PROBABILISTIC REASONING OVER TIME

Today

- Reading
  - AIMA Chapter 15.1-15.2, 15.5
- Goals
  - Case study: Latent Dirichlet allocation
  - Reasoning with uncertainty over time
  - Types of inference
    - Filtering, prediction, smoothing, most likely explanation

Case Study: Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a Bayesian network that describes a hypothetical process of generating a document

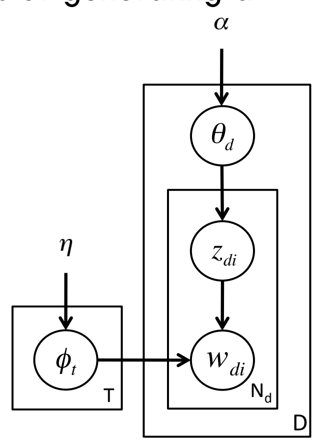
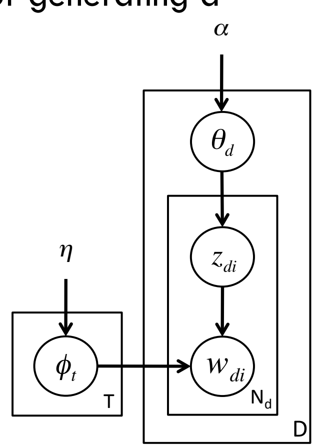


Plate notation is a compact representation of a BN where boxes (i.e. plates) are analogous to for-loops

Case Study: LDA

Latent Dirichlet Allocation is a Bayesian network that describes a hypothetical process of generating a document



- Similarities/differences to past examples?
- What are the independencies encoded in the Bayesian Network?

Case Study: Inference in LDA

- Marginalize out θ and ϕ
- Use Gibbs sampling to draw samples from the posterior distribution:

$$p(z|w) \propto p(z,w)$$

- Each sample is an assignment of words to topics
- We want the most likely assignment, i.e. the assignment of words to topics that has the highest probability

Case Study: Latent Dirichlet Allocation

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

Case Study: Latent Dirichlet Allocation

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Modeling uncertainty over time

- Sometimes, we want to model a *dynamic* process: the value of the random variables change over time
  - Price of a stock
  - Patient stats, e.g. blood pressure, heart rate, blood sugar levels
  - Traffic on California highways
  - Pollution, humidity, temperature, rain fall, storms
  - Sensor tracking and detection

Modeling uncertainty over time

- Tracy got a new job working at the Coop. She works the late shift and doesn't get off until 2am. When she works the late shift, I often observe her eyes are red the next day. But sometimes she stays up late doing homework, and her eyes are red anyways.

- What are questions we might be interested in asking?
- How can we model this domain as a Bayesian network?

Modeling uncertainty over time

- Suppose we also know that if Tracy works the late shift one night she is less likely to work the late shift the next night.

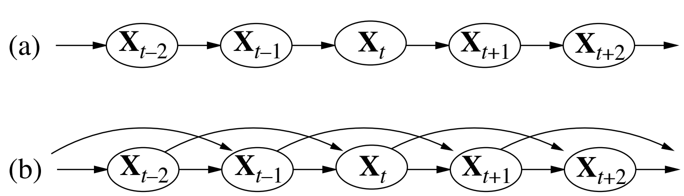
- How does this change the model?

States and Evidence

- Model a dynamic process as a series of time slices
- Each time slice contains a set of random variables
  - We observe the value of some random variables called the evidence. Often denoted as E_t
  - We don't observe the value of some random variables called the state. Often denoted as X_t

Transition Model

- We're often interested in reasoning about the state variables X_t given the history $X_{0:t-1}$
- Markov Assumption: the state variable X_t depends on a *bounded* subset of $X_{0:t-1}$
  - First order Markov Process: $P(X_t | X_{0:t-1}) = P(X_t | X_{t-1})$
  - Second order Markov Process: $P(X_t | X_{0:t-1}) = P(X_t | X_{t-1}, X_{t-2})$



Transition Model

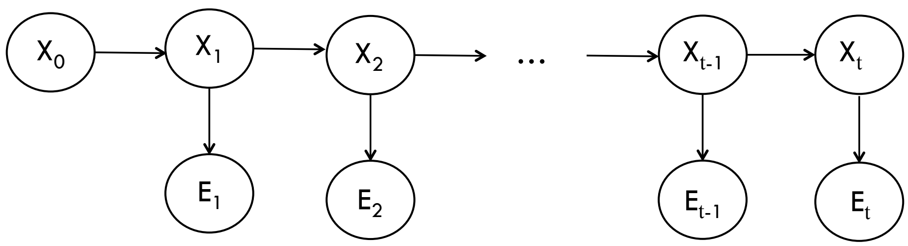
- We're often interested in reasoning about the state variables X_t given the history $X_{0:t-1}$
- Stationarity Assumption: the conditional distribution $P(X_t | X_{t-1})$ is the same for all t
  - Need to specify only one conditional distribution

Sensor (emission) model

- The state variables are responsible for generating (emitting) the evidence variables
- Sensor Markov Assumption: the evidence at time t is independent of every other random variable given the state at time t
  - As a result, your state should encompass all relevant information for specifying the evidence

Hidden Markov Model

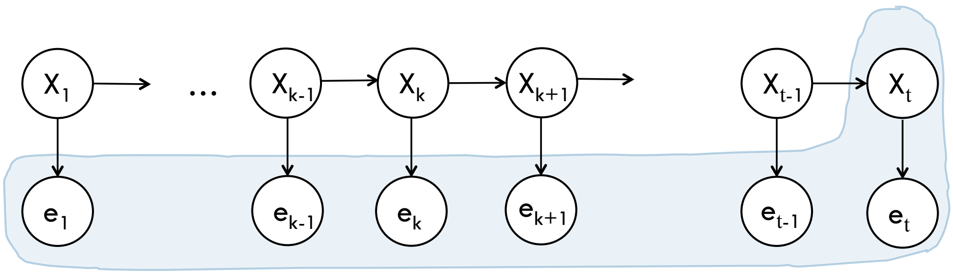
- Hidden Markov Models involve three things:
  - Transition model: $P(X_t | X_{t-1})$
  - Emission (evidence) model: $P(E_t | X_t)$
  - Prior probability: $P(X_0)$



Inference Tasks

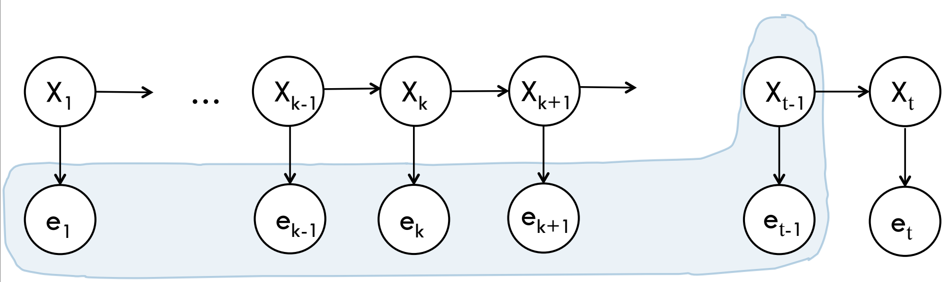
- Filtering: $P(X_t | e_{1:t})$
  - Decision making in the here and now
- Prediction: $P(X_{t+k} | e_{1:t})$
  - Trying to plan the future
- Smoothing: $P(X_k | e_{1:t})$ for $0 \leq k < t$
  - Gives a better (smoother) estimate than filtering by taking into account future evidence
- Most Likely Explanation (MLE): $\underset{x_{1:t}}{\operatorname{argmax}} P(x_{1:t} | e_{1:t})$
  - e.g., speech recognition, sketch recognition

Filtering: $P(X_t \mid e_{1:t})$



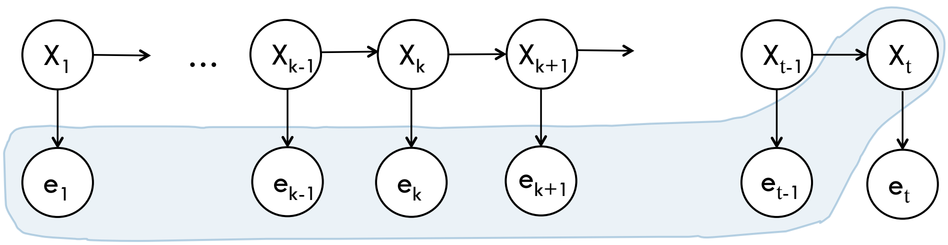
- A recursive state estimation algorithm

Filtering: $P(X_t \mid e_{1:t})$



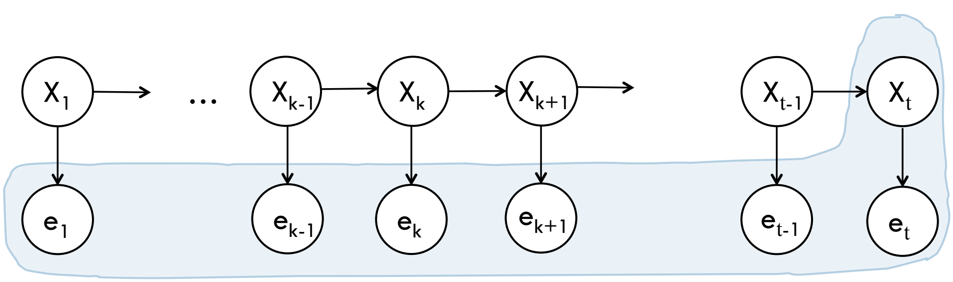
- Assume we already have $p(X_{t-1} \mid e_{1:t-1})$

Filtering: $P(X_t | e_{1:t})$



- Update from state X_{t-1} to X_t

Filtering: $P(X_t | e_{1:t})$



- Then incorporate the new evidence E_t

The Forward Algorithm

$$\begin{aligned} p(X_t|e_{1:t}) &= p(X_t|e_{1:t-1}, e_t) \\ &\propto p(e_t|X_t, e_{1:t-1})\ p(X_t|e_{1:t-1}) \\ &= \underbrace{p(e_t|X_t)}_{\text{Incorporate evidence}}\ \underbrace{p(X_t|e_{1:t-1})}_{\text{Update state}} \end{aligned}$$

The Forward Algorithm

$$\begin{aligned} p(X_t|e_{1:t}) &= p(X_t|e_{1:t-1}, e_t) \\ &\propto p(e_t|X_t, e_{1:t-1})\ p(X_t|e_{1:t-1}) \\ &= p(e_t|X_t)\ p(X_t|e_{1:t-1}) \end{aligned}$$

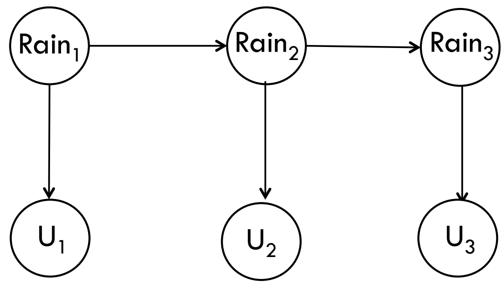
$$\begin{aligned} &= p(e_t|X_t) \sum_{X_{t-1}} p(X_t, X_{t-1}|e_{1:t-1}) \\ &= p(e_t|X_t) \sum_{X_{t-1}} p(X_t|X_{t-1}, e_{1:t-1})\ p(X_{t-1}|e_{1:t-1}) \\ &= \underbrace{p(e_t|X_t)}_{\text{Emission}} \sum_{X_{t-1}} \underbrace{p(X_t|X_{t-1})\ p(X_{t-1}|e_{1:t-1})}_{\text{Transmission + recursion}} \end{aligned}$$

Filtering Example

$p(R_0)$ = <0.5, 0.5>

R_{t-1}	$p(R_t \mid R_{t-1})$
T	0.7
F	0.3

R_t	$p(U_t \mid R_t)$
T	0.9
F	0.2



$$p(X_t|e_{1:t}) \propto p(e_t|X_t) \sum_{X_{t-1}} p(X_t|X_{t-1})\, p(X_{t-1}|e_{1:t-1})$$

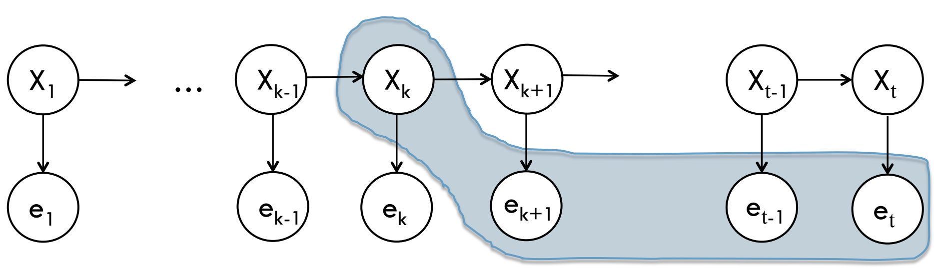
Prediction

- Compute $p(X_{t+k} \mid e_{1:t})$ for $k > 0$
- Given the equations for filtering, can you figure out how to do prediction?

Smoothing: $p(X_k | e_{1:t})$ for $1 \leq k < t$

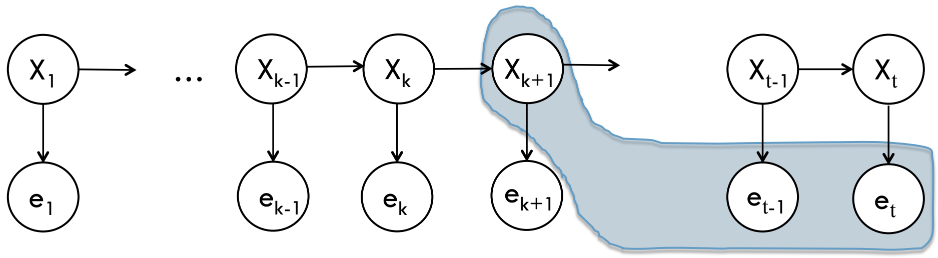
$$
\begin{aligned}
p(X_k|e_{1:t}) &= p(X_k|e_{1:k}, e_{k+1:t}) \\
&\propto p(X_k, e_{k+1:t}|e_{1:k}) \\
&= p(e_{k+1:t}|X_k, e_{1:k})\; p(X_k|e_{1:k}) \\
&= p(e_{k+1:t}|X_k)\; \underbrace{p(X_k|e_{1:k})}_{\text{Forward Algorithm}}
\end{aligned}
$$

The Backward Algorithm



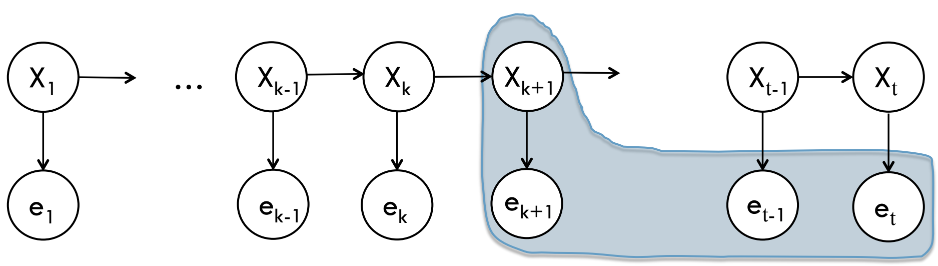
- A recursive state estimation algorithm

The Backward Algorithm



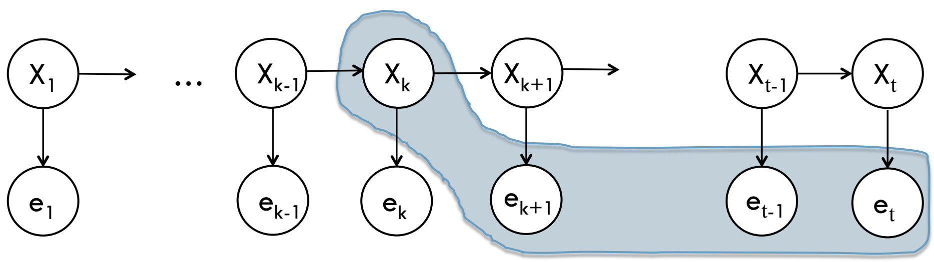
- Assume we have $p(X_{k+1} | e_{k+2:t})$

The Backward Algorithm



- Incorporate evidence via $p(e_{k+1} | X_{k+1})$

The Backward Algorithm



- Update the state via $p(X_{k+1} | X_k)$

Smoothing: $p(X_k | e_{1:t})$ for $1 \leq k < t$

The Forward Backward Algorithm

$$\begin{aligned} p(X_k|e_{1:t}) &= p(X_k|e_{1:k}, e_{k+1:t}) \\ &\propto p(X_k, e_{k+1:t}|e_{1:k}) \\ &= p(e_{k+1:t}|X_k, e_{1:k})\ p(X_k|e_{1:k}) \\ &= p(e_{k+1:t}|X_k)\ \underbrace{p(X_k|e_{1:k})}_{\text{Forward Algorithm}} \end{aligned}$$

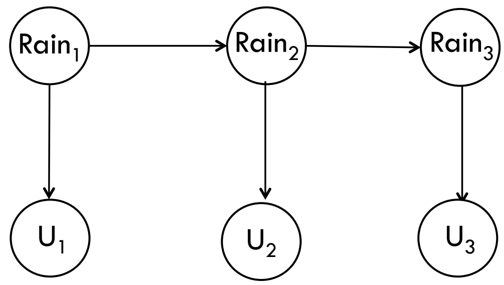
$$\begin{aligned} p(e_{k+1:t}|X_k) &= \sum_{X_k+1} p(e_{k+1:t}, X_{k+1}|X_k) \\ &= \sum_{X_k+1} p(e_{k+1:t}|X_{k+1})\ p(X_{k+1}|X_k) \\ &= \sum_{X_k+1} \underbrace{p(e_{k+1}|X_{k+1})}_{\text{Emission}}\ \underbrace{p(e_{k+2:t}|X_{k+1})}_{\text{Recursion}}\ \underbrace{p(X_{k+1}|X_k)}_{\text{Transmission}} \end{aligned}$$

Smoothing Example

$p(R_0) = \langle 0.5, 0.5 \rangle$

R_{t-1}	$p(R_t \mid R_{t-1})$
T	0.7
F	0.3

R_t	$p(U_t \mid R_t)$
T	0.9
F	0.2



$P(r_1 \mid u_1)$	$P(r_2 \mid u_1, u_2)$	$P(r_1 \mid u_1, u_2)$
0.818	0.883	?

Most Likely Explanation

- Find the state sequence that makes the observed evidence sequence most likely

$$\underset{X_{1:t}}{\operatorname{argmax}}\, P(X_{1:t} \mid e_{1:t})$$

- Recursive formulation:
  - The most likely state sequence for $X_{1:t}$ is the most likely state sequence for $X_{1:t-1}$ followed by the transition to X_t
  - Equivalent to Filtering algorithm except summation replaced with max
  - Called the Viterbi Algorithm

Dynamic Bayesian Networks

- Any BN that represents a temporal probability distribution using state variables and evidence variables is called a Dynamic Bayesian Network

- A Hidden Markov Model is the simplest type of DBN
  - State is represented by a single variable
  - Evidence is represented by a single variable
  - Applications
    - speech recognition
    - handwriting recognition
    - gesture recognition

Approximate Inference in Dynamic BN

- Recall approximate inference algorithms from previous lecture
  - Direct sampling, rejection sampling, likelihood weighting
  - Gibbs sampling

- Likelihood weighting applied to DBN (with some modifications) is known as a Particle filter

Particle Filtering

- Likelihood weighting fixes the evidence variables and samples only the non-evidence variables
- Introduces a weight to correct for the fact that we're sampling from the prior distribution instead of the posterior distribution

$$\text{weight} = p(e_1 \mid \text{Parents}(e_1)) * p(e_2 \mid \text{Parents}(e_2)) \ldots$$

Particle Filtering

- Initialize
  - Draw N particles (i.e. samples) for X_0 from the prior distribution $p(X_0)$
- Propagate
  - Propagate each particle forward by sampling $X_{t+1} \mid X_t$
- Weight
  - Weight each particle by $p(e_{t+1} \mid X_{t+1})$
- Resample
  - Generate N new particles by sampling proportional to the weights. The new particles are unweighted

Particle Filtering

- Particles: track samples of states rather than an explicit distribution