COMP90051 **Statistical Machine Learning**

Semester 2, 2016

Lecturer: Trevor Cohn

21. Independence in PGMs; Example PGMs

THE UNIVERSITY OF MELBOURNE

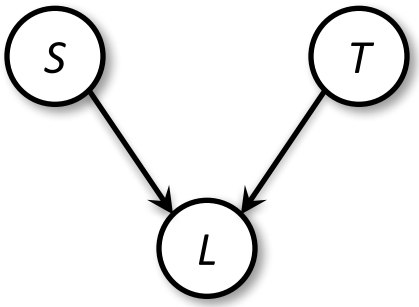
Independence

PGMs encode assumption of statistical independence between variables.

Critical to understanding the capabilities of a model, and for efficient inference.

Recall: Directed PGM

- Nodes
- Edges (acyclic)



- Random variables
- Conditional dependence
  - Node table: $\Pr(child|parents)$
  - Child directly depends on parents
- Joint factorisation

$$\Pr(X_1, X_2, \ldots, X_k) = \prod_{i=1}^{k} \Pr\left(X_i | X_j \in parents(X_i)\right)$$

Graph encodes:

- independence assumptions
- parameterisation of CPTs

Independence relations (D-separation)

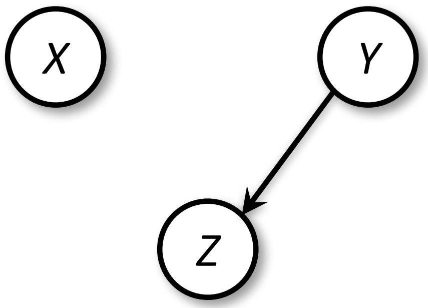
- Important *independence* relations between RV's
  - ***Marginal independence*** *P(X, Y) = P(X) P(Y)*
  - ***Conditional independence*** *P(X, Y | Z) = P(X | Z) P(Y | Z)*
- Notation $\boldsymbol{A \perp B \mid C}$**:**
  - *RVs in set A are independent of RVs in set B, when given the values of RVs in C.*
  - *Symmetric: can swap roles of A and B*
  - $\boldsymbol{A \perp B}$ denotes marginal independence, $C = \emptyset$
- Independence captured in graph structure
  - *Caveat*: dependence does not follow *in general* when X and Y are not independent

Marginal Independence

- Consider graph fragment

- What [marginal] independence relations hold?
  - $X \perp Y$?

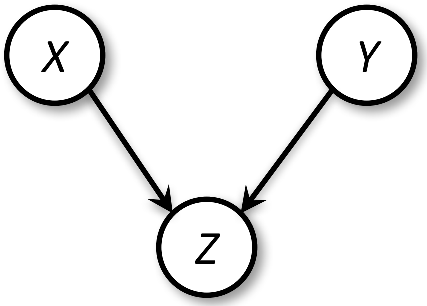
    Yes – $P(X, Y) = P(X)\ P(X)$
- What about $X \perp Z$, where Z connected to Y?



Marginal Independence

- Consider graph fragment

  Marginal independence denoted $\boldsymbol{X \perp Y}$

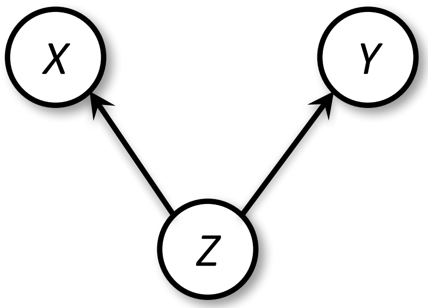


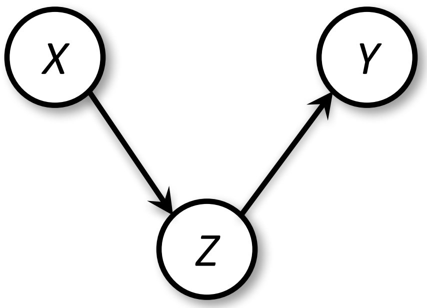
- What [marginal] independence relations hold?
  * $X \perp Z$?

    No – $P(X,Z) = \sum_Y P(X)P(Y)P(Z|X,Y)$
  * $X \perp Y$?

    Yes – $P(X,Y) = \sum_Z P(X)P(Y)P(Z|X,Y) = P(X)P(Y)$

Marginal Independence





Are X and Y marginally dependent? (X ⊥ Y?)

$P(X,Y) = \sum_Z P(Z)P(X|Z)P(Y|Z)$ … No

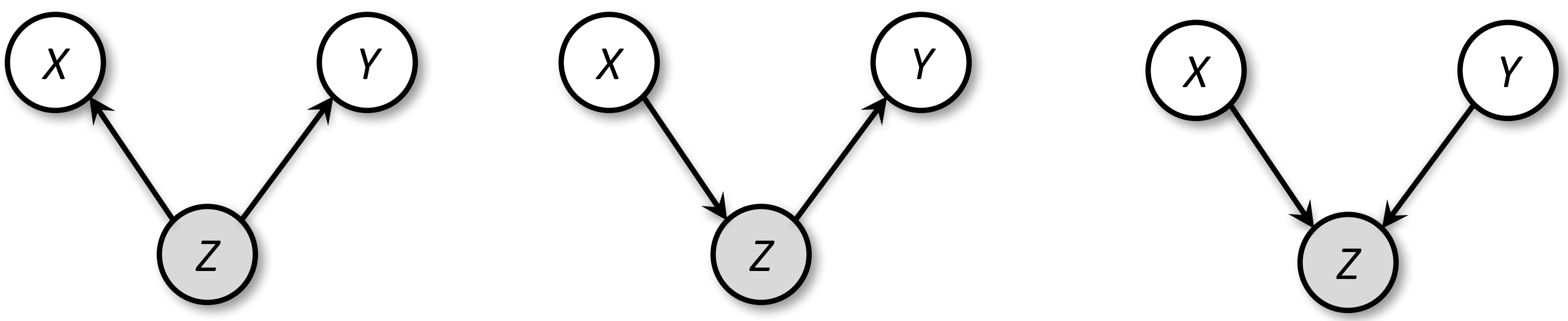
$P(X,Y) = \sum_Z P(X)P(Z|X)P(Y|Z)$ … No

Marginal Independence

- Marginal independence **can** be read off graph
  - however, must account for edge directions
  - relates (loosely) to *causality*:
    if edges encode causal links, can X affect (cause) Y?
- General rules, X and Y are linked by:
  - no edges, in any direction → independent
  - intervening node with incoming edges from X and Y (aka *head-to-head*) → independent
  - *head-to-tail, tail-to-tail* → not (necessarily) independent
- … generalises to longer chains of intermediate nodes (coming)

Conditional independence

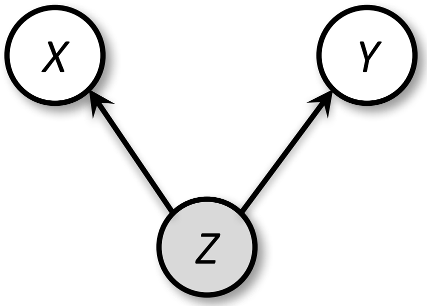
- What if we know the value of some RVs? How does this affect the in/dependence relations?
- Consider whether $X \perp Y | Z$ in the canonical graphs

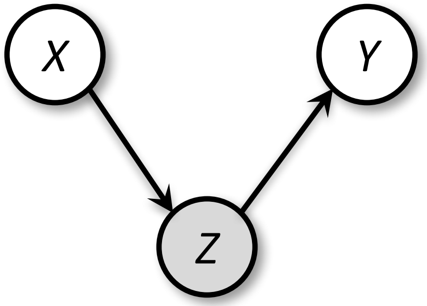


  - Test by trying to show P(X,Y|Z) = P(X|Z) P(Y|Z).

Conditional independence

$$P(X,Y|Z) = \frac{P(Z)P(X|Z)P(Y|Z)}{P(Z)}$$
$$= P(X|Z)P(Y|Z)$$

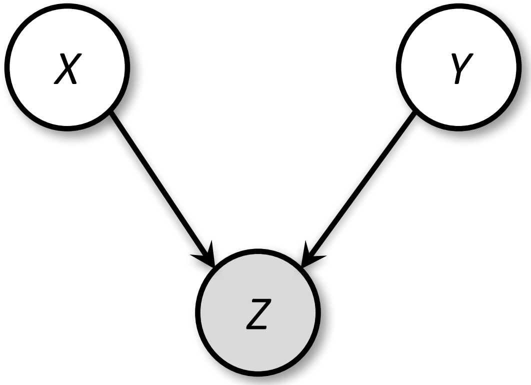




$$P(X,Y|Z) = \frac{P(X)P(Z|X)P(Y|Z)}{P(Z)}$$
$$= \frac{P(X|Z)P(Z)P(Y|Z)}{P(Z)}$$
$$= P(X|Z)P(Y|Z)$$

Conditional independence

- So far, just graph separation… Not so fast!
  - cannot factorise the last canonical graph

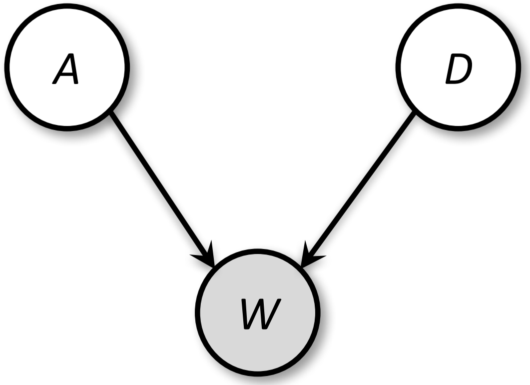


- Known as explaining away: value of Z can give information linking X and Y
  - E.g., X and Y are binary coin flips, and Z is whether they land the same side up. Given Z, then X and Y become completely dependent (deterministic).
  - A.k.a. Berkson's paradox

N.b., Marginal dependence ≠ conditional independence!

Explaining away

- The washing has fallen off the line (W). Was it aliens (A) playing? Or next door's dog (D)?



A	Prob
0	0.999
1	0.001

D	Prob
0	0.9
1	0.1

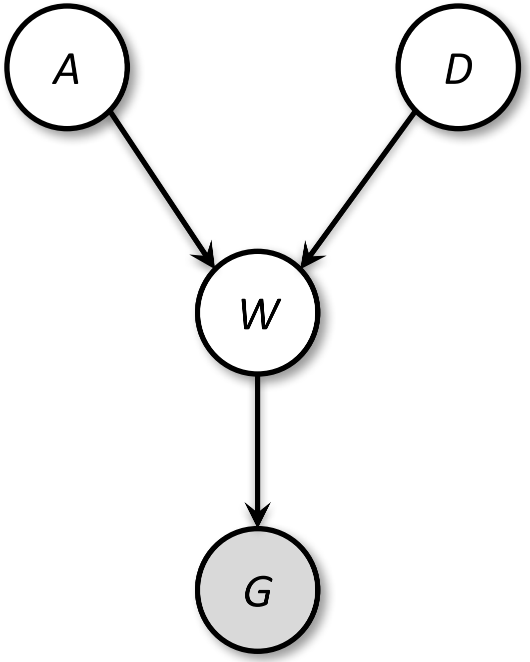
A	D	P(W=1 |A,D)
0	0	0.1
0	1	0.3
1	0	0.5
1	1	0.8

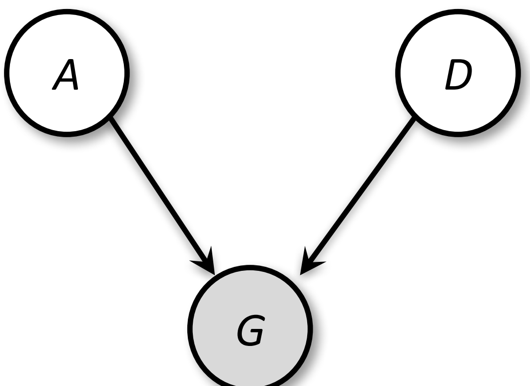
- Results in conditional posterior
  - P(A=1|W=1) = 0.004
  - P(A=1|D=1,W=1) = 0.003
  - P(A=1|D=0,W=1) = 0.005

Explaining away II

- Explaining away also occurs for *observed children* of the head-head node
  - attempt factorise to test A ⊥ D | G

$$P(A, D|G) = \sum_W P(A)P(D)P(W|A, D)P(G|W)$$
$$= P(A)P(D)P(G|A, D)$$





"D-separation" Summary

- Marginal and cond. independence can be read off graph structure
  * marginal independence relates (loosely) to *causality*: if edges encode causal links, can X affect (cause or be caused by) Y?
  * conditional independence less intuitive
- How to apply to larger graphs?
  * based on paths separating nodes, i.e., do they contain nodes with head-to-head, head-to-tail or tail-to-tail links?
  * can all [undirected!] paths connecting two nodes be blocked by an independence relation?

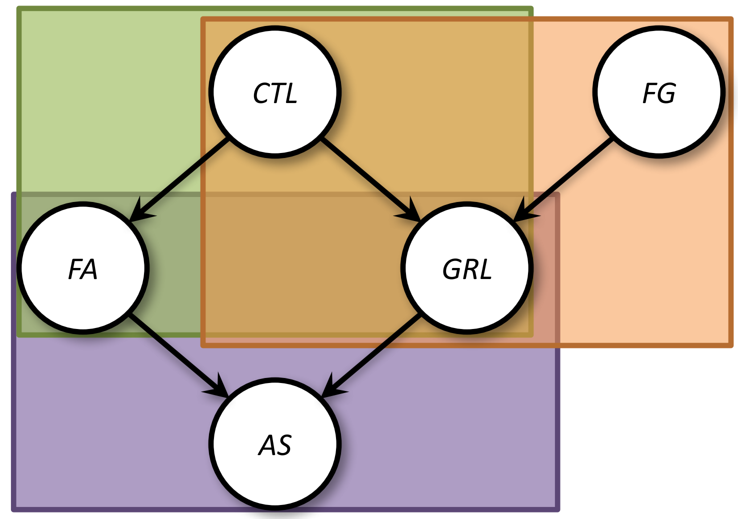
D-separation in larger PGM

- Consider pair of nodes
  FA $\perp$ FG?

  Paths:

  FA – CTL – GRL – FG

  FA – AS – GRL – FG



- Paths can be blocked by independence
- More formally see "**Bayes Ball**" algorithm which formalises notion of d-separation as reachability in the graph, subject to specific traversal rules.

What's the point of d-separation?

- Designing the graph
  - understand what independence assumptions are being made; not just the obvious ones
  - informs trade-off between expressiveness and complexity
- Inference with the graph
  - computing of conditional / marginal distributions must respect in/dependences between RVs
  - affects complexity (space, time) of inference

Markov Blanket

- For an RV what is the minimal set of other RVs that make it *conditionally independent* from the rest of the graph?
  - what conditioning variables can be safely dropped from $P(X_j \mid X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$?
- Solve using d-separation rules from graph
- Important for predictive inference (e.g., in pseudolikelihood, Gibbs sampling, etc)

Undirected PGMs

Undirected variant of PGM, parameterised by arbitrary positive valued functions of the variables, and global normalisation.

A.k.a. Markov Random Field.

Undirected vs directed

Undirected PGM

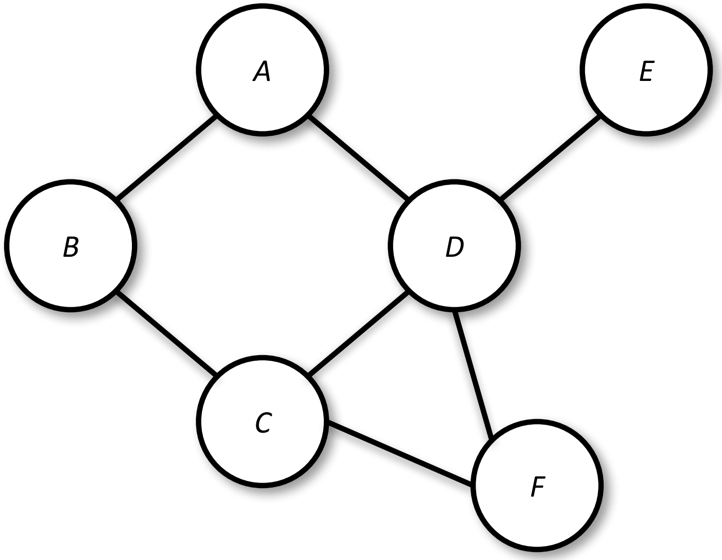
- Graph
  - Edges undirected
- Probability
  - Each node a r.v.
  - Each clique C has "factor" $\psi_C(X_j : j \in C) \geq 0$
  - Joint $\propto$ product of factors

Directed PGM

- Graph
  - Edged directed
- Probability
  - Each node a r.v.
  - Each node has conditional $p(X_i | X_j \in parents(X_i))$
  - Joint $=$ product of cond'ls

Key difference = normalisation

Undirected PGM formulation



- Based on notion of
  - *Clique: a set of fully connected nodes (e.g., A-D, C-D, C-D-F)*
  - *Maximal clique: largest cliques in graph (not C-D, due to C-D-F)*
- Joint probability defined as

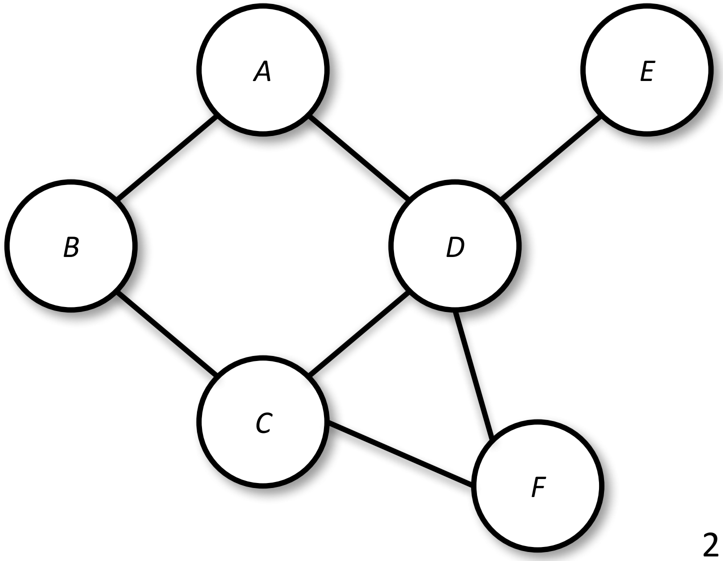
$$P(a,b,c,d,e,f) = \frac{1}{Z}\psi_1(a,b)\psi_2(b,c)\psi_3(a,d)\psi_4(d,c,f)\psi_5(d,e)$$

  - where ψ is a positive function and Z is the normalising 'partition' function

$$Z = \sum_{a,b,c,d,e,f} \psi_1(a,b)\psi_2(b,c)\psi_3(a,d)\psi_4(d,c,f)\psi_5(d,e)$$

d-separation in U-PGMs

- Good news! Simpler dependence semantics
  - conditional independence relations = graph connectivity
  - if all paths between nodes in set X and Y pass through an observed nodes Z then $X \perp Y | Z$
- For example $B \perp D | \{A, C\}$
- Markov blanket of node = its immediate neighbours



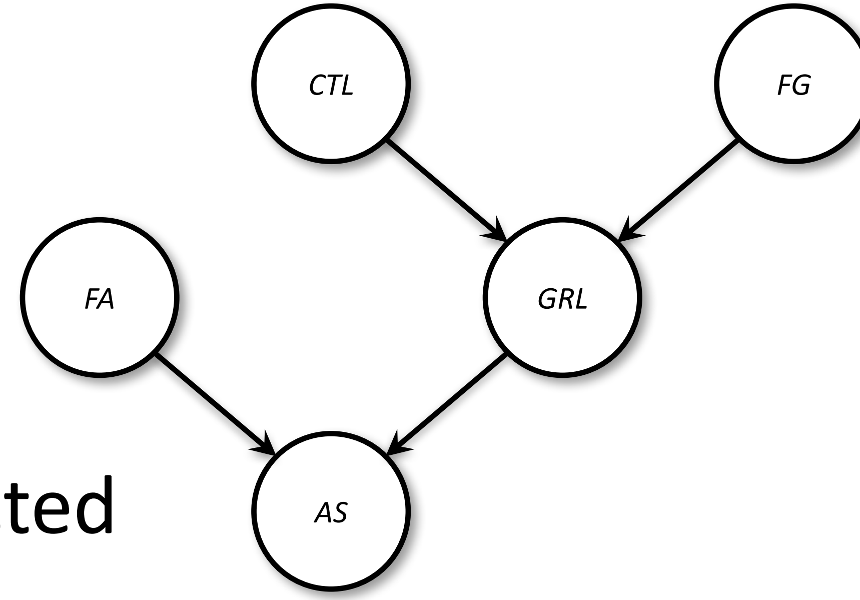
Directed to undirected

- Directed PGM formulated as

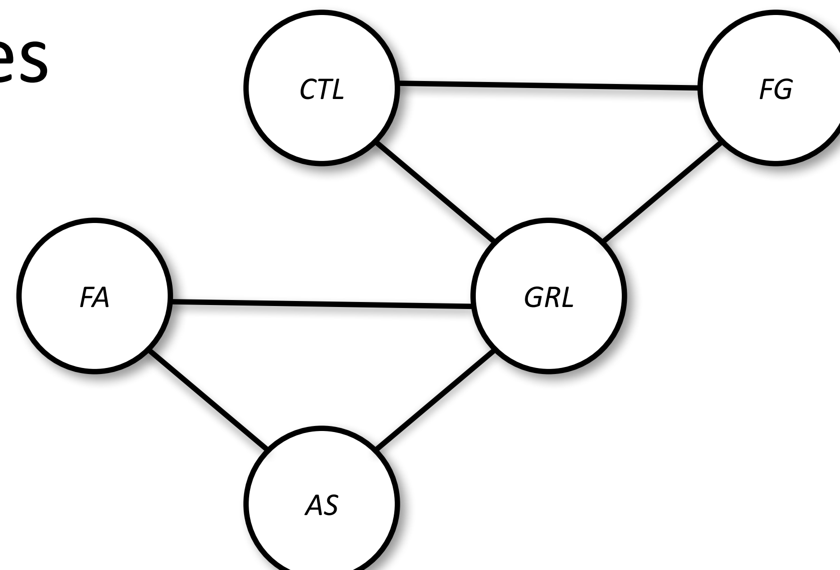
$$P(X_1, X_2, \ldots, X_k) = \prod_{i=1}^{k} Pr(X_i | X_{\pi_i})$$

where **π** indexes parents.

- Equivalent to U-PGM with
  * each conditional probability term is included in one factor function, ψ_c
  * clique structure links *groups of variables,* i.e., $\{\{X_i\} \cup X_{\pi_i}, \forall i\}$
  * normalisation term trivial, Z = 1



1. copy nodes
2. copy edges, undirected
3. 'moralise' parent nodes



Why U-PGM?

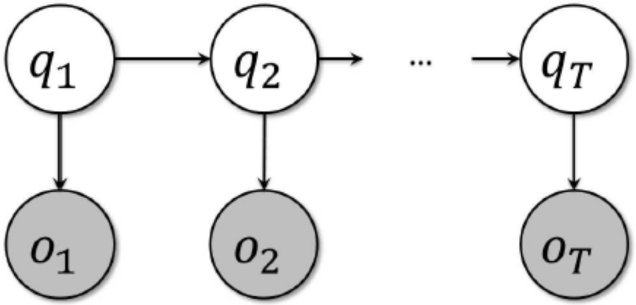
- Pros
  - generalisation of D-PGM
  - simpler means of modelling without the need for per-factor normalisation
  - general inference algorithms use U-PGM representation (supporting both types of PGM)
- Cons
  - (slightly) weaker independence
  - calculating global normalisation term (Z) intractable in general (but tractable for chains/trees, e.g., CRFs)

Example PGMs

The hidden Markov model (HMM); lattice Markov random field (MRF)

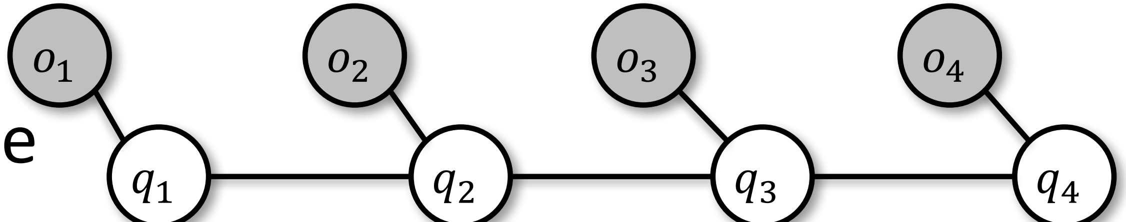
The HMM (and Kalman Filter)

- Sequential observed outputs from hidden state



$A = \{a_{ij}\}$ transition probability matrix; $\forall i : \sum_j a_{ij} = 1$
$B = \{b_i(o_k)\}$ output probability matrix; $\forall i : \sum_k b_i(o_k) = 1$
$\Pi = \{\pi_i\}$ the initial state distribution; $\sum_i \pi_i = 1$

- The Kalman filter same with continuous Gaussian r.v.'s
- A CRF is the undirected analogue

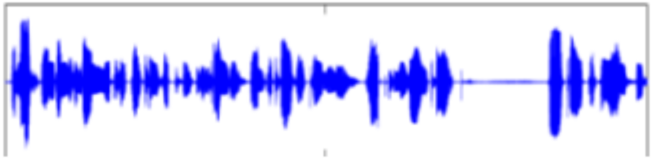


HMM Applications

- NLP – part of speech tagging: given words in sentence, infer hidden parts of speech

  "I love Machine Learning" → noun, verb, noun, noun

- Speech recognition: given waveform, determine phonemes



- Biological sequences: classification, search, alignment
- Computer vision: identify who's walking in video, tracking

Fundamental HMM Tasks

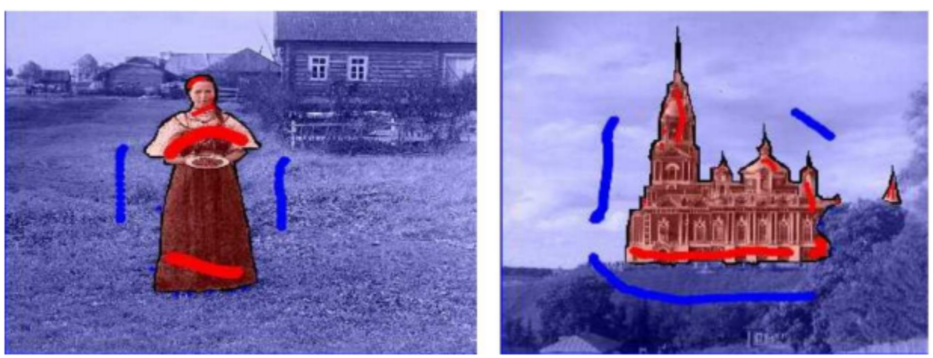
HMM Task	PGM Task
Evaluation. Given an HMM μ and observation sequence O, determine likelihood $\Pr(O\|\mu)$	Probabilistic inference
Decoding. Given an HMM μ and observation sequence O, determine most probable hidden state sequence Q	MAP point estimate
Learning. Given an observation sequence O and set of states, learn parameters A, B, Π	Statistical inference

Computer Vision

Hidden square-lattice Markov random fields

Pixel labelling tasks in Computer Vision

Semantic labelling (Gould et al. 09)

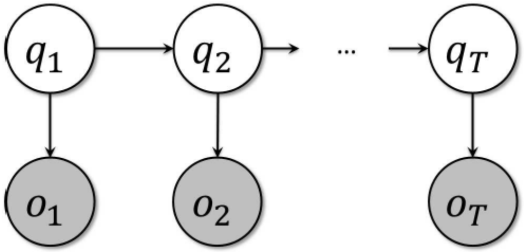


Interactive figure-ground segmentation (Boykov & Jolly 2011)

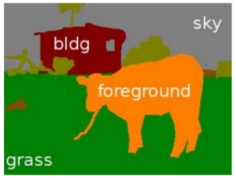
Denoising (Felzenszwalb & Huttenlocher 04)

What these tasks have in common

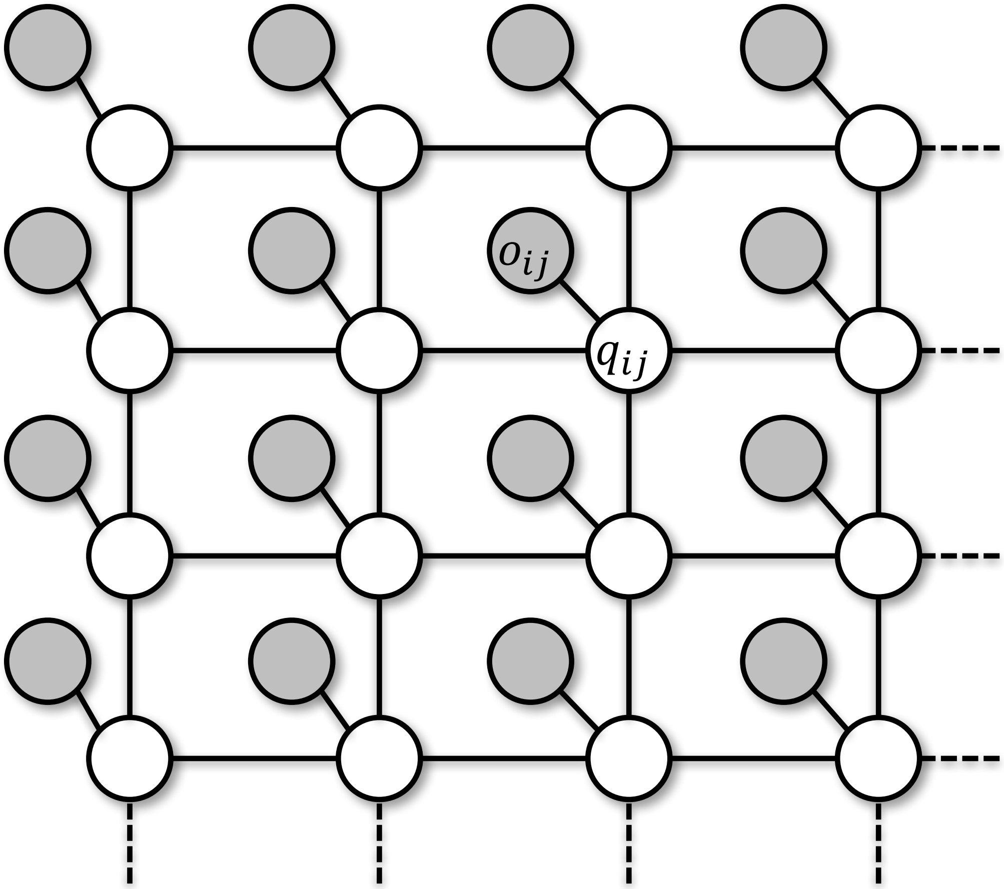
- Hidden state representing semantics of image
  - Semantic labelling: Cow vs. tree vs. grass vs. sky vs. house
  - Fore-back segment: Figure vs. ground
  - Denoising: Clean pixels
- Pixels of image
  - What we observe of hidden state
- Remind you of HMMs?



A hidden square-lattice MRF

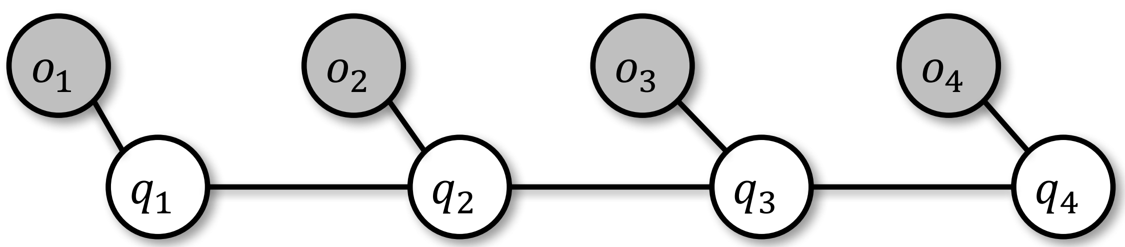
- Hidden states: square-lattice model
  - Boolean for two-class states 
  - Discrete for multi-class 

  - Continuous for denoising 
- Pixels: observed outputs
  - Continuous e.g. Normal



Application to sequences: CRFs

- Conditional Random Field: Same model applied to sequences
  - observed outputs are words, speech, amino acids etc
  - states are tags: part-of-speech, phone, alignment…
- CRFs are discriminative, model *P(Q|O)*
  - versus HMM's which are generative, *P(Q,O)*
  - undirected PGM more general and expressive



Summary

- Notion of independence, ‘d-separation’
  * marginal vs conditional independence
  * explaining away, Markov blanket
  * undirected PGMs & relation to directed PGMs
- Share common training & prediction algorithms (coming up next!)