COMP 64101
Reasoning and Learning under Uncertainty

Omar Rivasplata

Department of Computer Science, University of Manchester
Manchester Centre for AI Fundamentals

Lecture 3 - Part 1

MANCHESTER
1824

Manchester Centre for
AI Fundamentals

Topics: Probabilistic Graphical Models (PGMs)

I know it's [Tuesday]. It's a good day for math!

–Inspired by Max Mintz, UPenn.

- Intro to PGMs (§ 4.1)
- Directed graphical models (Bayes nets) (§ 4.2)
- Undirected graphical models (Markov random fields) (§ 4.3)
- More goodies on PGMs (§ 4.4 – 4.7)

Nota bene: Section numbers refer to Murphy (2023).

4.1 Introduction

What is a Probabilistic Graphical Model (PGM)?

- PGMs are underpinned by graph structures to account for conditional dependence and independence in a set of random variables:
  - Nodes represent random variables.
  - Edges represent conditional dependence.
  - Lack of edges represents conditional independence (CI).
- Some kinds of graphical models:
  - directed graphical models.
  - undirected graphical models.
  - some combination of directed and undirected.
- PGMs provide a convenient formalism for defining joint distributions on sets of random variables, accounting for conditional dependence.

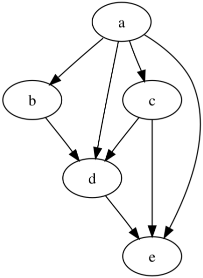
To do: Look up and learn the mathematical definition of graph.

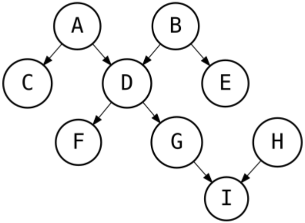
4.2 Directed Graphical Models

- a.k.a. **belief networks** or **belief nets**.
- a.k.a. **Bayesian networks** or **Bayes nets**.

  (but there's nothing Bayesian about Bayes nets)

- Model is based on a directed acyclic graph (DAG).





4.2.1 Representing the joint distribution

- Nodes are ordered in topological ordering: Parents before children.
- In a DAG we define the ordered Markov property: each node is conditionally independent of all its predecessors (takeaway parent nodes) in the ordering given its parents:

$$x_i \perp\!\!\!\perp \mathbf{x}_{\mathrm{pred}(i)\setminus\mathrm{pa}(i)} \mid \mathbf{x}_{\mathrm{pa}(i)}$$

  where

  - $\mathrm{pa}(i)$ is the set of parent nodes of i,
  - $\mathrm{pred}(i)$ is the set of all predecessors of i.

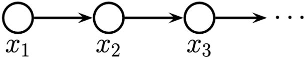
- The corresponding joint distribution $p(\mathbf{x}_{1:N}) = p(x_1, x_2, \dots, x_N)$ is then as follows:

$$p(\mathbf{x}_{1:N}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\mathrm{pa}(i)})$$

- $p(x_i|\mathbf{x}_{\mathrm{pa}(i)})$ is the conditional probability distribution (CPD) for node i.

Example: Markov chains

- A first-order Markov chain:

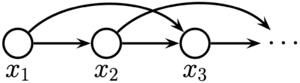


  - Joint distribution for $x_1, x_2, \ldots, x_T$ is:

    $p(\mathbf{x}_{1:T}) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1})$.

  - **To do:** Read about conditional probability table (CPT) for discrete random variables, also called transition probability matrix.

- A second-order Markov chain:



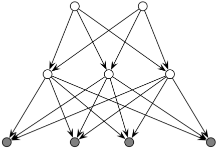
  - Joint distribution for $x_1, x_2, \ldots, x_T$ is:

    $p(\mathbf{x}_{1:T}) = p(x_1, x_2) \prod_{t=3}^{T} p(x_t | x_{t-2}, x_{t-1})$.

Example: The "student" network

- **To do:** Read this example (Murphy (2023), pp. 145 - 146).
  - Make sure to understand the **explaining away** effect.
  - This is also called **Berkson's paradox**.

Example: Sigmoid belief nets

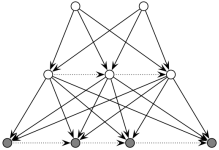


- Hierarchical latent variable model with 2 layers:
  - Joint distribution for $\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2$ is:

$$p(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2) = p(\mathbf{z}_2) p(\mathbf{z}_1 | \mathbf{z}_2) p(\mathbf{x} | \mathbf{z}_1)$$

$$= \prod_{k_2=1}^{K_2} p(z_{2,k_2}) \prod_{k_1=1}^{K_1} p(z_{1,k_1} | \mathbf{z}_2) \prod_{i=1}^{D} p(x_i | \mathbf{z}_1)$$

  - When the latent variables $(\mathbf{z}_1, \mathbf{z}_2)$ are binary, and the latent CPDs are logistic regression models, this is called a **sigmoid belief net**.



- With autoregressive connections within each layer:
  - **To do:** Joint distribution for $\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2$.

4.2.3 Gaussian Bayes nets

- **To do:** Read about this (Murphy (2023), p. 148).

4.2.4 Conditional Independence (CI) properties

- Notation: $\mathbf{x}_A \perp\!\!\!\perp \mathbf{x}_B \mid \mathbf{x}_C$
  - $\mathbf{x}_A$ is conditionally independent of $\mathbf{x}_B$ given $\mathbf{x}_C$. (In the given graph.)
  - Always relative to a given graph.
  - Sometimes the notation $\perp\!\!\!\perp_G$ is used, where G is the given graph.
  - Safe to write $\perp\!\!\!\perp$ if G is obvious from the context.
- $I(G)$ the set of all CI statements encoded by the graph G.
- $I(p)$ the set of all CI statements that hold true in some distribution p.
- G is an **I-map** (independence map) for $p \quad \Leftrightarrow \quad I(G) \subset I(p)$

  (Equivalent to say: p is Markov w.r.t. G.)
- **Minimal I-map:** There is no sub-graph G' that is an I-map of p.

4.2.4.1 Global Markov properties

- We say an undirected path P is **d-separated** by a set of nodes C iff at least one of the following conditions hold:
    - P contains a chain or **pipe**, $s \to m \to t$ or $s \leftarrow m \leftarrow t$, where $m \in C$.
    - P contains a tent or **fork** $_{s} \swarrow {}^{m} \searrow {}_{t}$ where $m \in C$.
    - P contains a **collider** (also called **v-structure**) ${}^{s} \searrow {}_{m} \swarrow {}^{t}$ where m is not in C and neither is any descendant of m.
- Next, we say that a set of nodes A is d-separated from a different set of nodes B given a third observed set C iff each undirected path from every node $a \in A$ to every node $b \in B$ is d-separated by C.
- Finally, we define the CI properties of a DAG as follows:

  $$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C \quad \Leftrightarrow \quad A \text{ is d-separated from } B \text{ given } C.$$

  This is called the (directed) **global Markov property**.

4.2.4.1 Global Markov properties (cont'd)

- **To do:** Read **Bayes ball algorithm** (Murphy (2023), pp. 149-150).

4.2.4.1 Global Markov properties (cont'd)

- Chain structure $X \to Y \to Z$
  - $p(x, z|y) = p(x|y)p(z|y)$ so then $X \perp\!\!\!\perp Z \mid Y$
  - "observing a node separates its predecessor and successor"
- Tent structure $X \leftarrow Y \to Z$
  - $p(x, z|y) = p(x|y)p(z|y)$ so then $X \perp\!\!\!\perp Z \mid Y$
  - "observing a root node separates its children nodes"
- Collider structure $X \to Y \leftarrow Z$
  - $p(x, z|y) \neq p(x|y)p(z|y)$ so then $X \not\!\perp\!\!\!\perp Z \mid Y$
  - "observing a common child node makes its parent nodes dependent"

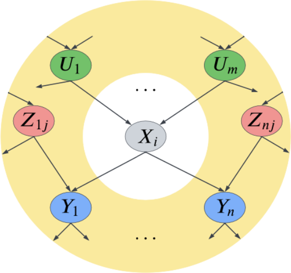
  - Notice that X and Z are (unconditionally) independent.
  - However, conditioning on Y makes X and Z become dependent.
  - **Berkson's paradox**, **explaining away**, or **inter-causal reasoning**.

4.2.4.3 Markov blankets

- $\mathrm{mb}(i)$ denotes the **Markov blanket** of node i.

- By definition, $\mathrm{mb}(i)$ is the smallest set of nodes that renders node i conditionally independent of all the other nodes in the graph.

- Fact: $\mathrm{mb}(i) = \mathrm{ch}(i) \cup \mathrm{pa}(i) \cup \mathrm{copa}(i)$



4.2.4.4 Other Markov properties

- The (directed) **local Markov property:** $i \perp\!\!\!\perp \mathrm{nd}(i) \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i)$
  - $\mathrm{nd}(i) =$ non-desdendants of node i.
- The **ordered Markov property:** $i \perp\!\!\!\perp \mathrm{pred}(i) \setminus \mathrm{pa}(i) \mid \mathrm{pa}(i)$
  - $\mathrm{pred}(i) =$ predecessors of node i.
- The **factorization property:** Any distribution p that is Markov w.r.t. the graph can be factored as

$$p(\mathbf{x}_{1:N}) = \prod_{i=1}^{N} p(x_i | \mathbf{x}_{\mathrm{pa}(i)})$$

4.2.5 Generation (sampling)

- To generate samples from a probabilistic model given by a DAG, follow the topological order:
  - generate parents before children
  - then generate a value for each node given the value of its parents.
- This is called **ancestral sampling**.
- This generates i.i.d. samples of $\mathbf{x}_{1:N} = (x_1, \ldots, x_N)$
  - i.i.d. = independent, identically distributed
- According to the distribution p defined by the graph.

4.2.6 Inference

- Graph G (a DAG) with N nodes.

- In the context of PGMs, the term “inference” refers to the task of computing a posterior distribution over a set of query nodes Q given the observed values for a set of visible nodes V, while marginalizing over the irrelevant nuisance variables, $R = \{1, \ldots, N\} \setminus \{Q, V\}$:

$$p(Q|V) = \frac{p(Q, V)}{p(V)} = \frac{\sum_R p(Q, V, R)}{p(V)}$$

- In the parametric case: $p = p_{\boldsymbol{\theta}}$ for some parameter vector $\boldsymbol{\theta}$.

- Single node case: $p(i|V)$ is called the posterior marginal for node i

4.2.7 Learning

- Possible to learn the graph structure from data.
- Possible to learn the associated distribution from data.

- **Parameter learning:** $p(\mathbf{x}_{1:N}) = p_{\boldsymbol{\theta}}(\mathbf{x}_{1:N}) =: p(\mathbf{x}_{1:N}|\boldsymbol{\theta})$

- Postulate prior $p(\boldsymbol{\theta})$
- Observe data $\mathcal{D}$
- Compute posterior $p(\boldsymbol{\theta}|\mathcal{D})$

- Can be done by treating $\boldsymbol{\theta}$ as a hiddel variable.

- **To do:** Read 4.2.7.k for k = 1,...,6. (Murphy (2023), pp. 157-161).

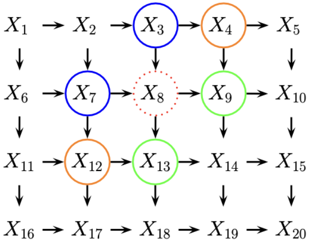
4.2.8 Stuff

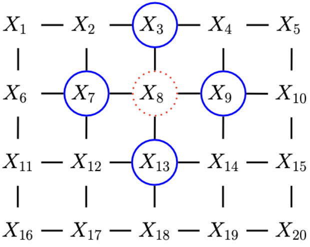
- **To do:** Read about **plate notation**, **factor analysis**, **naive Bayes classifier** and so on (Murphy (2023), pp. 162-164).

4.3 Undirected Graphical Models

- a.k.a. **Markov random fields**, or **Markov networks**.
- Do not require us to specify edge orientations, and are more natural for some problems such as image analysis and spatial statistics.
- Relax the restrictive arrow directions from DAGs.

- Model is based on an **undirected graph** (e.g. right side below).





4.3.1 Representing the joint distribution

- There's no topological ordering in an undirected graph.
- Associate a potential function with each maximal clique of the graph.
    - A **clique** is a set of nodes that are all neighbors of each other.
    - A **maximal clique** is a clique which cannot be made any larger without losing the clique property.
- $\psi_c(\mathbf{x}_c, \boldsymbol{\theta}_c)$ potential function associated with (maximal) clique c.
    - Potential function is any nonnegative function.
    - $\mathbf{x}_c$ are the variables in clique c.

4.3.1 Representing the joint distribution (cont'd)

- **Hammersley-Clifford theorem:** Suppose a joint distribution satisfies the CI properties implied by the undirected graph G. Then

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c, \boldsymbol{\theta}_c)$$

  - $\mathcal{C}$ is the set of all the (maximal) cliques of the graph G.
  - $Z(\boldsymbol{\theta})$ is the **partition function**: normalizing factor s.t. $\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\theta}) = 1$.

- **Gibbs distribution:**

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-\mathcal{E}(\mathbf{x};\boldsymbol{\theta}))$$

  - $\mathcal{E}(\mathbf{x};\boldsymbol{\theta})$ is the energy of state $\mathbf{x}$, given by $\mathcal{E}(\mathbf{x};\boldsymbol{\theta}) = \sum_{c \in \mathcal{C}} \mathcal{E}(\mathbf{x}_c;\boldsymbol{\theta}_c)$
  - This kind of probability model is also called an **energy-based model**.

Examples

- 4.3.2 Fully visible MRFs (Ising, Potts, Hopfield, etc.)
  - **To do:** Read about these (Murphy (2023), pp. 166-171).
- 4.3.3 MRFs with latent variables (Boltzmann machines, etc.)
  - **To do:** Read about these (Murphy (2023), pp. 172-174).
- 4.3.4 Maximum entropy models
  - **To do:** Read about these (Murphy (2023), pp. 175-176).
- 4.3.5 Gaussian MRFs
  - **To do:** Read about these (Murphy (2023), pp. 177-179).

4.3.6 Conditional Independence (CI) properties

- Notation: $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C$
  - $\mathbf{X}_A$ is conditionally independent of $\mathbf{X}_B$ given $\mathbf{X}_C$. (In the given graph.)
  - Always relative to a given graph.
  - Sometimes the notation $\perp\!\!\!\perp_G$ is used, where G is the given graph.
  - Safe to write $\perp\!\!\!\perp$ if G is obvious from the context.
- Undirected graphs define CI relationships via simple graph separation:

  $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_C \quad \Leftrightarrow \quad C$ separates A from B in the graph G.

  This is called the (undirected) **global Markov property**.
- "C separates A from B in the graph G" means that, when we remove all the nodes in C, if there are no paths connecting any node in A to any node in B, then the CI property holds.

Other Markov properties (undirected graph case)

- Let $G = (V, E)$ be the given (undirected) graph.
  - V is the set of vertices or nodes.
  - E is the set of edges joining pairs of nodes.

- **Markov blankets:** defined the same as for the directed graph case.

- The (undirected) **local Markov property:** $t \perp\!\!\!\perp V \setminus \mathrm{cl}(t) \mid \mathrm{mb}(t)$
  - $\mathrm{cl}(t) := \mathrm{mb}(t) \cup \{t\}$ is the **closure** of node t.
  - "a node's Markov blanket is its set of immediate neighbor"

- **Pairwise Markov property:** $s \perp\!\!\!\perp t \mid V \setminus \{s, t\} \quad \Leftrightarrow \quad G_{s,t} = 0.$
  - $G_{s,t} = 0$ means there is no edge between s and t.
  - This property says that two nodes are conditionally independent given the rest iff there is no direct edge between them.

Converting a DPGM to UPGM

- Determinining CI relationships in UPGMs is easier than in DPGMs:
  - don't have to worry about the directionality of the edges.
  - can use simple graph separation, instead of d-separation.

- It is possible to convert a DPGM to a UPGM, so that we can infer CI relationships for the DPGM using simple graph separation.

- Basically, need three cases:
  - **Chain** $X \rightarrow Y \rightarrow Z$. Simply drop the orientation (the arrow heads).
  - **Tent** $X \leftarrow Y \rightarrow Z$. Simply drop the orientation (the arrow heads).
  - **Collider** $X \rightarrow Y \leftarrow Z$. Add an edge between the co-parents X and Z, and drop the orientation (the arrow heads).

- This process is called **moralization**.
- To determine if $A \perp\!\!\!\perp B \mid C$ holds, first we form the **ancestral graph** of the DAG with respect to $U = A \cup B \cup C$. We then moralize this ancestral graph, and apply the graph separation rules for UPGMs.

Example

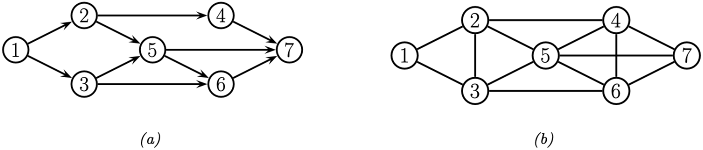


Figure 4.23: (a) A DPGM. (b) Its moralized version, represented as a UPGM.

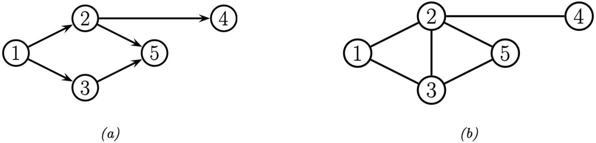


Figure 4.25: (a) The ancestral graph induced by the DAG in Figure 4.23(a) wrt $U = \{X_2, X_4, X_5\}$. (b) The moralized version of (a).

- e.g. can conclude from the latter: $X_4 \perp\!\!\!\perp X_5 \mid X_2$

Generation (sampling), Inference, Learning

- **4.3.7 Generation (sampling)**
  - Unlike with DPGMs, it can be quite slow to sample from UPGMs.
  - It is common to use MCMC methods for generating from UPGMs.
- **4.3.8 Inference**
  - Again, this is about computing the posterior distribution over nodes.
  - More details in Part 3 (Chapter 9).
- **4.3.9 Learning**
  - Mainly about estimating the parameters for MRFs.
  - Computing the MLE can be computationally expensive.
  - Computing the posterior over the parameters, $p(\boldsymbol{\theta}|\mathcal{D})$, is even harder, because of the additional normalizing constant $p(\mathcal{D})$.
  - A **doubly intractable** case.
  - Point estimation methods such as MLE and MAP can be tractable.

More goodies on PGMs

- **4.4 Conditional random fields (CRFs)**
  - 4.4.1 1d CRFs
  - 4.4.2 2d CRFs
- **4.5 Comparing directed and undirected PGMs**
  - 4.5.1 CI properties
  - 4.5.2 Converting between a directed and undirected
  - 4.5.3 Conditional directed vs undirected PGMs, label bias problem
  - 4.5.4 Combining directed and undirected graphs
  - 4.5.5 Comparing directed and undirected Gaussian PGMs
- **4.6 PGM extensions**
  - 4.6.1 Factor graphs
  - 4.6.1.1 Bipartite factor graphs
  - 4.6.1.2 Forney factor graphs

4.7 Structural causal models

- Based on DAGs.
- A **structural causal model** (SCM) is a triple $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{F})$ where
  - $\mathcal{U} = \{U_i : i = 1, \dots, N\}$ is a set of **exogenous random variables**.
  - $\mathcal{V} = \{V_i : i = 1, \dots, N\}$ is a set of **endogenous random variables**.
  - $\mathcal{F} = \{f_i : i = 1, \dots, N\}$ is a set of functions s.t. $V_i = f_i(V_{\mathbf{pa}(i)}, U_i)$.
- Assumes the equations can be structured in a recursive way, dependency of nodes given the DAG.
- Assumes the model is causally sufficient, which means that $\mathcal{V}$ and $\mathcal{U}$ are all of the causally relevant factors.
- This is called the **causal Markov assumption**.

4.7.1 Example: causal impact of education on wealth

- X = the level of education of a person, on some numeric scale (say 0 = high school, 1 = college, 2 = graduate school).
- Y = person's wealth (at some moment in time).
- Z = person's debt incurred, based on their education.

- The SCM has the form (see the figure):
  - $X = f_x(U_x)$
  - $Z = f_z(X, U_z)$
  - $Y = f_y(X, Z, U_y)$

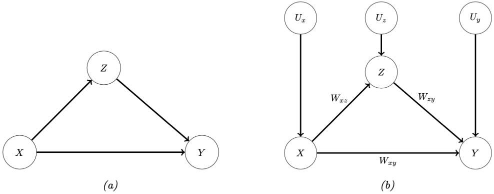


Figure 4.53: (a) PGM for modeling relationship between salary, education and debt. (b) Corresponding SCM.

4.7.2 Structural equation models

- A **structural equation model** (SEM) is a special case of a SCM, in which all the functional relationships are linear, and the prior on the noise (exogenous variables) is Gaussian.
- SEM version of previous example:
  - $X = U_x$
  - $Z = c_z + w_{x,z}X + U_z$
  - $Y = c_y + w_{x,y}X + w_{z,y}Z + U_y$
- If $p(u_x) = \mathcal{N}(u_x|0, \sigma_x^2)$, $p(u_y) = \mathcal{N}(u_y|0, \sigma_y^2)$, $p(u_z) = \mathcal{N}(u_z|0, \sigma_z^2)$; then the model can be converted to the following Gaussian DGM:
  - $p(x) = \mathcal{N}(x|0, \sigma_x^2)$
  - $p(z) = \mathcal{N}(z|c_z + w_{x,z}x, \sigma_z^2)$
  - $p(y) = \mathcal{N}(y|c_y + w_{x,y}x + w_{z,y}z, \sigma_y^2)$

4.7.3 Interventions, Do operator and augmented DAGs

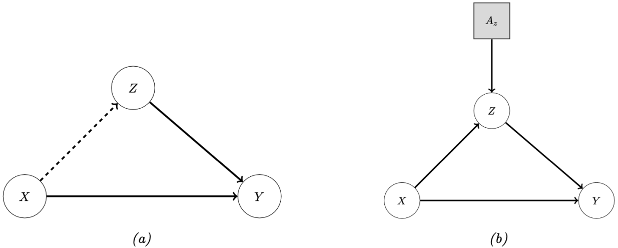


Figure 4.54: An SCM in which we intervene on Z. (a) Hard intervention, in which we clamp Z and thus cut its incoming edges (shown as dotted). (b) Soft intervention, in which we change Z's mechanism. The square node is an "action" node, using the influence diagram notation from Section 34.2.

4.7.4 Counterfactuals

- **Observational studies:** predicting effects of causes.
- **Interventional studies:** predicting causes of effects.

- Example:
  - I took the aspirin and my headache did go away.
  - Did taking the aspirin cause my headache to go away?
  - **Counterfactual question:** "if I had not taken the aspirin, would my headache have gone away anyway?"

- In counterfactual reasoning, the aim is to estimate $p(Y^{a'}|\mathrm{do}(a), y)$, to answer the question: "what is the probability distribution over outcomes Y if I were to do a', given that I have already done a and observed outcome y ?"
- Counterfactual reasoning requires strictly more assumptions than reasoning about interventions.

References

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL http://probml.github.io/book2.