Lecture 19. PGM Independence; Example PGMs

COMP90051 Statistical Machine Learning

Semester 1, 2021
Lecturer: Trevor Cohn

Outline

- Formalising independence and graph structure
  - Independence and conditional independence
  - Explaining away
  - Notions of 'd-separation' & Markov blanket
- Example PGMs and applications

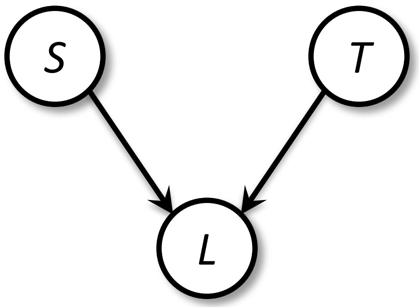
Independence

PGMs encode assumption of statistical independence between variables.

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

Recap: Directed PGM

- Nodes
- Edges (acyclic)



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

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

Graph encodes:
- independence assumptions
- parameterisation of CPTs

Independence relations

- Important *independence* relations between RV's
  - *Marginal independence* $P(X, Y) = P(X)\, P(Y)$
  - *Conditional independence* $P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$
- Notation $A \perp B \mid C$:
  - 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*
  - $A \perp B$ denotes marginal independence, i.e., $C = \emptyset$

Reading independence off PGM

- Independence relations captured in graph structure
- How to read these independence relations off the graph?
  - PGM specifies the form of the joint distribution, can simply write out the joint and attempt to factorise. If joint factorises then we have independence
  - Even easier, independence can be read off graph directly, based on connecting paths between nodes
- **Caveat*:*** in general, "*A, B are **not** independent*" is not the same as saying "*A, B dependent*"

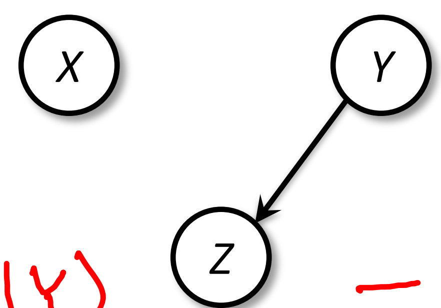
Marginal Independence

- Consider graph fragment

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

    Yes – $P(X, Y) = P(X)\,P(Y)$

- What about $X \perp Z$, where Z connected to Y?



$P(X,Z) = \sum_Y P(X)P(Y)P(Z|Y)$

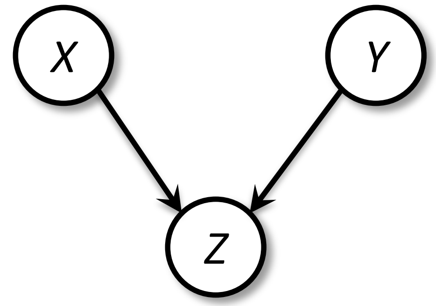
$= P(X) \sum_Z P(Y)P(Z|Y)$

– Yes

Marginal Independence

- Consider graph fragment

  *Marginal independence denoted **X⊥Y***



- What [marginal] independence relations hold?
  * X ⊥ Z?

    $P(X,Z) = \sum_Y P(X)\,P(Y)\,P(Z|X,Y)$ — No

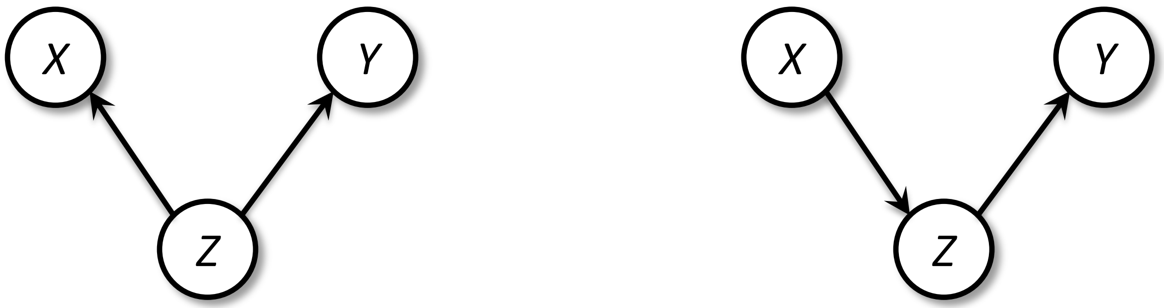
  * X ⊥ Y?

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

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

    $= P(X)\,P(Y)$ — Yes

Marginal Independence



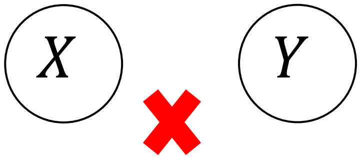
Are X and Y marginally dependent? (X ⊥ Y?)

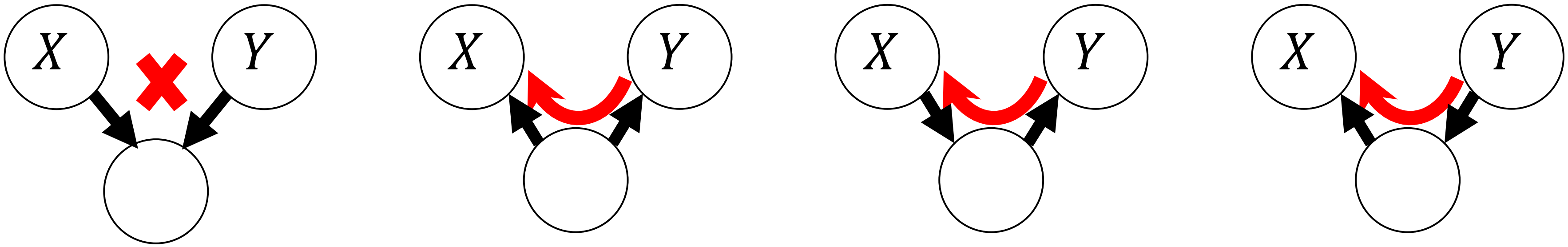
$P(X, Y) = \sum_Z P(Z)P(X|Z)P(Y|Z)$ … No

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

Marginal Independence

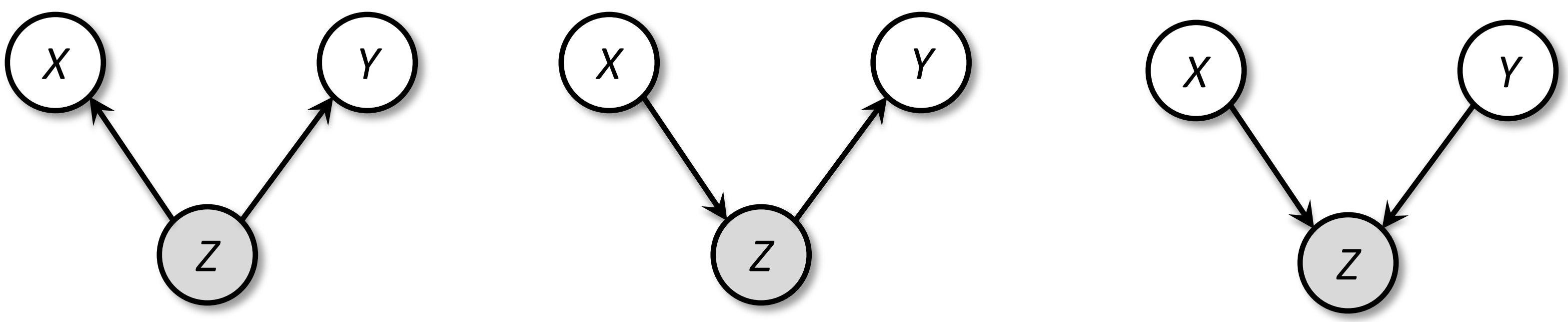
- Marginal independence **can** be read off graph
  - must account for edge directions
  - relates to *causality*: if edges encode causal links, can X *cause* or be *caused by* Y?
- Summary (*thus far*): Are X and Y independent?





Conditional independence

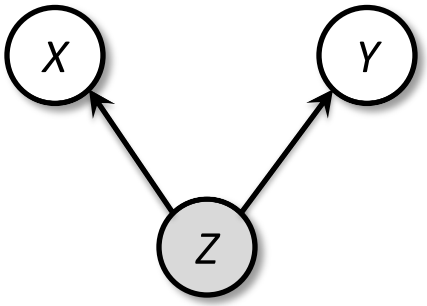
- What if we know the value of some RVs? How does this affect the (in)dependence relations?
- Consider whether $X \perp Y | Z$ in the canonical graphs

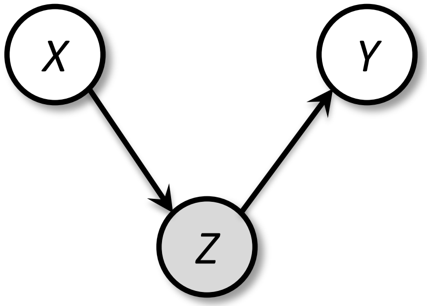


  - Shaded node = observed RV
  - Test by trying to show P(X,Y|Z) = P(X|Z) P(Y|Z).

Conditional independence

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

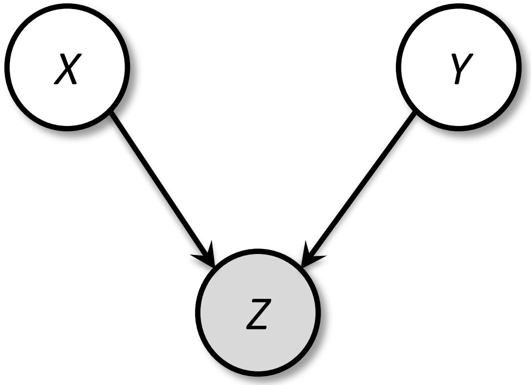




$$P(X,Y|Z) = \frac{P(X)P(Z|X)P(Y|Z)}{P(Z)}$$
$$= \frac{P(X|Z)P(Z)P(Y|Z)}{P(Z)}$$
$$= P(X|Z)P(Y|Z)$$

Explaining away

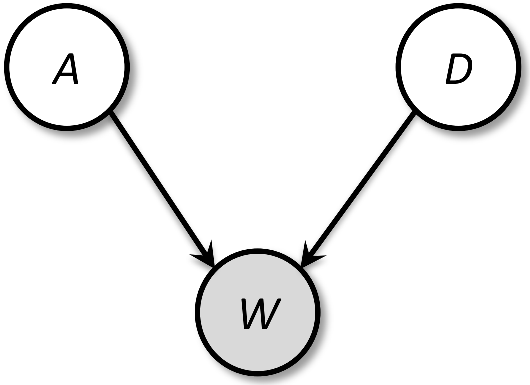
- So far, just graph separation… Not so fast!
  - cannot factorise the last canonical graph



- 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

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

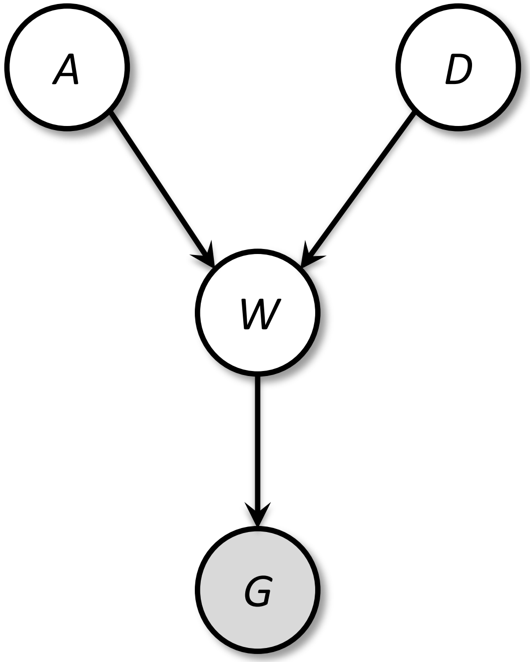
D	Prob
0	0.9
1	0.1

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

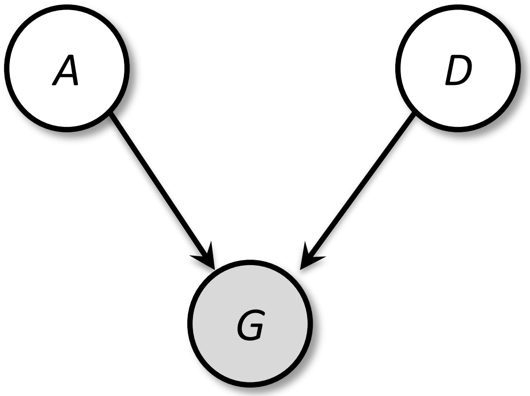
Explaining away II

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



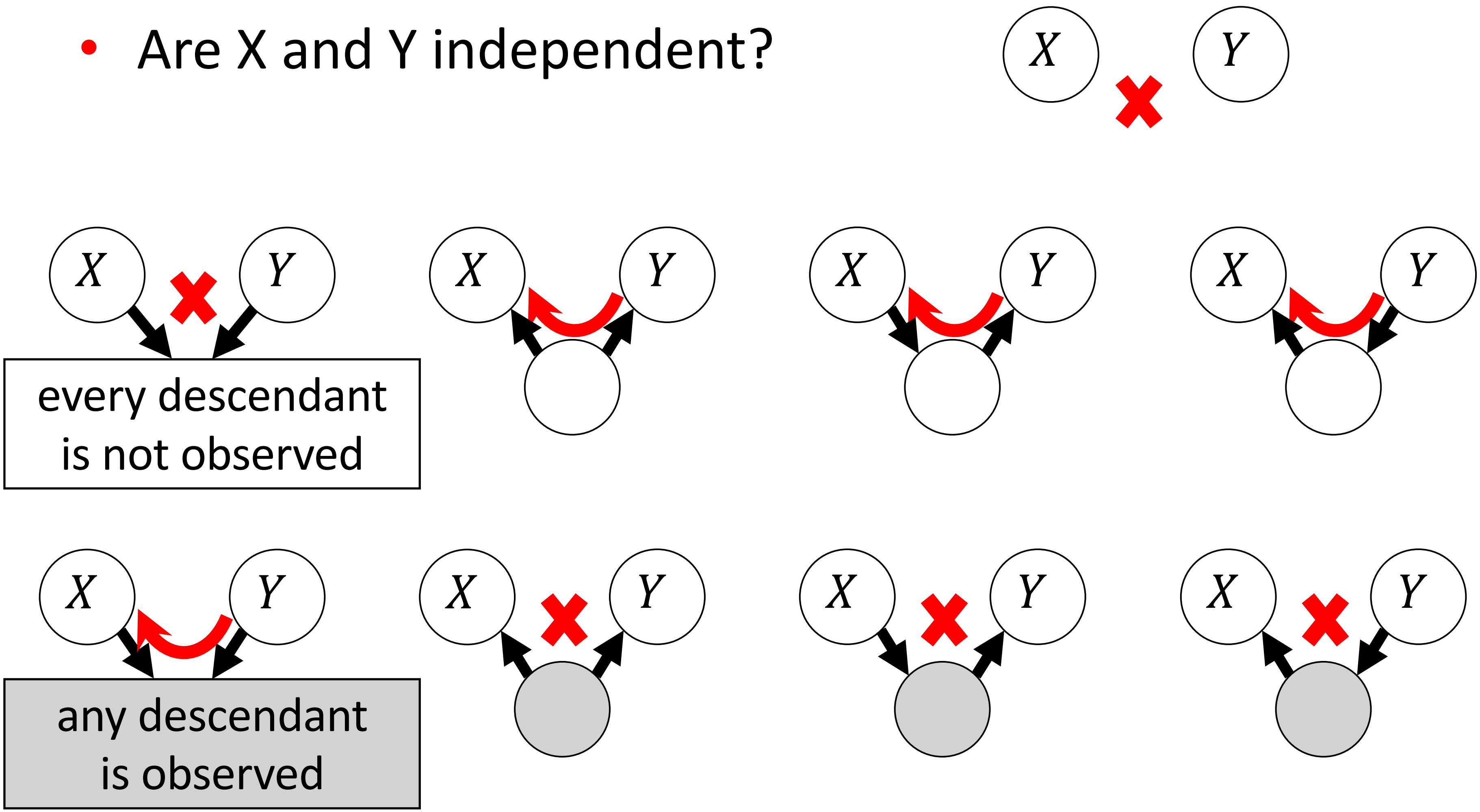
$$P(A, D|G) = \sum_W P(A)P(D)P(W|A, D)P(G|W)$$
$$= P(A)P(D)P(G|A, D)$$

*Effective PGM after marginalising over **W***



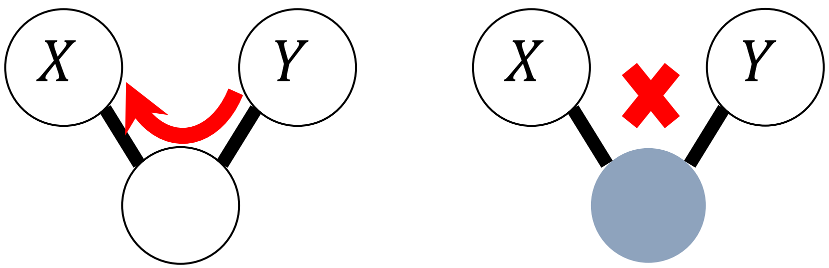
Summary: Independence in directed PGMs

- Are X and Y independent?



Independence in ***undirected*** PGMs

- Are X and Y independent?
  - Look at paths connecting Y and X; no path → independent



Mini Summary

- Notion of independence
  * marginal vs conditional independence
  * explaining away

Up next: a deeper dive into independence

D-Separation

Packaging the ideas of graph connectivity and independence into an algorithm which can be applied to a whole PGM

D-separation

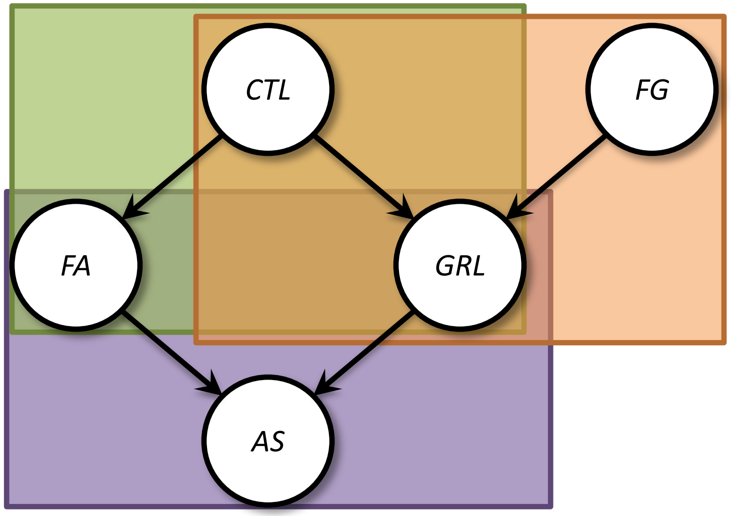
- Marginal and cond. independence can be read off graph structure for simple *canonical* fragments
- D-separation allows application of these ideas to larger graphs
  - based on *paths* separating nodes: do they include components which are known independence relations
  - can all *paths* be blocked by an independence relation?

D-separation in larger PGM

- Consider pair of nodes
  FA ⊥ FG?

  Paths:

  FA – CTL – GRL – FG
  FA – AS – GRL – FG



- Both 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)

Mini Summary

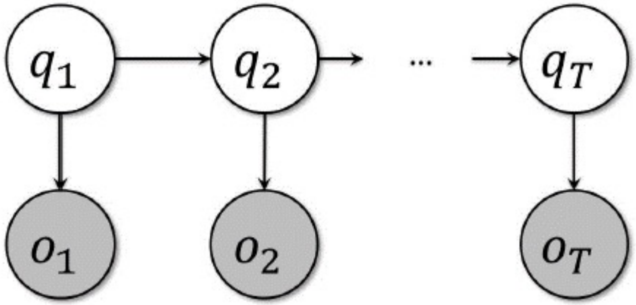
- Formalising independence and graph structure
  - Notions of ‘d-separation’;
  - Markov blanket
- Up next: example PGMs

Example PGMs

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

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; $\sum_i \pi_i = 1$

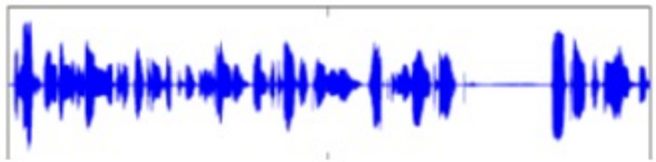
- The Kalman filter same with continuous Gaussian r.v.'s

HMM Applications

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

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

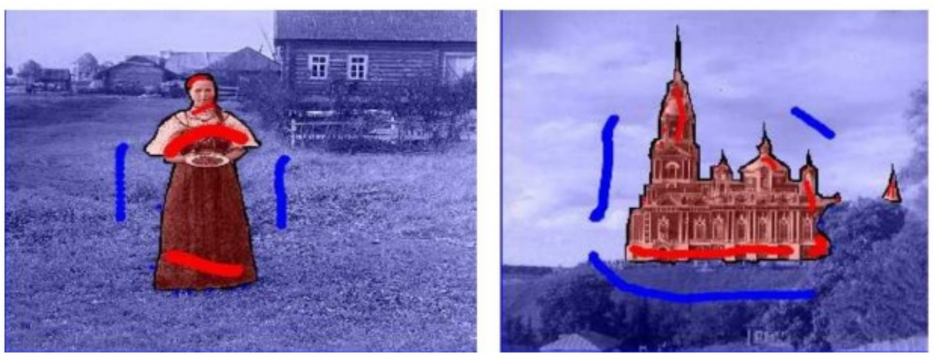
- Speech recognition: given waveform, determine phonemes



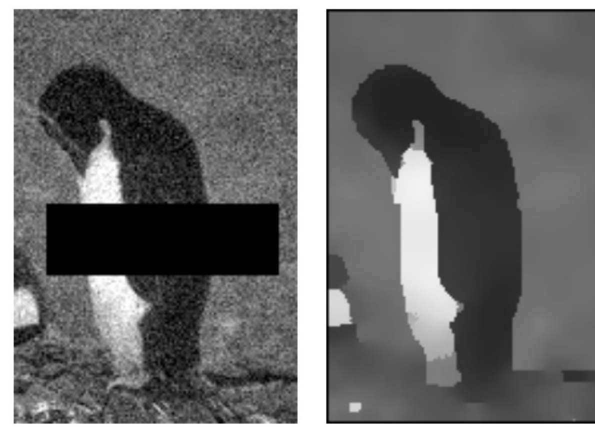
- Biological sequences: classification, search, alignment
- Computer vision: identify who's walking in video, tracking

Pixel labelling tasks in Computer Vision

Semantic labelling (Gould et al. 09)



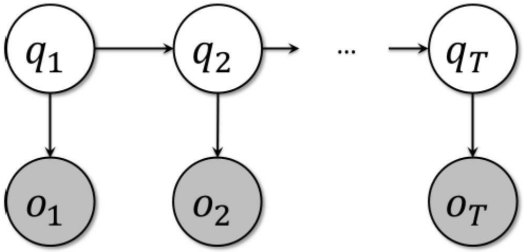
Interactive figure-ground segmentation (Boykov & Jolly 2011)



Denoising (Felzenszwalb & Huttenlocher 04)

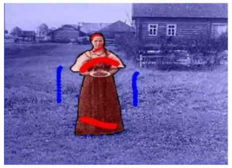
What these tasks have in common

- 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?

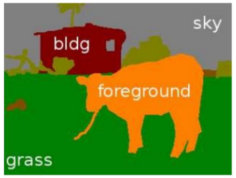


A hidden square-lattice Markov random field

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

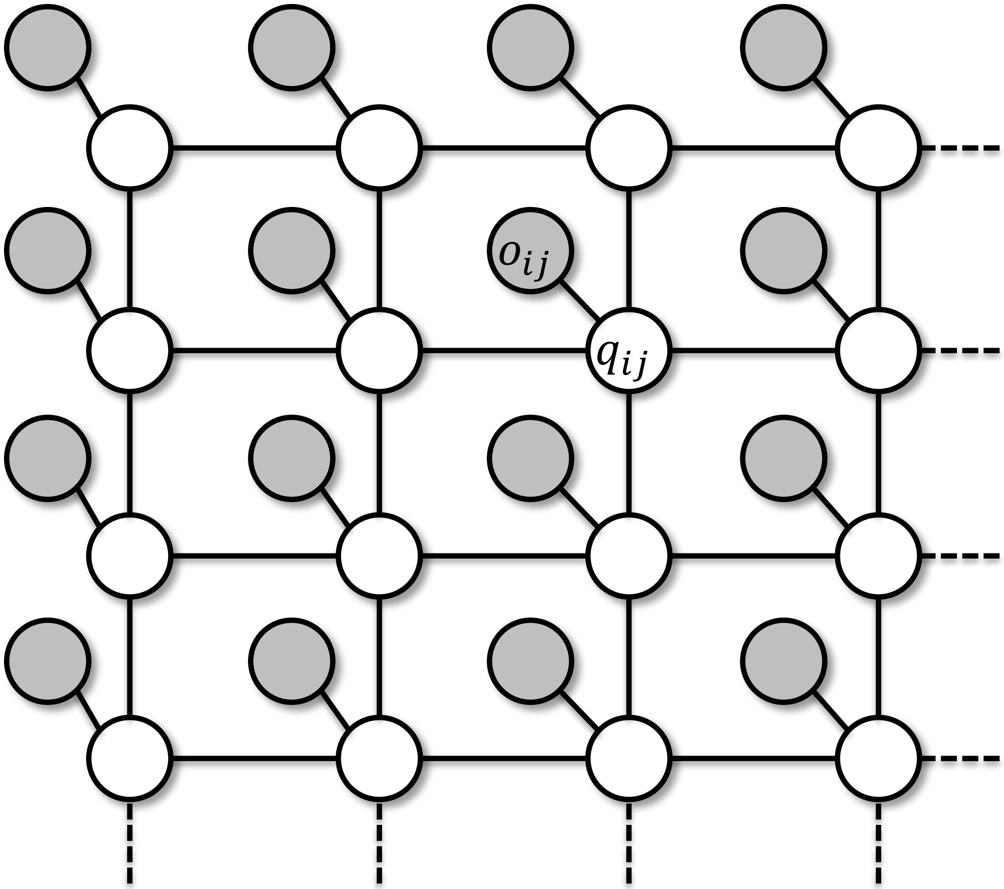


  - Discrete for multi-class



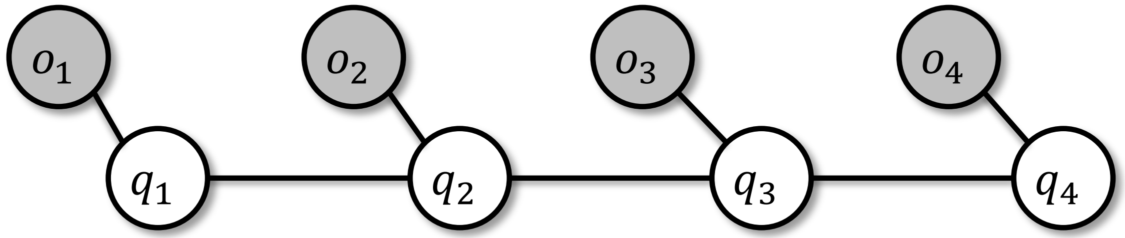
  - Continuous for denoising

- 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

- Formalising independence and graph structure
  - Independence and conditional independence
  - Explaining away
  - Notions of ‘d-separation’ & Markov blanket
- Example PGMs and applications

Next time: elimination for probabilistic inference