Graphical Models

Saeed Saremi

Assigned reading: Ch. 11

October 31, 2024

outline

- Markov chains
- Hidden Markov models (HMM)
- Probabilistic graphical models (PGM)
  - reading conditional independence from directed graphs
  - atomic independence structures on triples
  - d(irected)-separation
- HMM $\subset$ PGM

Markov chains

- So far in the course we have assumed independent and identically distributed (i.i.d.) datasets. But we now consider a sequence of random variables

$$X_{1:T} := (X_1, X_2, \ldots, X_T)$$

where the random variables are dependent $X_t \not\perp\!\!\!\perp X_{t'}$:

$$P(X_t \mid X_{t'}) \neq P(X_t).$$

- The (first order) Markov condition:

  In predicting the future, the past doesn't matter, only the present.

- In other words,

$$P(X_{t+1} \mid X_{1:t}) = P(X_{t+1} \mid X_t).$$

? Prove the following:

$$P(X_{1:T}) = P(X_1) \prod_{t=1}^{T-1} P(X_{t+1} \mid X_t)$$

? How many parameters are needed to identify a Markov chain? Assume $\mathcal{X} = [K]$.

(time) homogenous Markov chains

- Consider our discrete setup where $X_t \in \mathcal{X} = [K]$
- For many problems of practical interest we assume the transition probabilities $P(X_{t+1} = c \mid X_t = c')$ does not depend on (time) t:

$$P(X_{t+1} = k' \mid X_t = k) = A_{kk'}$$

- Such Markov chains are referred to as *time homogenous*.
- The matrix A is referred to as transition probability matrix.

? What is the dimension of A?

? What are the properties the transition probability matrix A should satisfy?

? Does A have to be symmetric?

? Define the row vector π_t to be the distribution of $P(X_t)$:

$$\pi_t(k) = P(X_t = k).$$

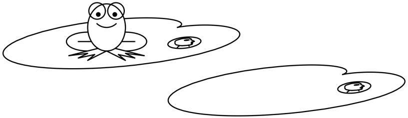
(a) Prove that

$$\pi_{t+1} = \pi_t A.$$

(b) What is π_t in terms of π_1?

example: randomly jumping frog

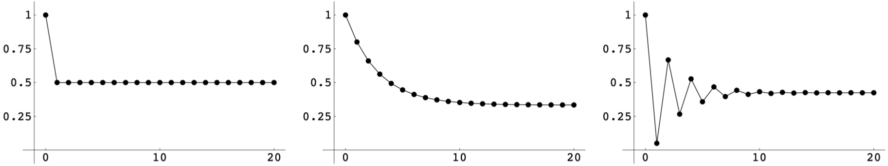
- Consider a randomly jumping frog with $\mathcal{X} = \{e, w\}$. Whenever he tosses heads, he jumps to the other lily pad:



- Say the coin on the east pad has probability p of landing heads up, while the coin on the west pad has probability q of landing heads up.

? Can we write $P(X_{1:T})$ as a Markov chain? Is it homogenous? If so, what is A?

? The probability of being on the east pad (started from the east pad) plotted below versus time in three different scenarios. Let's guess if $p > q$ or not in each case.



(a) (b) (c)

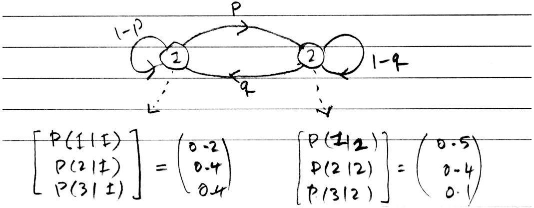
the chirping-jumping frog HMM

A Markov chain is useful when we need to compute a probability for a sequence of observable events. In many cases, however, the events we are interested in are hidden: we don't observe them directly. A hidden Markov model (HMM) allows us to talk about both observed events and hidden events that we think of as "causal" factors in our probabilistic model.

- Our jumping frog emits $y \in \mathcal{Y} = [L]$ chirps when he jumps on each lily pad.
- In this example we do not observe the frog location, we only hear the chirps.
- The emission of y chirps is probabilistic captured by

$$P(Y_t = l \mid X_t = k)$$

  and is shown diagrammatically[1] as:



[1]Soon we will learn a more elegant way to describe an HMM diagramatically.

A hidden Markov model has the following ingredients:

- A set of K states in $[K] = \{1, \ldots, K\}$.
- A transition probability matrix $A \in \mathbb{R}^{K \times K}$, whose rows must add to 1.
- A sequence of observations $Y_{1:T} = (Y_1, \ldots, Y_T)$ with $Y_t \in \mathcal{Y}$, where $|\mathcal{Y}| = L$.
- A sequence of observation likelihoods, also called emission probabilities,

$$B_{kl} = P(Y_t = l \mid X_t = k),$$

  i.e., the probability of an observation $Y_t = l$ being generated from a state $X_t = k$.
- An initial probability distribution over states denoted by π_1.

There are three fundamental problems concerning HMMs:

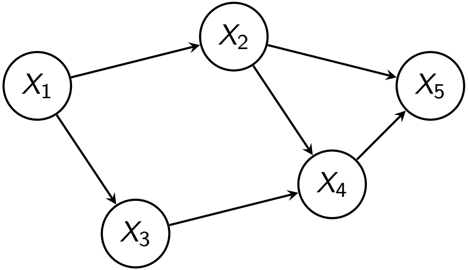
- Likelihood: Given an HMM (A, B, π) and the observation $Y_{1:T}$, determine the likelihood $P(Y_{1:T})$.
- Decoding: Given an observation sequence $Y_{1:T}$ and an HMM (A, B, π) discover the mostly likely hidden state sequence $X_{1:T}$.
- Learning: Given an observation sequence $Y_{1:T}$ and the initial probabilities π, learn the HMM parameters (A, B).

For directed acyclic graphs (DAGs) with d nodes representing random variables $X_{1:d} = (X_1, \ldots, X_d)$, the joint probability distribution is factorized as

$$P(X_{1:d}) = \prod_{i=1}^{d} P(X_i \mid \mathrm{pa}(X_i)),$$

where $\mathrm{pa}(X_i)$ is the set of nodes that point to X_i ("pa" stands for parents).

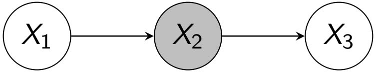
- **?** Write down $P(X_{1:5})$ in the example below:



- **?** How many parameters are needed to identify $P(X_{1:5})$ in the above example?
- **?** Prove that $P(X_{1:d})$ is NOT a proper probability in the presence of cycles.

atomic graphs I

The case of single and two node graphs are trivial. The first level of complexity starts with three-node graphs. We start with the following (Markov) chain configuration[2]:



Reading the joint distribution from the graph, we have

$$P(X_{1:3}) = \prod_{i=1}^{d} P(X_i \mid \mathrm{pa}(X_i)) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2).$$

? Prove $X_1 \perp\!\!\!\perp X_3 \mid X_2$, i.e. given X_2, X_1 and X_3 are independent:

$$P(X_1, X_3 \mid X_2) = P(X_1 \mid X_2)P(X_3 \mid X_2)$$

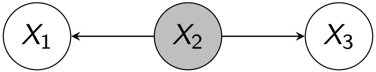
? In addition, show the chain is Markovian:

$$P(X_3 \mid X_{1:2}) = P(X_3 \mid X_2)$$

[2]In the book this is called head-to-tail.

atomic graphs II

We continue with following fan-OUT (tail-to-tail) configuration:



Reading the joint distribution from the graph, we have

$$P(X_{1:3}) = \prod_{i=1}^{d} P(X_i \mid \mathrm{pa}(X_i)) = P(X_1 \mid X_2)P(X_2)P(X_3 \mid X_2).$$

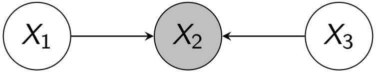
- **?** Prove $X_1 \perp\!\!\!\perp X_3 \mid X_2$, i.e. given X_2, X_1 and X_3 are independent:

$$P(X_1, X_3 \mid X_2) = P(X_1 \mid X_2)P(X_3 \mid X_2)$$

- **?** Argue that X_1 and X_3 are not marginally independent: $X_1 \not\perp\!\!\!\perp X_3$

atomic graphs III

We finish with following fan-IN (head-to-head) configuration:



Reading the joint distribution from the graph, we have

$$P(X_{1:3}) = \prod_{i=1}^{d} P(X_i \mid \mathrm{pa}(X_i)) = P(X_1)P(X_2 \mid X_1, X_3)P(X_3).$$

- **?** Prove $X_1 \perp\!\!\!\perp X_3$, i.e. X_1 are X_3 marginally independent:

$$P(X_1, X_3) = P(X_1)P(X_3)$$

- **?** Argue that after observing X_2, X_1 and X_3 are not guaranteed to be independent[3]:

$$X_1 \not\!\perp\!\!\!\perp X_3 \mid X_2$$

[3] This is sometimes referred to as the *explaining away* phenomenon.

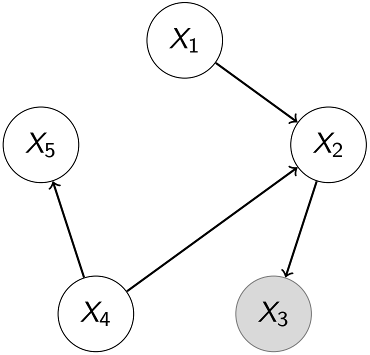
- A, B, and C are arbitrary non-intersecting set of nodes in a general PGM.
- We wish to know whether $A \perp\!\!\!\perp B \mid C$ holds or not.
- Construct all possible (undirected) paths from any node in A to any node in B.
- A path (with respect to C) is set to be blocked if it includes a node such that
  - the arrows form a $\in$ chain/fan-OUT config at the node and the node $\in C$.
  - a fan-IN config at the node, and the node (and all its descendants) $\notin C$
- If all paths from A to B are blocked, then A is **d-separated** from B by C.

Theorem (d-separation)

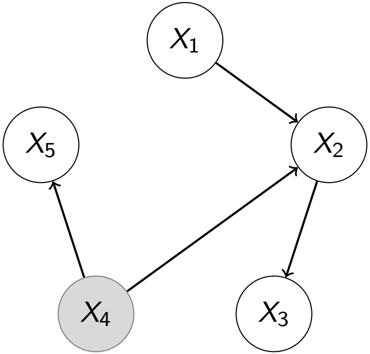
If A is d-separated from B by C on a directed graphical model, then the joint distribution over all the variables in the graph will satisfy

$$A \perp\!\!\!\perp B \mid C$$

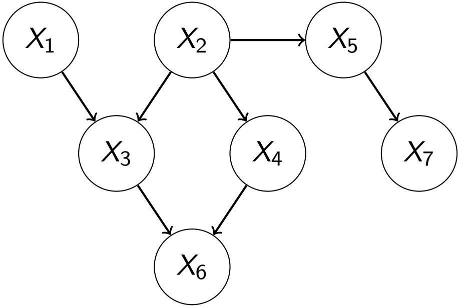
$X_1 \not\perp\!\!\!\perp X_5 \mid X_3$



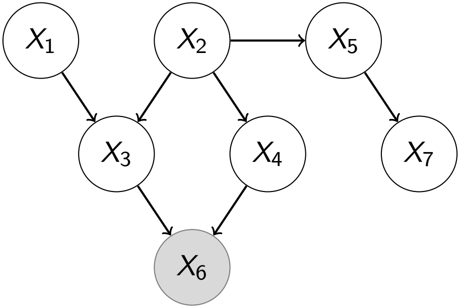
$X_1 \perp\!\!\!\perp X_5 \mid X_4$



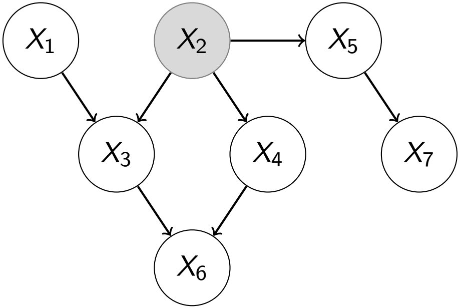
$X_1 \perp\!\!\!\perp X_2$



$X_1 \not\perp\!\!\!\perp X_2 \mid X_6$

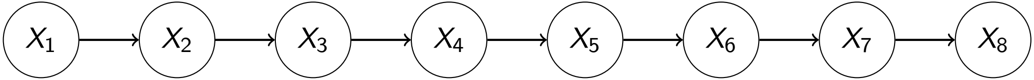


$X_3 \perp\!\!\!\perp X_7 \mid X_2$ **?**



Markov meets Pearl

- Equipped with this graphical language we can represent Markov chains as



? What can we say about $X_{t_1} \perp\!\!\!\perp X_{t_2}$? What about $X_{t_1} \perp\!\!\!\perp X_{t_2} \mid X_{t_3}$?

- The corresponding hidden Markov model is represented by: