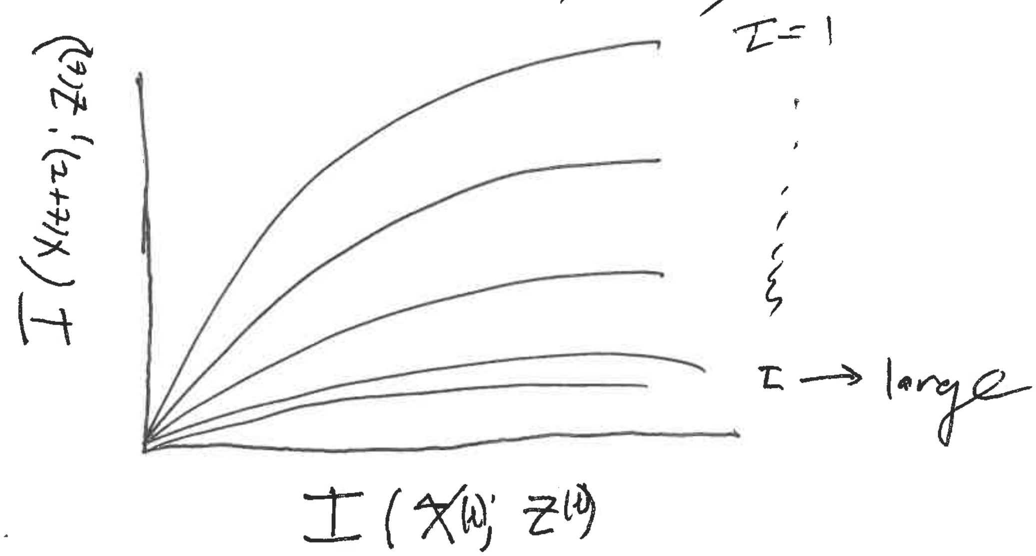
$Z(t)$ → coarse-grained representation

$\underline{I(X(t); Z(t))}$ ← Minimize

$\underline{I(X(t+\tau); Z(t))}$ ← Maximize

$$p^*(Z|X(t)) = \underset{p(z|x(t))}{\arg\max}\; I(Z(t); X(t+\tau)) - \beta I(X(t); Z(t))$$

↳ Solve iteratively using Blahut-Arimoto



Blahut - Arimoto

$$J = I\big(\overset{Y}{x(t+\tau)}\,;\,\overset{Z}{z(t)}\big) - \beta\, I\big(\overset{X}{x(t)}\,;\,\overset{Z}{z(t)}\big)$$

1) $$p(z|x) = \frac{p(z)}{Z(\beta,x)} \exp\left[-\beta D_{KL}\left(p(y|x) \,\|\, p(y|z)\right)\right]$$

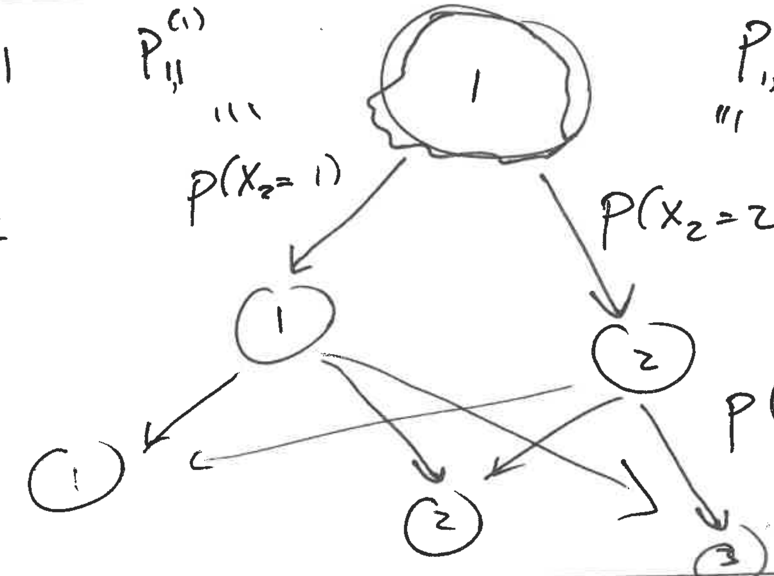
2) $$p(z) = \sum_x p(z|x)\,p(x)$$

3) $$p(y|z) = \sum_y p(y|x)\,p(z|x)\,p(x)$$

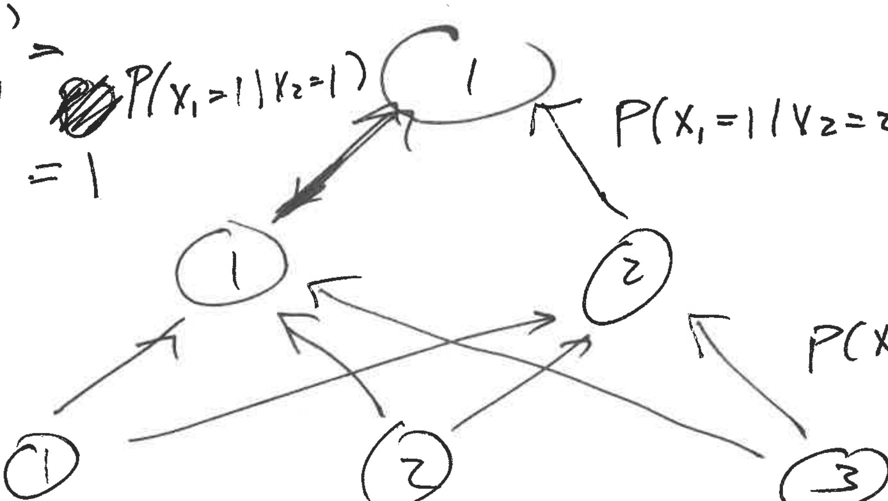
<u>Assume:</u> $p(x,y)$ Known

How to measure hierarchy?

→ Think about "treeness"



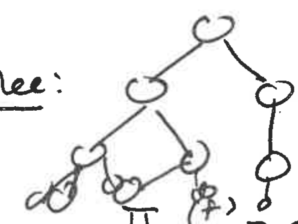
$\mathcal{H}_f$ = entropy of paths



$\mathcal{H}_b$ = entropy of paths

$$\Upsilon = \frac{\mathcal{H}_f - \mathcal{H}_b}{\mathcal{H}_f}$$

w/ Tree:

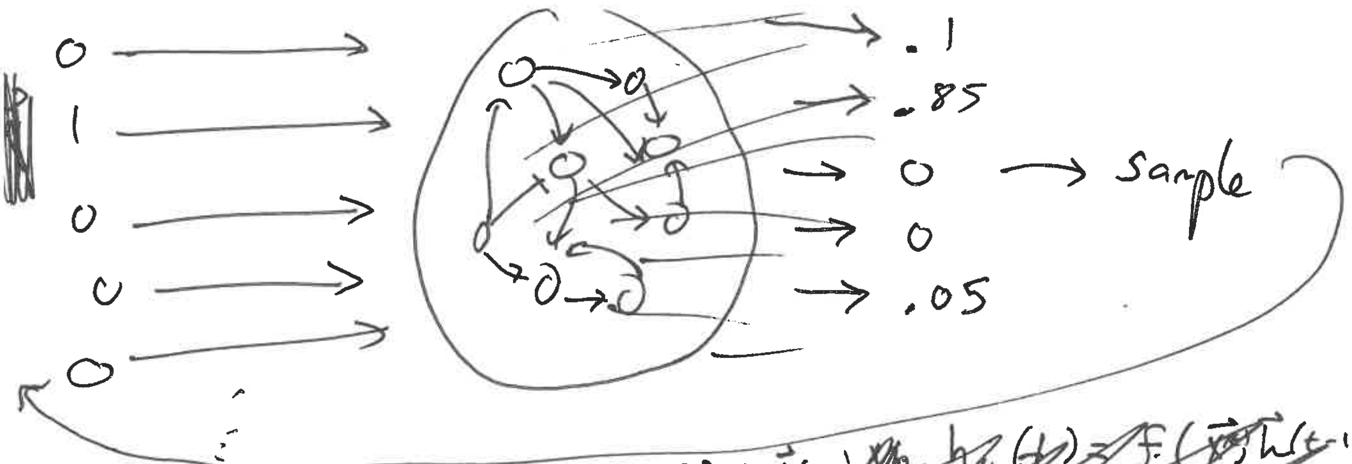


$\Pi_v^{(f)} = p^{(1)}_{11}\, p^{(2)}_{12}\, p^{(3)}_{22}$

$$\mathcal{H}_f = -\sum_{v \in V} p(\Pi_v^{(f)}) \log p(\Pi_v^{(f)})$$

$$\mathcal{H}_b = -\sum_j Q_j^{(end)} \cdot \sum_{w \in W_j} p(\Pi_w^{(b)}) \log p(\Pi_w^{(b)})$$

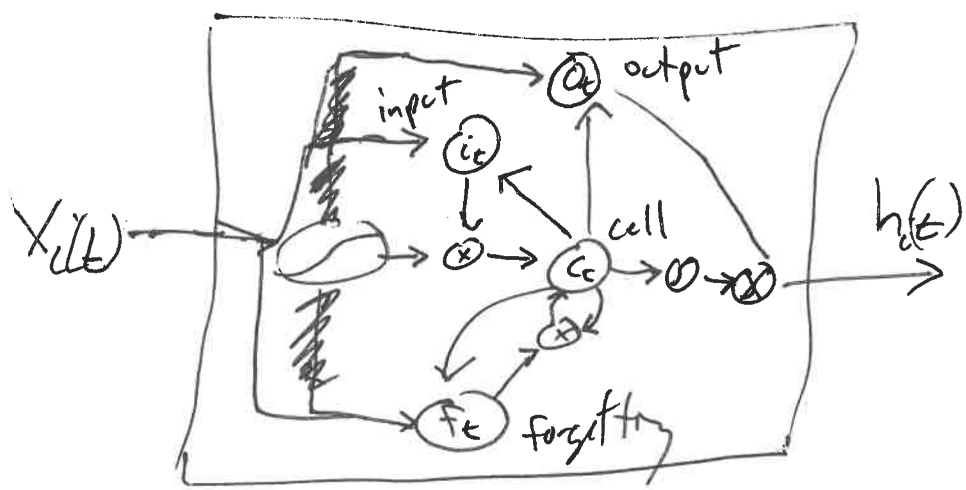
Long Short-Term Memory



$h_i(t) = f(\vec{x}(t), \vec{h}(t-1), p_i)$ $\cancel{h_i(t) = f_i(\vec{x}, \vec{h}(t-1))}$ $\cancel{x_i(t+1) = g_i(\vec{h}(t))}$

$\vec{x}_j(t+1) = g(\vec{h}(t))$

ReLU $\rightarrow$ $h_i = \begin{cases} 0 & \text{if } \vec{W}^{(i)} \cdot \vec{X}_i + b_i < 0 \\ \vec{W}^{(i)} \cdot \vec{X}_i + b_i & \text{o.w.} \end{cases}$



input $\rightarrow$ controls flow

output $\rightarrow$ controls output value

forgetting $\rightarrow$ sets time scale of integration

GRU $\rightarrow$ no output gate

Tuning Field

$F_{ij} = E[h_i(t) \mid B=j]$