Time series analysis:
Markov and Hidden Markov processes

Banyuls school on « Advanced Computational Analysis for Behavioral and Neurophysiological Recordings »

G. Debrégeas and R. Monasson

1. Random walks and Markov processes
2. Bayesian inference of transition rates
   illustration: Inference of diffusion coefficient
3. Hidden Markov Models and the 3 basic questions
   Illustration: The Cookie Monster problem
   Decoding of trajectories from HPC activity

Markov dynamical processes

Markov process = **stochastic dynamical process without memory**

Markov dynamical processes

Markov process = **stochastic dynamical process without memory**

Example: want to model weather dynamics, day after day
intrisically stochastic!

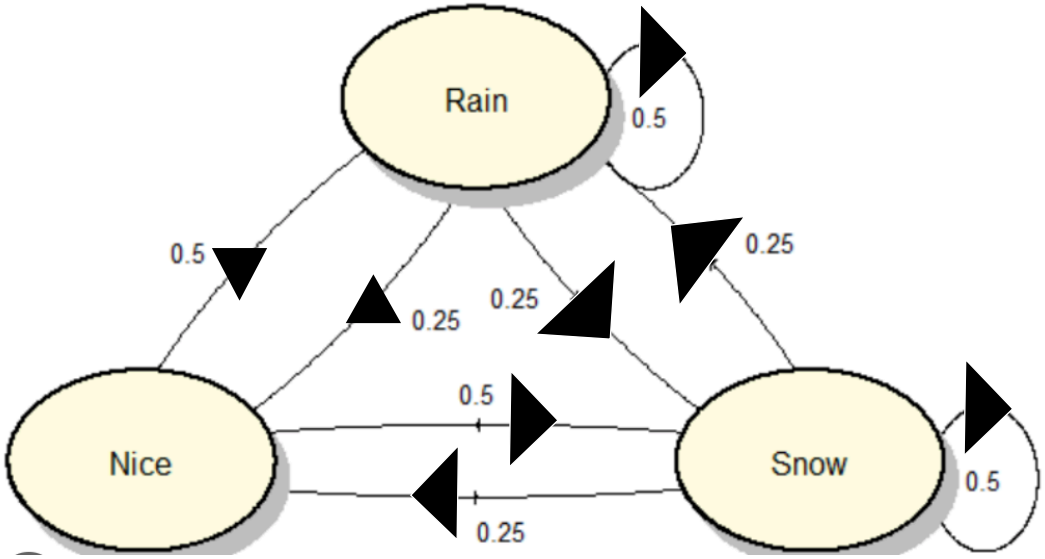
first, identify possible states, e .g. weather(day *n*) = nice, rain, snow
second, define transition probabilities from weather(n) to weather(n+1)

Markov dynamical processes

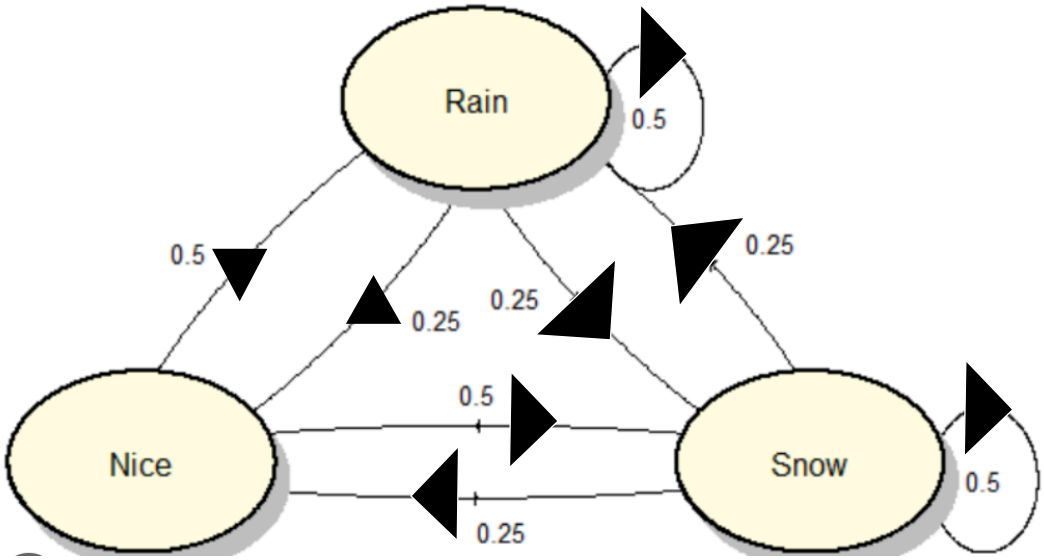
Markov process = **stochastic dynamical process without memory**

Example: want to model weather dynamics, day after day
intrisically stochastic!

first, identify possible states, e .g. weather(day *n*) = nice, rain, snow
second, define transition probabilities from weather(n) to weather(n+1)



Markov dynamical processes



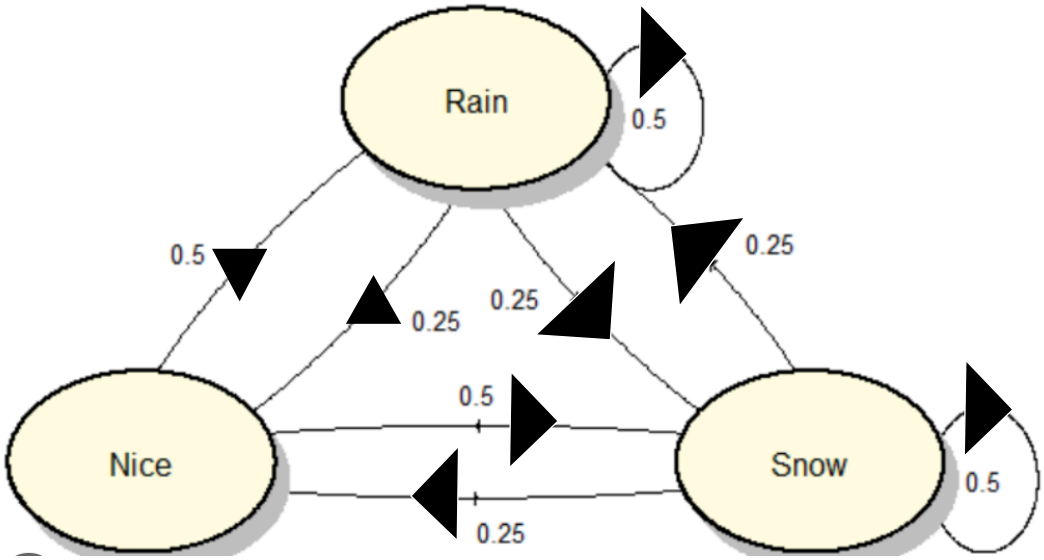
Transition probability = conditional probability of w(n+1) given w(n), e.g.

$$\Omega(nice \mid rain) = 0.25$$

Be careful here!! $\Omega(w' \mid w)$ means $\Omega(w \rightarrow w')$

Of course: $\Omega(w' \mid w) \geq 0, \sum_{w'} \Omega(w' \mid w) = 1$

Markov dynamical processes

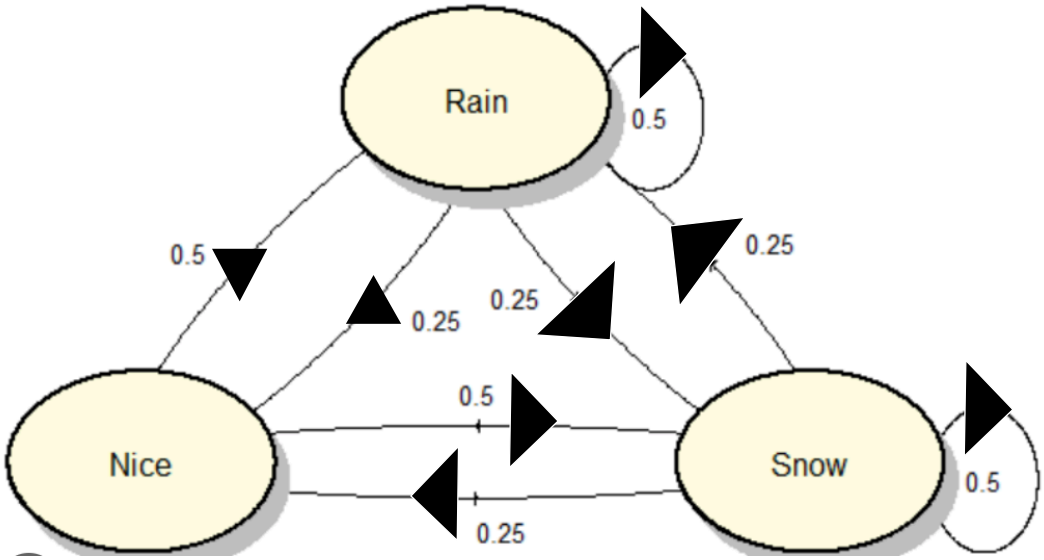


Fundamental question: if I wait long enough, what is going to happen?

Answer: under very general conditions, states will be visited with stationary probability

$$P_{stat}(w) \geq 0, \; such \; that \; P_{stat}(w') = \sum_{w'} \Omega(w' \mid w) \times P_{stat}(w)$$

Markov dynamical processes



Fundamental question: if I wait long enough, what is going to happen?

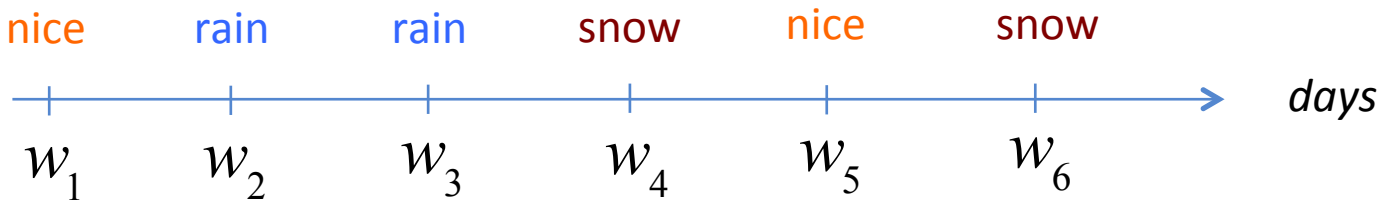
Answer: under very general conditions, states will be visited with stationary probability

$$P_{stat}(w) \geq 0, \; such \; that \; P_{stat}(w') = \sum_{w'} \Omega(w' \mid w) \times P_{stat}(w)$$

Exercise 4: *what is $P_{stat}(w)$ here?*

Markov dynamical processes: inference

Goal: infer the transition probabilities from the observation of one (or more) trajectory



- What are the values of $\Omega(w' \mid w)$?
- How accurate is the inference depending on the number of observations?

Bayesian framework for Markov process

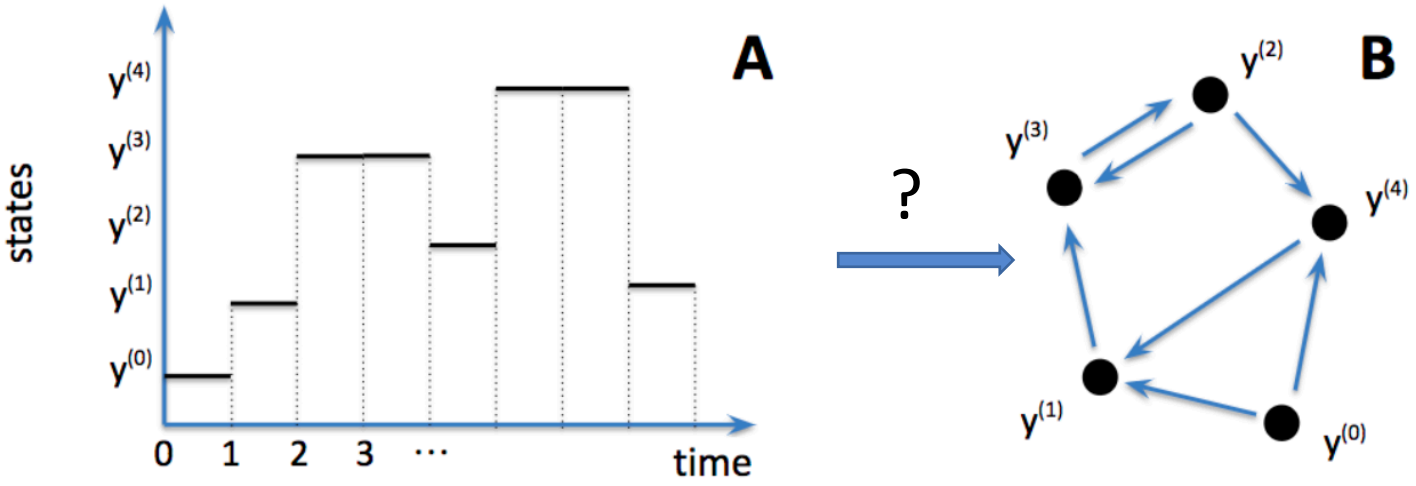
We consider a stochastic dynamical process for the evolution of the system, in which the conditional probability to be in a state w_{t+1} at time t + 1 only depends on the state w_t at time t.

Observations: $\{w_0, w_1, w_2, w_3, \ldots, w_T\}$

Likelihood of trajectory: $P\left(\{w_1, w_2, w_3, \ldots, w_T\} \middle| w_0\right) = \prod_{t=1,\ldots,T-1} \Omega\left(w_{t+1} \middle| w_t\right)$

Prior over transition probabilites? We know that $\Omega(w' \mid w) \geq 0, \sum_{w'} \Omega(w' \mid w) = 1$

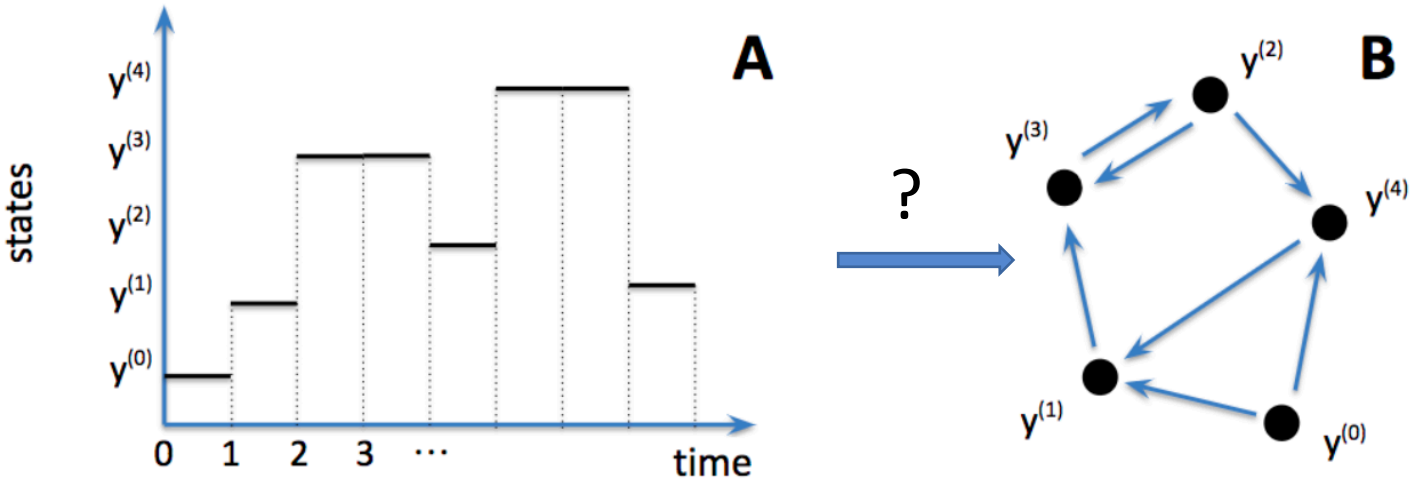
Inference of the transition matrix



$$P\left(\left\{w_1, w_2, w_3, \ldots, w_T\right\} \middle| w_0\right) = \prod_{t=1,\ldots,T-1} \Omega\left(w_{t+1} \middle| w_t\right) = \prod_{y^{(a)}, y^{(b)}} \Omega\left(y^{(b)} \middle| y^{(a)}\right)^{N\left(y^{(a)} \to y^{(b)}\right)}$$

Where $N\left(y^{(a)} \to y^{(b)}\right)$ is the number of transitions from $y^{(a)}$ to $y^{(b)}$ along the trajectory.

Inference of the transition matrix



The maximum likelihood estimator (MLE) for the Ω matrix is obtained by maximising the log-probability of the trajectory under the constraint that the sum of elements along each row must be equal to one:

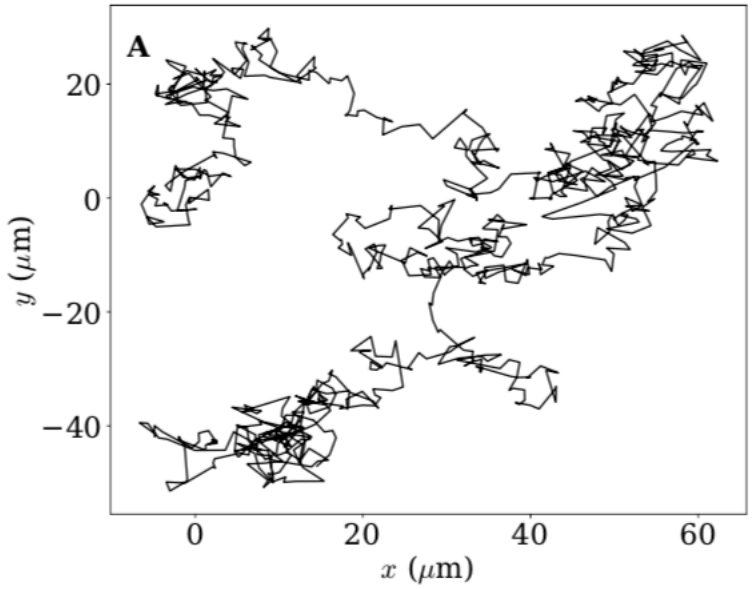
$$\Omega^{MLE} = \underset{\Omega}{\operatorname{argmax}} \left\{ \sum_{y,y'} \mathcal{N}(y \to y') \, \log \Omega(y \to y') - \sum_{y} \lambda(y) \left(\sum_{y'} \Omega(y \to y') - 1 \right) \right\}$$

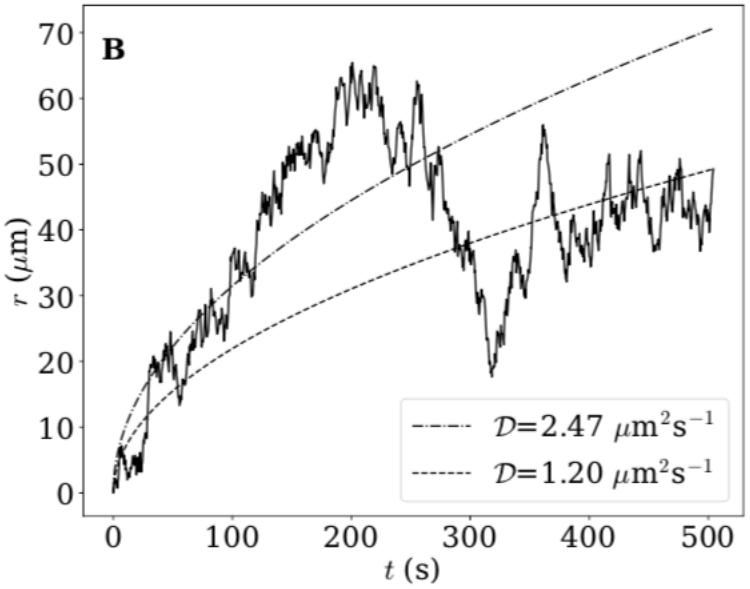
$$\frac{\mathcal{N}(y \to y')}{\Omega^{MLE}(y \to y')} = \lambda(y) \quad \Longrightarrow \quad \Omega^{MLE}(y \to y') = \frac{\mathcal{N}(y \to y')}{\sum_{z} \mathcal{N}(y \to z)} \, .$$

One can add regularizations: eg. pseudo-counts for non observed transitions

Application: global inference of a single parameter

We consider a particle undergoing diffusive motion in the plane, with position $\boldsymbol{r}(t) = (x(t), y(t))$ at time t. The diffusion coefficient (supposed to be isotropic) is denoted by $\mathcal{D}$, and we assume that the average velocity vanishes. Measurements give access to the positions (x_i, y_i) of the particles at times t_i, where i is a positive integer running from 1 to M.





Application: global inference of a single parameter

Displacements between successive times:

$$\delta x_i = x_{i+1} - x_i \ , \qquad \delta y_i = y_{i+1} - y_i \ , \qquad \delta t_i = t_{i+1} - t_i \ .$$

Likelihood:

$$p(\{\delta x_i, \delta y_i\}|\mathcal{D}; \{\delta t_i\}) = \prod_{i=1}^{M-1} \frac{1}{4\pi\mathcal{D}\,\delta t_i} e^{-\frac{\delta x_i^2}{4\,\mathcal{D}\,\delta t_i} - \frac{\delta y_i^2}{4\,\mathcal{D}\,\delta t_i}}$$

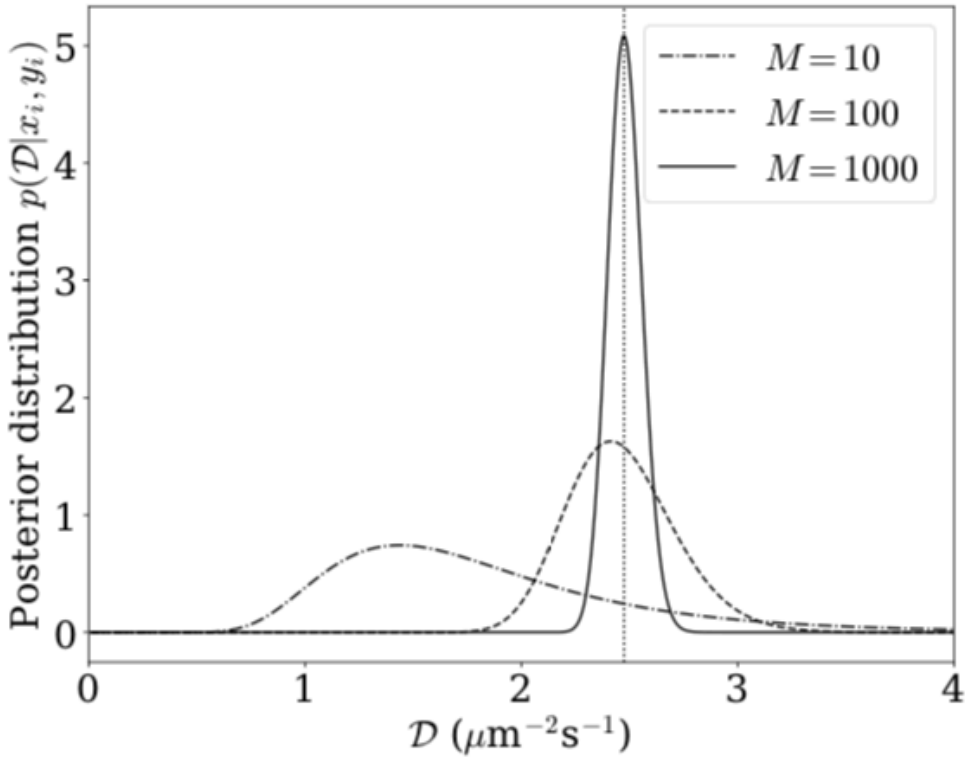
Prior? Flat over $\mathcal{D}$ ...

Bayes' in

$$p(\mathcal{D}|\{\delta x_i, \delta y_i\}; \{\delta t_i\}) = \frac{p(\{\delta x_i, \delta y_i\}|\mathcal{D}; \{\delta t_i\})p(\mathcal{D})}{\int_0^\infty \mathrm{d}\mathcal{D}' \, p(\{x_i, y_i\}|\mathcal{D}'; \{\delta t_i\})p(\mathcal{D}')}$$

Application: global inference of a single parameter

Posterior distribution:



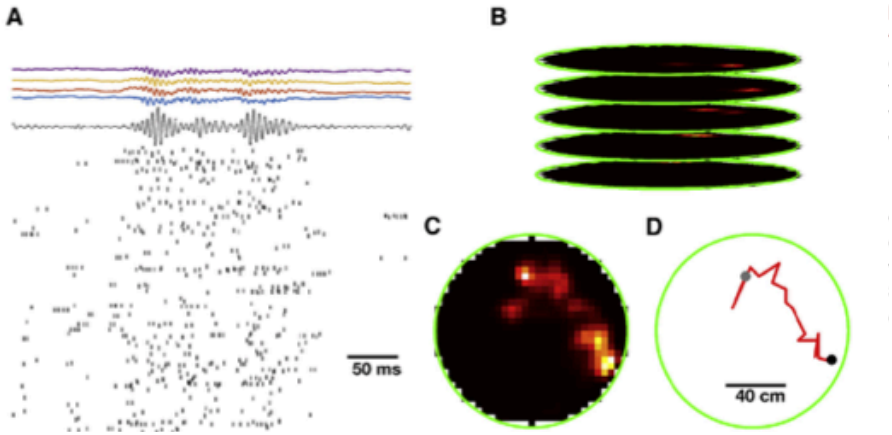
Cocco, Zamponi, RM
From Statistical Physics to Data-Driven Modelling

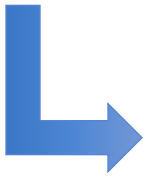
General statement:

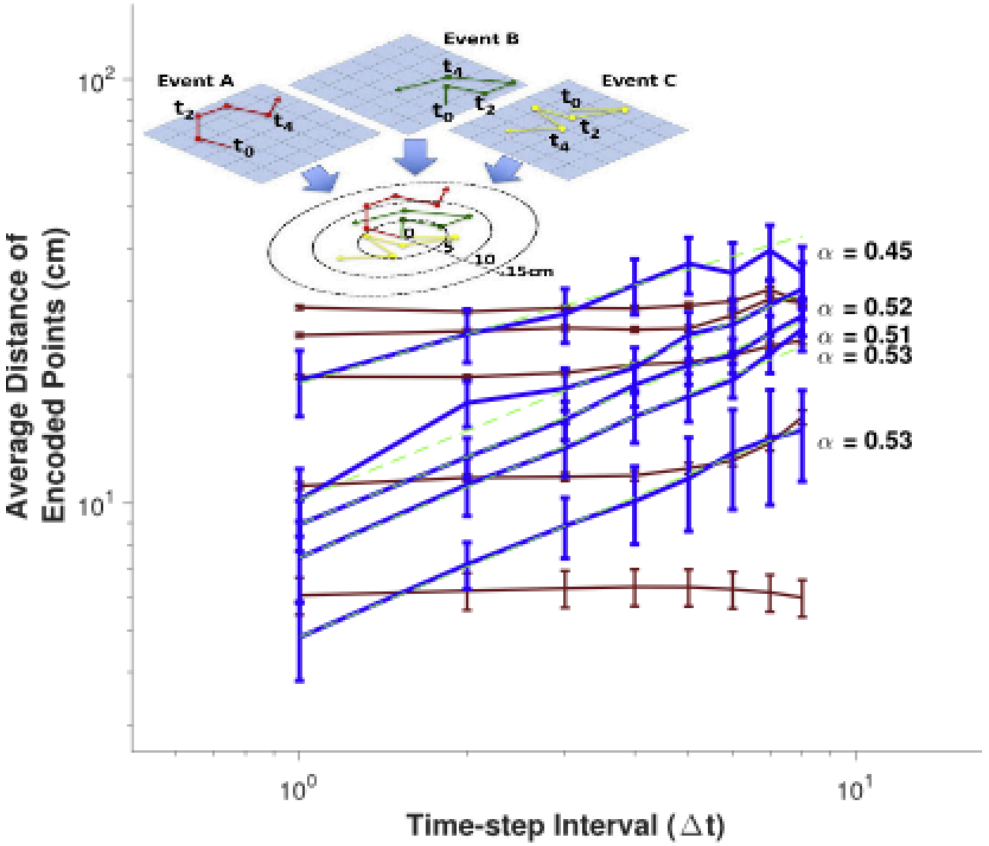
Statistical uncertainty over inferred parameter (here, $\mathcal{D}$) decays with the number M of data as $M^{-1/2}$

Virtual trajectories during sleep and diffusion

Stella et al., Neuron (2019)







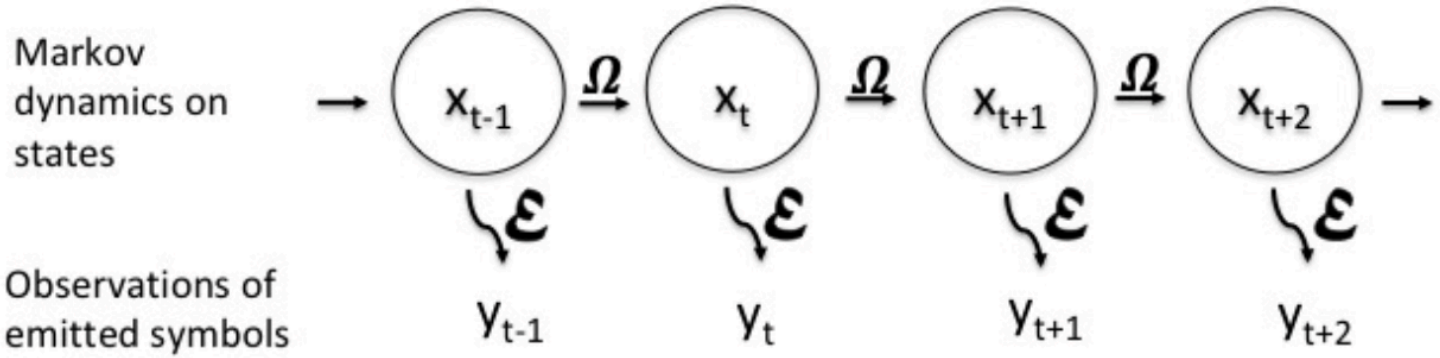
Virtual trajectories are compatible with Brownian diffusion ⋯

Hidden Markov Models (HMM)

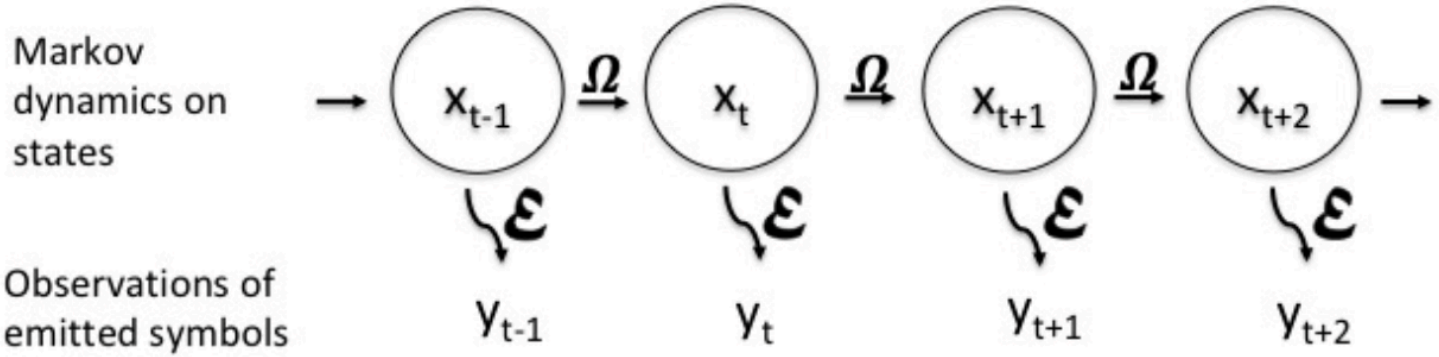
In many situations, dynamical states are not directly accessible:

- Observations are corrupted by measurement noise or finite resolution
  (ex: position of particle is not perfectly measured)

- Observations depend on internal states of the system under study, but do not characterize them completely
  (ex: behavioral states can be organized in patterns/sequences, with transitions between these patterns = internal states)



Hidden Markov Models (HMM)



A HMM is defined by:

- the set of possible states x(i), i = 1,...,D and the D×D transition matrix from one state to another, Ω(x(i) → x(i′));

- the set of possible symbols y(j), j = 1,...,Q and the Q × D emission matrix E(y(j)|x(i)) defining the conditional probability of a symbol given a state.

Neither Ω nor E vary with time.

Hidden Markov Models (HMM)

There are three fundamental issues with HMM:

Problem 1. Given a transition matrix Ω, an emission matrix E, and an initial state x_0 how can we **compute the probability p(y)** of a time sequence of observed symbols **y** = $(y^1,y^2,...,y^M)$?

Hidden Markov Models (HMM)

There are three fundamental issues with HMM:

Problem 1. Given a transition matrix Ω, an emission matrix E, and an initial state x_0 how can we **compute the probability p(y)** of a time sequence of observed symbols **y** = $(y^1,y^2,...,y^M)$?

Problem 2. Given Ω , E, what is the conditional probability **p(x|y)** of a sequence of hidden states, **x** = $(x^1, x^2,. . ., x^M)$, **given a sequence of observed symbols y**? In particular, what is the trajectory of states $\mathbf{x}^{\mathbf{MAP}}$ (**y**) with highest probability?

Hidden Markov Models (HMM)

There are three fundamental issues with HMM:

Problem 1. Given a transition matrix Ω, an emission matrix E, and an initial state x_0 how can we **compute the probability p(y)** of a time sequence of observed symbols **y** = $(y^1,y^2,...,y^M)$?

Problem 2. Given Ω , E, what is the conditional probability **p(x|y)** of a sequence of hidden states, **x** = $(x^1, x^2,. . ., x^M)$, **given a sequence of observed symbols y**? In particular, what is the trajectory of states $\mathbf{x}^{\mathbf{MAP}}$ (**y**) with highest probability?

Problem 3. Given a sequence of symbols **y**, how can we **infer Ω, E**?

Problem 1 & dynamic programming

Problem 1. Given a transition matrix Ω, an emission matrix E , and an initial state x_0 how can we compute the probability p(y) of a sequence of observed symbols **y** ?

$$p(\boldsymbol{y}) = \sum_{x_1, x_2, \ldots x_M} p(x_1, \ldots, x_M | x_0) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t)$$

$$= \sum_{x_1, x_2, \ldots x_M} \prod_{t=1}^{M} \Omega(x_{t-1} \to x_t) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t) \ .$$

Looks really bad: summation over D^M sequences of states …

Problem 1 & dynamic programming

Problem 1. Given a transition matrix Ω, an emission matrix E , and an initial state x_0 how can we compute the probability p(y) of a sequence of observed symbols **y** ?

$$p(\boldsymbol{y}) = \sum_{x_1, x_2, \ldots x_M} p(x_1, \ldots, x_M | x_0) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t)$$

$$= \sum_{x_1, x_2, \ldots x_M} \prod_{t=1}^{M} \Omega(x_{t-1} \to x_t) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t) \ .$$

Looks really bad: summation over D^M sequences of states …

But can be done in time linear in M because the underlying process is Markov, i.e. local in time !!

We define:

$$\mathcal{M}_t(x', x) \equiv \Omega(x \to x') \, \mathcal{E}(y_t | x') \ .$$

These can be seen as the elements of a DxD matrix, which explicitly depends on time t through the symbol y_t.

Problem 1 & dynamic programming

Problem 1. Given a transition matrix Ω, an emission matrix E , and an initial state x_0 how can we compute the probability p(y) of a sequence of observed symbols **y** ?

$$p(\boldsymbol{y}) = \sum_{x_1, x_2, \ldots x_M} p(x_1, \ldots, x_M | x_0) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t)$$

$$= \sum_{x_1, x_2, \ldots x_M} \prod_{t=1}^{M} \Omega(x_{t-1} \to x_t) \prod_{t=1}^{M} \mathcal{E}(y_t | x_t) \ .$$

We define:

$$\mathcal{M}_t(x', x) \equiv \Omega(x \to x') \, \mathcal{E}(y_t | x') \ .$$

$$p(\boldsymbol{y}) = \sum_{x_1, \ldots, x_M} \mathcal{M}_M(x_M, x_{M-1}) \times \ldots \times \mathcal{M}_t(x_t, x_{t-1}) \ldots \times \mathcal{M}_2(x_2, x_1) \times \mathcal{M}_1(x_1, x_0)$$

$$= \sum_{x_M} \left[\mathcal{M}_M \cdot \ldots \cdot \mathcal{M}_t \cdot \ldots \mathcal{M}_2 \cdot \mathcal{M}_1 \right](x_M, x_0) \ .$$

We have reduced the computation of p(y) to that of the product of M different matrices that can be carried out in time M × D^2: huge gain with respect to D^M

Problem 2 & Viterbi Algorithm

Problem 2. Given Ω , E, what is the conditional probability **p(x|y)** of a sequence of hidden states, **x** = (x^1, x^2,. . ., x^M), **given a sequence of observed symbols y**? In particular, what is the trajectory of states $\mathbf{x}^{\mathbf{MAP}}$ (**y**) with highest probability?

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x}) \times p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{\mathcal{M}_M(x_M, x_{M-1}) \times \ldots \times \mathcal{M}_t(x_t, x_{t-1}) \times \ldots \times \mathcal{M}_1(x_1, x_0)}{\sum_{x_M} \left[\mathcal{M}_M \cdot \ldots \cdot \mathcal{M}_1\right](x_M, x_0)} .$$

Trajectory of hidden states with highest probability?

Looks bad again, we have to look through D^M possibilities to maximize the numerator …

Problem 2 & Viterbi Algorithm

Problem 2. Given Ω , E, what is the conditional probability **p(x|y)** of a sequence of hidden states, **x** = (x^1, x^2,. . ., x^M), **given a sequence of observed symbols y**? In particular, what is the trajectory of states $\mathbf{x}^{\mathbf{MAP}}$ (**y**) with highest probability?

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x}) \times p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{\mathcal{M}_M(x_M, x_{M-1}) \times \ldots \times \mathcal{M}_t(x_t, x_{t-1}) \times \ldots \times \mathcal{M}_1(x_1, x_0)}{\sum_{x_M} \left[\mathcal{M}_M \cdot \ldots \cdot \mathcal{M}_1\right](x_M, x_0)} .$$

Trajectory of hidden states with highest probability?

Looks bad again, we have to look through D^M possibilities to maximize the numerator …

Key observation:
A state, say, x_t, appears in two matrix elements only, $M_t(x_t, x_{t-1})$ & $M_{t+1}(x_{t+1}, x_t)$
If we fix the values of x_{t-1} and x_{t+1} the optimisation over x_t is easy to perform.
This observation can be turned into an efficient recursive algorithm, with running time linear in M and D, invented by Viterbi (1967)

Problem 2 & Viterbi Algorithm

Problem 2. Given Ω , E, what is the conditional probability **p(x|y)** of a sequence of hidden states, **x** = (x^1, x^2,. . ., x^M), **given a sequence of observed symbols y**? In particular, what is the trajectory of states $\mathbf{x}^{\mathbf{MAP}}$ (**y**) with highest probability?

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x}) \times p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{\mathcal{M}_M(x_M, x_{M-1}) \times \ldots \times \mathcal{M}_t(x_t, x_{t-1}) \times \ldots \times \mathcal{M}_1(x_1, x_0)}{\sum_{x_M} \left[\mathcal{M}_M \cdot \ldots \cdot \mathcal{M}_1\right](x_M, x_0)} .$$

Trajectory of hidden states with highest probability?

Looks bad again, we have to look through D^M possibilities to maximize the numerator …

Key observation:
A state, say, x_t, appears in two matrix elements only, $M_t(x_t, x_{t-1})$ & $M_{t+1}(x_{t+1}, x_t)$
If we fix the values of x_{t-1} and x_{t+1} the optimisation over x_t is easy to perform.
This observation can be turned into an efficient recursive algorithm, with running time linear in M and D, invented by Viterbi (1967)

Problem 2 & Viterbi Algorithm

Suppose that we know, for each possible state x_t at time t, the most likely sequence $\{x^*_0, \ldots, x^*_{t-1}\}$ at the previous times.

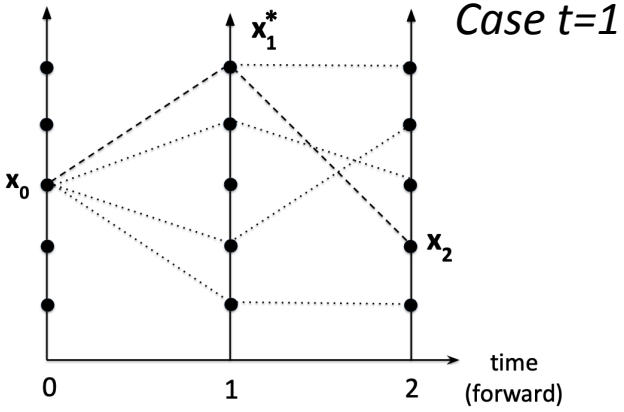
Then, to determine the best value of the state x_t for each possible state x_{t+1} at time t+1 we have to maximize the part of the log probability that depends on x_t:

$$\log \Omega\left(x^*_{t-1}(x_t) \rightarrow x_t\right) + \log E\left(y_t \mid x_t\right) + \log \Omega\left(x_t \rightarrow x_{t+1}\right)$$

This is easy to do since there are only D possibilities for x_t (but we have to repeat this maximization step D times, once for each x_{t+1})

We then obtain $x^*_t(x_{t+1})$,

and iterate until the ending time.

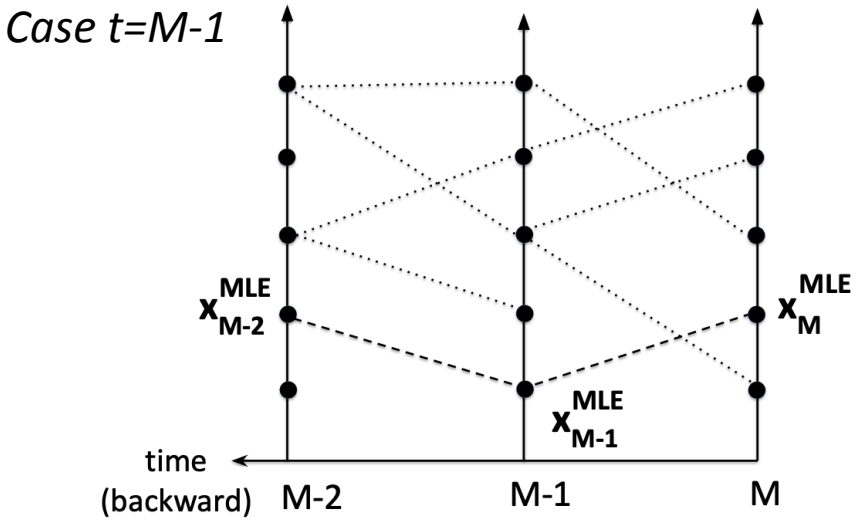


Problem 2 & Viterbi Algorithm

At the end, we find the best x_M through maximization of

$$\log \Omega\left(x^*_{M-1}(x_M) \rightarrow x_M\right) + \log E\left(y_M \mid x_M\right)$$

Then, we have to backtrack in reverse time all the way down to time t=0, using the tables $x^*_t(x_{t+1})$ determines in the forward pass



Problem 3 & Expectation-Maximization

Problem 3. Given a sequence of symbols **y**, how can we **infer Ω, E**?

This is the hardest problem!

- Can be "solved" using an interative EM scheme:

  Estimates of parameters -> Sampling of trajectories of states and symbols
  -> New estimate of parameters

- Guarantee that likelihood increases at each step (local convergence)

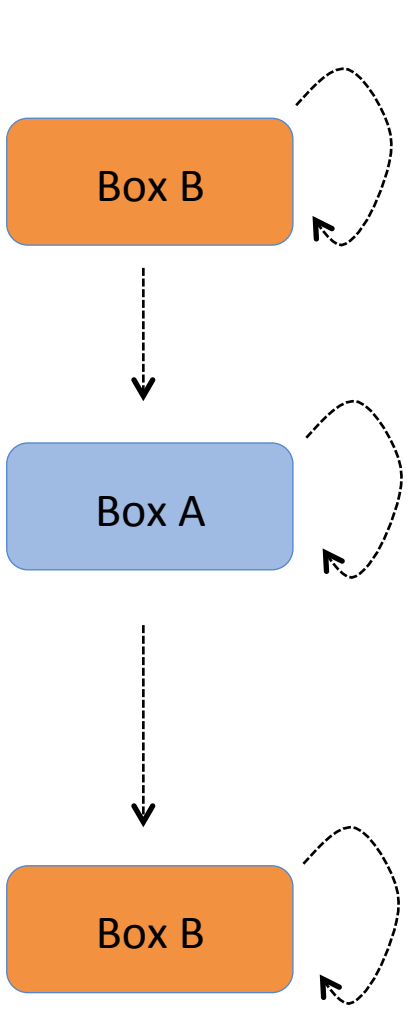
- But no guarantee that best maximum is reached. In practice, many restarts …

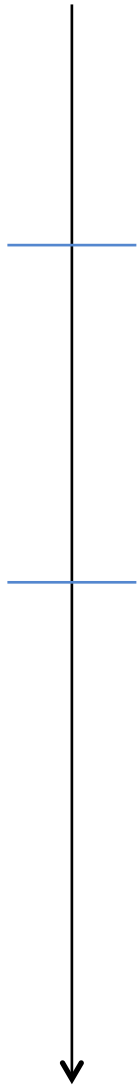
Application: Cookie Monster is striking again!

Me want cookie!

Application: Cookie Monster is striking again!

Me want cookie!





Plain
Chocolate
C
C
P
P
C
P
C
P
P
P
C
C
C
P
P
C

The transition and emission matrices

$$\Omega\left(B \rightarrow B\right) = 0.8,\ \Omega\left(B \rightarrow A\right) = 0.2$$

$$\Omega\left(A \rightarrow A\right) = 0.5,\ \Omega\left(A \rightarrow B\right) = 0.5$$

$$E\left(P \mid A\right) = E(C \mid A) = 0.5,\ E\left(P \mid B\right) = 0.25, E(C \mid B) = 0.75$$

The transition and emission matrices

$$\Omega\left(B \to B\right) = 0.8,\ \Omega\left(B \to A\right) = 0.2$$

$$\Omega\left(A \to A\right) = 0.5,\ \Omega\left(A \to B\right) = 0.5$$

$$E\left(P \mid A\right) = E(C \mid A) = 0.5,\ E\left(P \mid B\right) = 0.25, E(C \mid B) = 0.75$$

There are two types of matrices $\mathcal{M}_t(x', x) \equiv \Omega(x \to x')\, \mathcal{E}(y_t | x')$.

$$M_{y=P} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ \frac{1}{8} & \frac{1}{5} \end{pmatrix}, \quad M_{y=C} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ \frac{3}{8} & \frac{3}{5} \end{pmatrix}$$

The transition and emission matrices

$$\Omega\left(B \to B\right) = 0.8,\ \Omega\left(B \to A\right) = 0.2$$

$$\Omega\left(A \to A\right) = 0.5,\ \Omega\left(A \to B\right) = 0.5$$

$$E\left(P \mid A\right) = E(C \mid A) = 0.5,\ E\left(P \mid B\right) = 0.25, E(C \mid B) = 0.75$$

There are two types of matrices $\mathcal{M}_t(x', x) \equiv \Omega(x \to x')\, \mathcal{E}(y_t | x')$.

$$M_{y=P} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ \frac{1}{8} & \frac{1}{5} \end{pmatrix}, \quad M_{y=C} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ \frac{3}{8} & \frac{3}{5} \end{pmatrix}$$

<u>Exercise 5:</u> Please check I haven't made a mistake here!

Problem 1

What is the probability of, say, **P, C, C, C, P, P** ?

$$\Pr = (1,1)^T \cdot M_P \cdot M_P \cdot M_C \cdot M_C \cdot M_C \cdot M_P \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

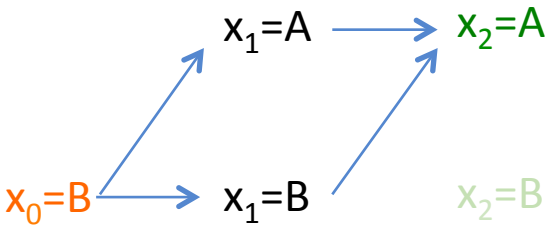
Last state is unknown, so we sum over all possibilities

I assume Cookie Monster's initial state is B

One finds $\Pr = 0.00985...$

Problem 2

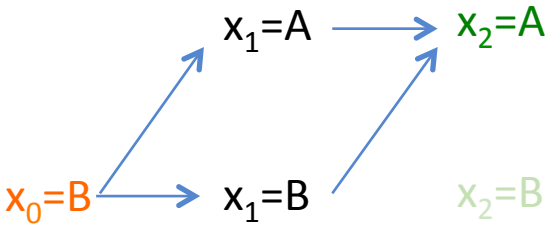
What is the most likely sequence of states corresponding to the observations **P, C** ? (M=2 …)

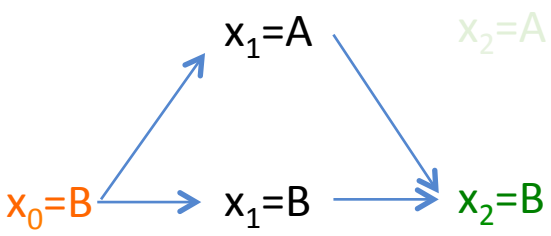


$$M_C(A,A)\cdot M_P(A,B)=\frac{1}{40}>M_C(A,B)\cdot M_P(B,B)=\frac{1}{50}\rightarrow x_1^*(x_2=A)=A$$

Problem 2

What is the most likely sequence of states corresponding to the observations **P, C** ? (M=2 …)





$$M_C(A,A)\cdot M_P(A,B)=\frac{1}{40}>M_C(A,B)\cdot M_P(B,B)=\frac{1}{50}\rightarrow x_1^*(x_2=A)=A$$

$$M_C(B,A)\cdot M_P(A,B)=\frac{3}{80}<M_C(B,B)\cdot M_P(B,B)=\frac{3}{25}\rightarrow x_1^*(x_2=B)=B$$

Problem 2

What is the most likely sequence of states corresponding to the observations **P, C** ? (M=2 …)

$$M_C(A,A)\cdot M_P(A,B)=\frac{1}{40}>M_C(A,B)\cdot M_P(B,B)=\frac{1}{50}\rightarrow x_1^*(x_2=A)=A$$

$$M_C(B,A)\cdot M_P(A,B)=\frac{3}{80}<M_C(B,B)\cdot M_P(B,B)=\frac{3}{25}\rightarrow x_1^*(x_2=B)=B$$

Then, at time M=2, the best state is B, so it is also B at time 1.

The most likely scenario is that Cookie Monster has taken the two cookies from the same box B

Solution to Exercise 4

The transition matrix, when states are ordered as Rain, Snow, Nice is

$$\Omega = \begin{pmatrix} 0.5 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0. \end{pmatrix}$$

There is one eigenvector corresponding to eigenvalue 1. Its normalized components are

$$P_{stat} = (2/5, 2/5, 1/5)$$

All the other eigenvectors have eigenvalues < 1 and are irrelevant at large times

Decoding of trajectory from HPC activity

- Continuity prior over trajectory: $P_{prior}(\{x_t\}) \propto \prod_t \exp\left[-(x_t - x_{t-\Delta t})^2 / (2v^2\Delta t^2)\right]$

  Here, v = typical velocity (hyper-parameter, should be chosen with care)

- Find most likely position:

MAP decoding $$\{x_t\}_{t=1,\ldots,T} = \text{argmax}\left[\prod_{t=1,\ldots,T} P\left(\{s_{i,t}\}\middle|x_t\right) \times P_{prior}\left(\{x_t\}\right)\right]$$

  - Brute force decoding impossible: exponential in *T*
  - Dynamic programming is efficient: linear in *T*, see Thursday