Chapter 15 (AIAMA)

Probabilistic Reasoning Over Time

Sukarna Barua
Assistant Professor, CSE, BUET

Temporal Probabilistic Models

- Static world (as we considered in Bayesian network):
  - Random variables have a fixed number of states/values.
  - Values of Random variables doesn't change over time
- Dynamic world (time is an important factor):
  - Random variables have a fixed number of states/values.
  - Values of Random variables change over time.

Temporal Probabilistic Models

- Dynamic world has a state at time t
  - State is composed of a set of random variables X_t
  - A snapshot of the state at time t is a set of values of X_t
- State is not observable
  - State is not directly observable.
  - A set of evidence variables E_t are observable at time t [evidences depends on state]
  - We may infer which state we are in from the evidence!

Temporal Probabilistic Models: Example

You want to know whether you have infection at time step t. You can measure fever, headache, stomachache at time step t.

- $X_t : \{Infection_t\}$
  - Values: Yes/No [*Unobservable by agent, hidden*]
- $E_t : \{Fever_t, Stomachache_t, Headache_t\}$
  - Values: Yes/No [*Observable by agent*]

Temporal Probabilistic Models

In a temporal probabilistic model, agent have:

- Environment: Partially observable
- Belief state: What is the current state as agent maintains/believes?
- Transition model: How the environment might evolve in the next time step
- Sensor model: How the observable events happen at world state?
- Decision: How the agent take action?
  - Evidence → Belief state → Decision

Hidden Markov Models

- A temporal probabilistic model may be called a Hidden Markov Model (HMM) when the state is represented by a discrete random variable:
- X_t: A single state variables at time t
  - Unobservable by agent [*hidden from the agent]*
- E_t: Set of evidence variables
  - Observable by agent [*known through percepts*]

Hidden Markov Models

- What happens if world state has multiple random variables?
  - Multiple random variables may be mapped to a single random variable
  - Example: <Burglary, Earthquake> makes up agent state both are Boolean.
  - Construct a single variable <BE> with four values {0,1,2,3} where
    - 0 means Burglary=T and Earthquake = T
    - 1 means Burglary=T and Earthquake = F
    - 2 means Burglary=F and Earthquake = T
    - 3 means Burglary=F and Earthquake = F

Hidden Markov Models: Example

A security guard inside a building needs to know whether it's raining outside. He can only see if someone coming in with/without an umbrella.

- $X_t : \{Rain_t\}$
  - Values: Yes/No [Unobservable by agent]
- $E_t : \{Umbrella_t\}$
  - Values: Yes/No [Observable by agent]

Transition Model

- Specifies the probability distribution of the state at time t, given the previous states:

  $P(X_t|X_{1:t-1})$

  - Assume the size of CPT when t is large [*exponentially large*]
  - Problematic as number of time steps increases
  - Not practical as current state may depend only on few previous states

Markov Assumption for Transition Model

- Assumption: Current state is independent of all states $X_{1:t-k-1}$ given the previous k number of states $X_{t-k:t-1}$:

$$P(X_t | X_{1:t-1}) = P(X_t | X_{t-k:t-1})$$

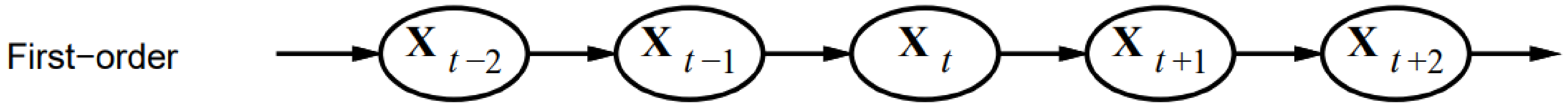
- Markov Process: Process satisfying Markov assumption.

  - Also known as Markov chains.

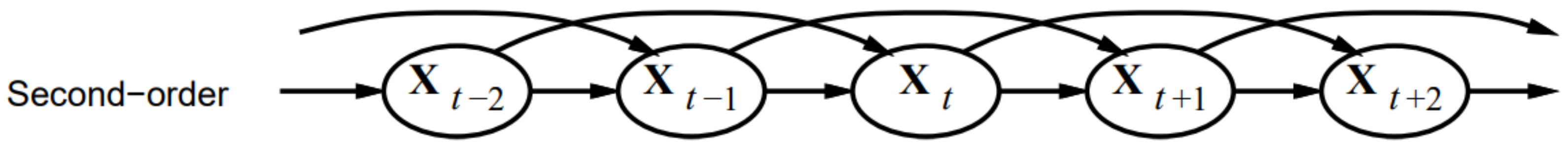
  - After Russian mathematician Andrei Markov

Order of Markov Process

- First Order Markov Process:
  - Current state is independent of all other states given only the previous state
  - $P(X_t|X_{1:t-1}) = P(X_t|X_{t-1})$
  - Transition model is a conditional distribution $P(X_t|X_{t-1})$



- For a second order Markov Process:
  - Transition model is a conditional distribution $P(X_t|X_{t-1}, X_{t-2})$



First Order Markov Process

- Stationary process: transition model do not change over time steps

  - $P(X_t|X_{t-1})$ is same for all time steps t.

  - $P(X_2|X_1) = P(X_3|X_2) = \cdots$

  - $P\left(X_t = x_j \middle| X_{t-1} = x_i\right) = a_i[j]$

  [a_{ij} *is the probability of state transitioning from* x_i *to* x_j]

Sensor/Emission Model

- Evidence values depend on current state as well as all previous states and evidence values
- Probability distribution of events E_t:

$$P(E_t | X_{1:t}, E_{1:t-1})$$

  - What is the probability that $Umbrella_t = true$ given all previous state and evidence values?
  - What is the size of CPT when t is large? [*exponentially large*]
  - Not practical from computational perspective

Markov Assumption for Sensor Model

- Assumption: Evidence at time t is independent of all previous states and events given the state at time t (current state).

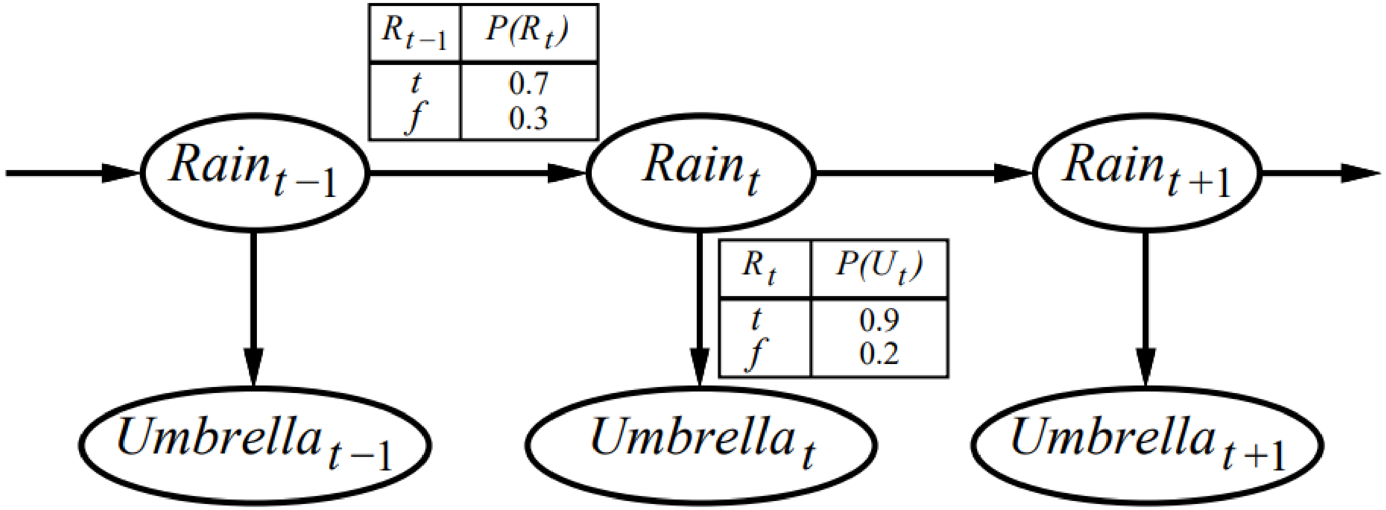
$$P(E_t|X_{1:t}, E_{1:t-1}) = P(E_t|X_t) \quad [evidence\ depend\ only\ on\ current\ state]$$

  - Evidence only depend on current state and is independent of all previous states and evidences
  - $P(E_t = e_k|X_t = x_i) = o_{i,k}$ [*probability of emitting output* o_k *from state* x_i]
- Also known as Observation/Emission Model

Example Markov Process

- For the umbrella example:

  - Transition model: $P(R_t|R_{t-1})$, sensor model: $P(U_t|R_t)$



Complete/Full Joint Distribution

- We have
  - $P(X_t|X_{t-1})$ [*transition model*]
  - $P(E_t|X_t)$ [*sensor model*]
- We also need
  - $P(X_1)$: The prior probability distribution of states at time step $t = 1$
- Complete joint distribution can be computed as:

$$P(X_{1:t}, E_{1:t}) = P(E_{1:t}|X_{1:t})P(X_{1:t}) = \prod_{i=1}^{t} P(E_i|X_i)P(X_i|X_{i-1})$$

[Assume $P(X_1|X_0) = P(\mathrm{X}_1)$ for notational convenience]

Complete/Full Joint Distribution

- Complete joint distribution derivation:

$$P(X_{1:t}, E_{1:t}) = P(E_{1:t}|X_{1:t})P(X_{1:t})$$

$$= P(E_t|E_{1:t-1}, X_{1:t})P(E_{1:t-1}|X_{1:t})P(X_t|X_{1:t-1}P(X_{1:t-1})$$

$$= P(E_t|X_t\ )P(E_{1:t-1}|X_{1:t})P(X_t|X_{1:t-1}P(X_{t-1}|X_{1:t-2})P(X_{1:t-2})$$

$$= \prod_{i=1}^{t} P(E_i|X_i) \times \prod_{1}^{t} P(X_i|X_{i-1})$$

$$= \prod_{i=1}^{t}[P(E_i|X_i)P(X_i|X_{i-1})]$$

Is First Order Markov Process Accurate?

- Sometimes true

  - For example, in a random walk along x – axis, position at time step t only depends on position at time step $t-1$.

- Sometimes not

  - For example, in our rain example, probability of raining at time step t may depend on several previous rainy days $t-1, t-2, \ldots$

Is First Order Markov Process Accurate?

- Sometimes not
  - For example, in our rain example, probability of raining at time step t only depend on whether it rained at time step $t-1$
- Solutions
  - Increase the order of the Markov process: $P(X_t|X_{t-1}, X_{t-2})$
  - Incorporate more state variables: $Temp_t, Humidity_t, Pressure_t, Session_t$, etc.

Inference in First Order Markov Process

- **Filtering query**: Compute probability distribution of current state given all observations to date.
  - $P(X_t|e_{1:t})$
  - Compute probability of raining (and not raining also!) today, given all umbrella observations taken so far
  - *Note the use capital and small letters: Capitals specify random variable and small letters specify values of random values.*
  - Required for decision making at current state

Inference in First Order Markov Process

- **Prediction query**: Compute probability distribution of a future state given all observations to date.
  - $P(X_{t+k}|e_{1:t})$
  - Compute probability of raining three days from now, given all umbrella observations taken so far
  - Required for decision making about future action

Inference in First Order Markov Process

- **Smoothing query**: Compute probability distribution of a past state given all observations to date.
  - $P(X_k|e_{1:t}), 0 \leq \mathrm{k} < \mathrm{t}$
  - Compute probability of raining last Wednesday, given all umbrella observations taken so far
  - Smoothing provides a better estimate than what was made before

Inference in First Order Markov Process

- **Most likely explanation query**: Given a sequence of observation, what is the most likely state sequence that have generated the observation sequence?
  - $P(X_{1:t}|e_{1:t})$
  - Umbrella was observed on first three days and absent on fourth, the most likely state sequence could be it rained first three days and did not rain on fourth.
  - Speech recognition: What is the sequence of words given a sequence of sounds?

Filtering

- Compute probability distribution of current state X_{t+1} given observation sequence $e_{1:t+1}$
- Agent maintains the probability distribution of current state X_t at time step t.
- As new evidence e_{t+1} comes up, agent updates its estimation of current state probabilities $P(X_{t+1})$

$$
\begin{aligned}
\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \quad \text{(dividing up the evidence)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}, \mathbf{e}_{1:t})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(using Bayes' rule)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(by the sensor Markov assumption).}
\end{aligned}
$$

Filtering

- α: a is a normalizing constant to make probabilities sum up to 1

$$\begin{aligned}
\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \quad \text{(dividing up the evidence)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}, \mathbf{e}_{1:t})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(using Bayes' rule)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(by the sensor Markov assumption).}
\end{aligned}$$

Filtering

$$
\begin{aligned}
\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \quad \text{(dividing up the evidence)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}, \mathbf{e}_{1:t})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(using Bayes' rule)} \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1})\, \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \quad \text{(by the sensor Markov assumption).}
\end{aligned}
$$

- How to calculate $P(X_{t+1}|e_{1:t})$?
  - Marginalize over X_t: $P(X_{t+1}) = \sum_{x_t} P(X_{t+1}, x_t) = \sum_{x_t} P(X_{t+1}|x_t)P(x_t)$

$$
\begin{aligned}
\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) &= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{e}_{1:t}) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) \\
&= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) \quad \text{(Markov assumption).}
\end{aligned}
$$

Filtering

$$\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) = \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{e}_{1:t}) P(\mathbf{x}_t \mid \mathbf{e}_{1:t})$$

$$= \alpha\, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) \quad \text{(Markov assumption).}$$

- $P(e_{t+1}|X_{t+1})$ comes from observation/sensor model [given]
- $P(X_{t+1}|x_t)$ comes from the transition model [given]
- $P(x_t|e_{1:t})$ is the probability distribution of states at time step t

  - This part is recurrence and can be computed recursively or iteratively [using dynamic programming approach]

Filtering

$$\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) = \alpha \, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t, \mathbf{e}_{1:t}) P(\mathbf{x}_t \mid \mathbf{e}_{1:t})$$

$$= \alpha \, \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} \mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t) P(\mathbf{x}_t \mid \mathbf{e}_{1:t}) \quad \text{(Markov assumption).}$$

- Let, $P(X_t|e_{1:t}) = \mathbf{f}_t$ [$\mathbf{f}_t$ *is a vector/array of probabilities*]
  - $P(X_t = x_i|e_{1:t}) = \mathbf{f}_t[\mathrm{i}]$ [$\mathbf{f}_t[\mathrm{i}]$ *is a single probability value*]

- Hence, $\mathbf{f}_t[i] = P(X_t = x_i|e_{1:t})$

$$= \alpha \, P(e_t \,|X_t = x_i) \textstyle\sum_j P(X_t = x_i \left| X_{t-1} = x_j \right) P(X_{t-1} = x_j|e_{1:t-1})$$

$$= \alpha \times \left(o_{i,k}\right) \times \textstyle\sum_j (a_{j,i})(\mathbf{f}_{t-1}[j])$$ [assume $e_{t+1} = e_k$ an output value]

Filtering: Forward Algorithm

- $\mathbf{f}_t[i] = \alpha \times \left(o_{i,k}\right) \times \sum_j \left(a_{j,i}\right)(\mathbf{f}_{t-1}[j])$
- $\mathbf{f}_t$ is known as forward probabilities
- How to compute forward probabilities up to time step t?
  - Start from $t = 1$ and compute $\mathbf{f}_1$ [base condition]
  - Compute going forward in time up to $\mathbf{f}_t$ using the recurrence
  - The algorithm is known as forward algorithm.

Filtering: Forward Algorithm

- $\mathbf{f}_t[i] = \alpha \times \left(o_{i,k}\right) \times \sum_j \left(a_{j,i}\right)(\mathbf{f}_{t-1}[j])$
- $\mathbf{f}_t$ is known as forward probabilities
- How to compute compute $\mathbf{f}_1$ [base condition]?
  - $\mathbf{f}_1[i] = P(X_1 = x_i | e_1)$

    $= \alpha P(e_1 | X_1 = x_i) P(X_1 = x_i)$

    $= \alpha \times o_{i,l} \times \pi_i$ [assume $\mathrm{e}_1 = \mathrm{e}_l$, π_i is the prior probability of state x_i]

Filtering: Example

- Compute $P(R_2|u_{1:2})$
- Day 1: $P(R_1|u_1) = \alpha P(u_1|R_1)P(R_1)$

  - $P(R_1)$ is the prior probability distribution of initial state [at time $t = 1$]

    - If both states are equally likely from START, $P(R_1) =< 0.5, 0.5 >$

  - $P(R_1|u_1)$ can now be calculated as:

$$\begin{aligned} \mathbf{P}(R_1 \mid u_1) &= \alpha\, \mathbf{P}(u_1 \mid R_1)\mathbf{P}(R_1) = \alpha \langle 0.9, 0.2 \rangle \langle 0.5, 0.5 \rangle \\ &= \alpha \langle 0.45, 0.1 \rangle \approx \langle 0.818, 0.182 \rangle \,. \end{aligned}$$

Filtering: Example

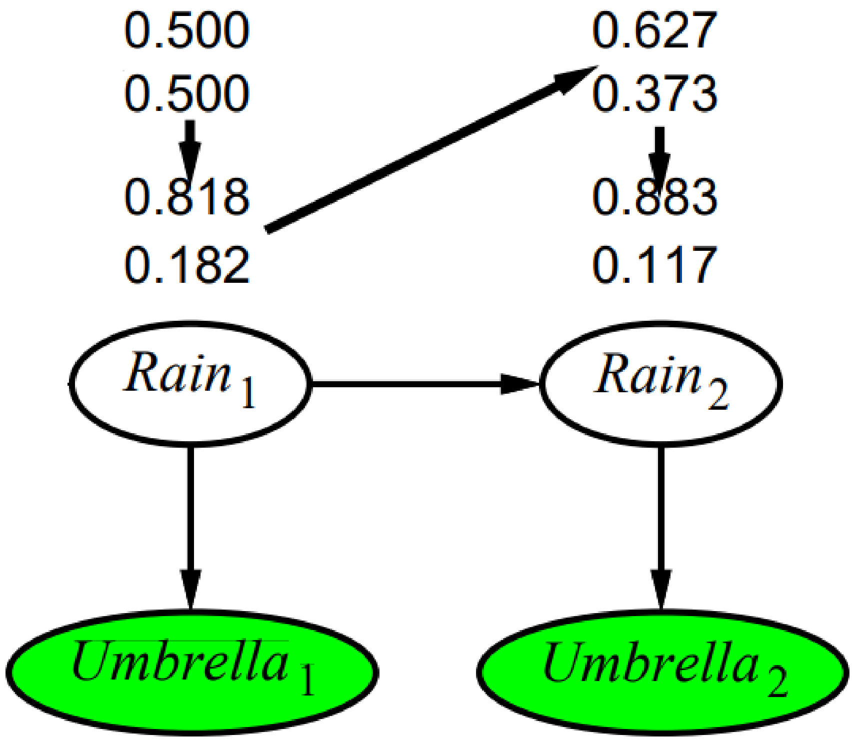
- Day 2: $P(R_2|u_{1:2}) = \alpha P(u_2|R_2)P(R_2|u_1) = \alpha P(u_1|R_1)\sum_{r_1} P(R_1|r_1)P(r_1|u_1)$
  - Can be calculated as:

$$\begin{aligned}\mathbf{P}(R_2 \mid u_1) &= \sum_{r_1} \mathbf{P}(R_2 \mid r_1)P(r_1 \mid u_1) \\ &= \langle 0.7, 0.3\rangle \times 0.818 + \langle 0.3, 0.7\rangle \times 0.182 \approx \langle 0.627, 0.373\rangle\end{aligned}$$

$$\begin{aligned}\mathbf{P}(R_2 \mid u_1, u_2) &= \alpha\,\mathbf{P}(u_2 \mid R_2)\mathbf{P}(R_2 \mid u_1) = \alpha\,\langle 0.9, 0.2\rangle\langle 0.627, 0.373\rangle \\ &= \alpha\,\langle 0.565, 0.075\rangle \approx \langle 0.883, 0.117\rangle\,.\end{aligned}$$

Filtering: Example

- Probability of rain increases at day 2 from day 1 [why?]



Prediction

- Compute probability distribution of a future state: $P(X_{t+k}|e_{1:t})$
- Can be computed using filtering:

  - First compute $P(X_t|e_{1:t})$ [*forward algorithm*]

  - Then compute as: $P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$.

  - Similarly, compute $P(X_{t+2}|e_{1:t})$, …, $P(X_{t+k}|e_{1:t})$

- Recursive/dynamic programming algorithm:

$$\mathbf{P}(\mathbf{X}_{t+k+1} \mid \mathbf{e}_{1:t}) = \sum_{\mathbf{x}_{t+k}} \mathbf{P}(\mathbf{X}_{t+k+1} \mid \mathbf{x}_{t+k}) P(\mathbf{x}_{t+k} \mid \mathbf{e}_{1:t}) \ .$$

Prediction: Don’t Go Too Much Ahead

- Recursive/dynamic programming algorithm:

$$\mathbf{P}(\mathbf{X}_{t+k+1} \mid \mathbf{e}_{1:t}) = \sum_{\mathbf{x}_{t+k}} \mathbf{P}(\mathbf{X}_{t+k+1} \mid \mathbf{x}_{t+k}) P(\mathbf{x}_{t+k} \mid \mathbf{e}_{1:t}) \ .$$

- Predicting too much ahead may be useless
  - $P(X_{t+k+1}|e_{1:t})$ will become fixed (stationary distribution of the Markov Process) after some time steps k
  - The time taken to reach the fixed point is known as Mixing Time.

- The more uncertainty in the transition model, the shorter will be the mixing time and the more future is obscured!

Likelihood of Evidence Sequence

- What is the likelihood of evidence sequence $e_{1:t}$?
- Compute as
  - $P(e_{1:t}) = \sum_{x_t} P(e_{1:t}, x_t)$

- $P(e_{1:t}, x_t)$ can be calculated recursively or using dynamic programming:

$$\begin{aligned} P(e_{1:t}, x_t) &= P(e_{1:t-1}, e_t, x_t) = P(e_t | x_t, e_{1:t-1}) P(x_t, e_{1:t-1}) \\ &= P(e_t | x_t) \sum_{x_{t-1}} P(x_t, x_{t-1}, e_{1:t-1}) \quad [\mathit{Markov\ assumption}] \\ &= P(e_t | x_t) \sum_{x_{t-1}} P(x_t | x_{t-1}) P(x_{t-1}, e_{1:t-1}) \end{aligned}$$

- $P(x_{t-1}, e_{1:t-1})$ can be computed recursively [using dynamic programming]
- This is similar to the forward algorithm [described earlier]