National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

# Lecture 8

# Generative Models for Discrete Data. Directed Graphical Models

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

# Content

- Problems statement;
- Chain rule;
- Conditional independence;
- Graphical models and terminology;
- Directed graphical models;
- Markov and hidden Markov models;
- Inference in GM;
- CI properties of DGMs.

# Problems statement

- How can we *compactly represent* the joint distribution $p(\mathbf{x}|\boldsymbol{\theta})$?
- How can we use the distribution *to infer* one set of variables given another in a reasonable amount of computation time?
- How can we *learn* the parameters of this distribution with a reasonable amount of data?

# Chain rule

By the chain rule of probability, we can always represent a joint distribution as follows, using any ordering of the variables:

$$\boxed{p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)\cdots p(x_V|x_{1:V-1})}$$

where $V$ – is the number of variables, the Matlab-like notation $1{:}V$ denotes the set $\{1,2,3\cdots V\}$, for brevity we have dropped the conditioning on the fixed parameters $\boldsymbol{\theta}$.

For example, suppose all variables have $K$ states. We can represent $p(x_V|x_{1:V-1})$ as a table of $O(K^V)$ numbers by writing $p(x_V = i_V|x_1 = i_1,\cdots x_{V-1} = i_{V-1}) = T_{i_1,i_2\cdots i_V}$; we say that $\mathbf{T}$ is a ***<u>stochastic matrix</u>*** or ***<u>conditional probability table</u>*** (CPT).

# Conditional independence

The key to efficiency representing large joint distributions is to make some assumptions about ***conditional independence*** (***CI***). Recall that $X$ and $Y$ are conditionally independent given $Z$, denoted $X \perp Y|Z$, if and only if the conditional joint can be written as a product of conditional marginals:

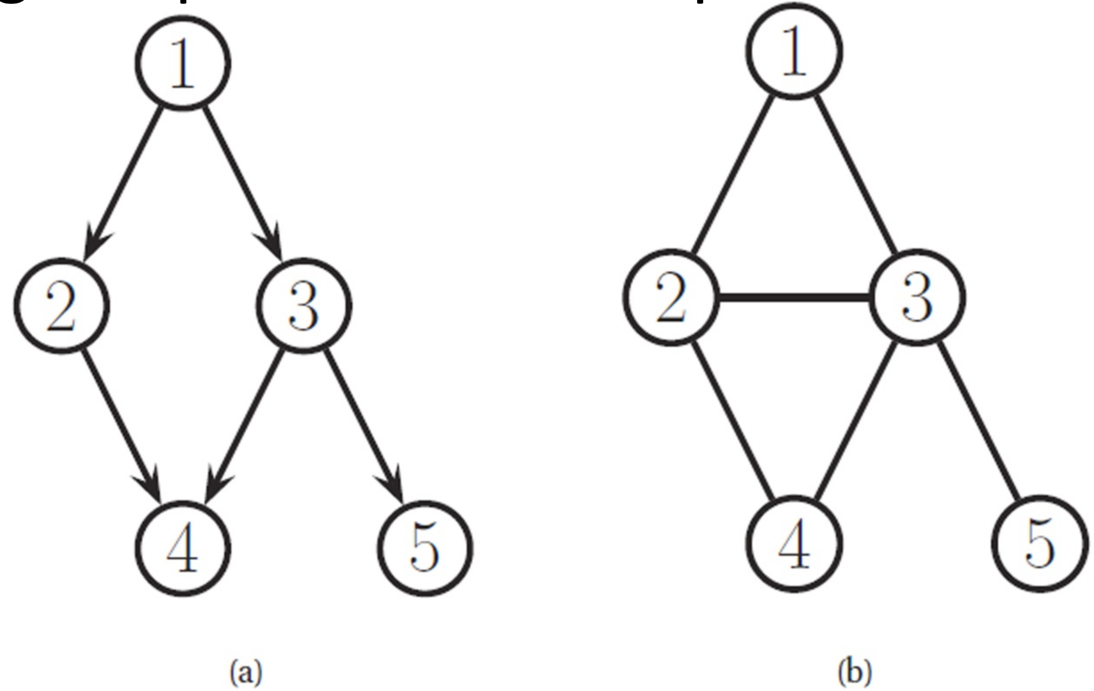
$$\boxed{X \perp Y|Z \Leftrightarrow p(X,Y|Z) = p(X|Z)p(Y|Z)}$$

Using assumption that $x_{t+1} \perp \mathbf{x}_{1:V-1}|x_t$ (***Markov first order assumption***) and chain rule we can write the joint distribution as

$$\boxed{p(\mathbf{x}_{1:V}) = p(x_1)\prod_{i=1}^{V} p(x_t|x_{t-1})}$$

# Graphical models

A ***graphical model*** (***GM***) is a way to represent a joint distribution by making CI assumption. In particular, the nodes in the graph represent random variables, and the (lack of) edges represent CI assumptions.



(a) A simple DAG on 5 nodes, numbered in topological order. Node 1 is the root, nodes 4 and 5 are the leaves. (b) A simple undirected graph, with the following maximal cliques: $\{1, 2, 3\}$, $\{2, 3, 4\}$, $\{3, 5\}$.

# Graph terminology (1/5)

A ***graph*** $G = (\mathcal{V}, \mathcal{E})$ consists of a set of ***nodes*** or ***vertices***, $\mathcal{V} = \{1{,}2, \cdots V\}$, and a set of ***edges***, $\mathcal{E} = \{(s,t){:}\, s,t \in \mathcal{V}\}$.

We can represent the graph by its ***adjacency matrix***, in which we write $G(s,t) = 1$ to denote $(s,t) \in \mathcal{E}$, that is, if $s \rightarrow t$ is an edge in the graph. If $G(s,t) = 1$ iff $G(t,s) = 1$, we say the graph is ***undirected***, otherwise it is ***directed***.

We usually assume $G(s,s) = 0$, which means there are no ***self loops***.

# Graph terminology (2/5)

- For a directed graph, the ***parent*** of a node is the set of all nodes that feed into it: $pa(s) \triangleq \{t: G(t,s) = 1\}$;
- For a directed graph, the ***children*** of a node is the set of all nodes that feed out of it: $ch(s) \triangleq \{t: G(s,t) = 1\}$;
- For a directed graph, the ***family*** of a node is the node and its parents, $fam(s) = \{s\} \cup pa(s)$;
- For a directed graph, a ***root*** is a node with no parents;
- For a directed graph, a ***leaf*** is a node with no children;
- For a directed graph, the ***ancestors*** are the parents, grand-parents, etc of a node. That is, the ancestor of $t$ is the set of nodes that connect to $t$ via a trail: $anc(t) \triangleq \{s: s \rightsquigarrow t\}$;
- For a directed graph, the ***descendants*** are the children, grand-children, etc of a node. That is, the descendants of $s$ is the set of nodes that can be reached via trails from $s$: $desc(s) \triangleq \{t: s \rightsquigarrow t\}$.

# Graph terminology (3/5)

- For any graph, we define the ***neighbors*** of a node as the set of all immediately connected nodes, $nbr(s) \triangleq \{t: G(s,t) = 1 \vee G(t,s) = 1\}$. For an undirected graph, we write $s \sim t$ to indicate that $s$ and $t$ are neighbors (so $(s,t) \in \mathcal{E}$);
- The ***degree*** of a node is the number of neighbors. For a directed graph, we speak of the ***in-degree*** and ***out-degree***, which count the number of parent and children;
- For any graph, we define a ***cycle*** or ***loop*** to be a series of nodes such that we can get back to where we started by following edges, $s_1 - s_2 - s_3 - \cdots - s_n - s_1, n \geq 2$. If the graph is directed, we may speak of a ***directed cycle***;

# Graph terminology (4/5)

- A ***directed acyclic graph*** (***DAG***) is a directed graph with no directed cycles;
- For a DAG, a ***topological ordering*** or ***total ordering*** is numbering of the nodes such that parents have lower numbers that their children;
- A ***path*** or ***trail*** $s \rightsquigarrow t$ is a series of directed edges leading from $s$ to $t$;
- An undirected ***tree*** is an undirected graph with no cycles. A directed tree is a DAG in which there are no directed cycles. IF we allow a node to have multiple parents, we call it a ***polytree***, otherwise we call it a ***moral directed tree***;
- A ***forest*** is a set of trees;

# Graph terminology (5/5)

- A (node-included) ***subgraph*** $G_A$ is a graph created by using the nodes in $A$ and theirs corresponding edges, $G_A = (\mathcal{V}_A, \mathcal{E}_A)$;
- For an undirected 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.

# Directed graphical models

A ***directed graphical model*** or ***DGM*** is a GM whose graph is a DAG.

These models are also called ***Bayesian*** or ***belief networks*** where term “belief” refers to subjective probability.

Given the topological ordering of DAGs, we define the ***ordered Markov property*** to be assumption that a node only depends on its immediate parents, not on all predecessors in the ordering:

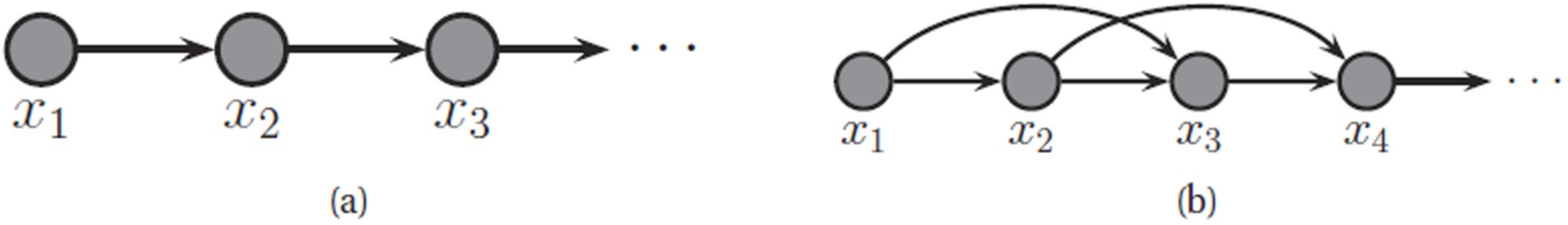
$$\boxed{x_s \perp \mathbf{x}_{pred(s)\setminus par(s)} | \mathbf{x}_{pa(s)}}$$

$$\boxed{p(\mathbf{x}_{1:V} | G) = \prod_{t=1}^{V} p\left(x_t | \mathbf{x}_{pa(t)}\right)}$$

# Markov and hidden Markov models (1/2)

The assumption that the immediate past, $x_{t-1}$, captures everything we need to know about the entire history, $\mathbf{x}_{1:t-2}$, is a bit strong. We can relax it a little by adding a dependence from $x_{t-2}$ to $x_t$ as well; this is called a second order Markov chain. The corresponding joint has the following form:

$$\boxed{p(\mathbf{x}_{1:T}) = p(x_1, x_2)p(x_3|x_1, x_2) \cdots p(x_t|x_{t-1}, x_{t-2}) = p(x_1, x_2) \prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})}$$

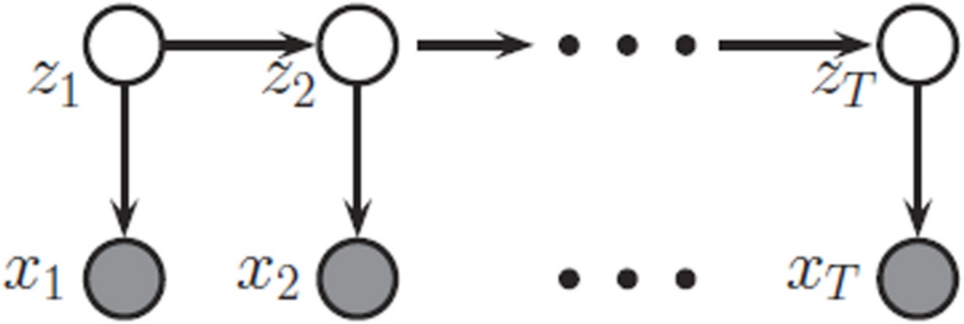


A first and second order Markov chain.

# Markov and hidden Markov models (2/2)

Unfortunately, even the second-order Markov assumption may be inadequate if there are long-range correlations amongst the observations. We can't keep building ever higher order models, since the number of parameters will blow up.

An alternative approach is to assume that there is an underlying hidden process, that can be modeled by a first-order Markov chain, but the data is a noisy observation of this process. The result is known as a ***hidden Markov model*** or ***HMM***.



A first-order HMM.

# Inference in GM

In general, we can pose the inference problem as follows. Suppose we have a set of correlated random variables with joint distribution $p(\mathbf{x}_{1:V}|\boldsymbol{\theta})$ (we assume that parameters $\boldsymbol{\theta}$ are known). Let us partition this vector into the ***visible parameters*** $\mathbf{x}_v$, which are observed, and the ***hidden variables*** $\mathbf{x}_h$, which are unobserved. Inference refers to computing the posterior distribution of the unknown given the known:

$$\boxed{p(\mathbf{x}_h|\mathbf{x}_v, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{p(\mathbf{x}_v|\boldsymbol{\theta})} = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{\sum_{\acute{\mathbf{x}}_h} p(\acute{\mathbf{x}}_h, \mathbf{x}_v|\boldsymbol{\theta})}}$$

In case only some hidden variables are of interest of use (query variables $\mathbf{x}_q$) from all hidden variables (nuisance variables $\mathbf{x}_u$), we can compute the desired marginal by marginalizing out the nuisance variables:

$$\boxed{p(\mathbf{x}_q|\mathbf{x}_v, \boldsymbol{\theta}) = \sum_{\mathbf{x}_u} p(\mathbf{x}_q, \mathbf{x}_u|\mathbf{x}_v, \boldsymbol{\theta})}$$

# CI properties of DGMs (1/2)

We say that $G$ is an ***I-map*** (independence map) for $p$, or that $p$ is ***Markov*** wrt $G$, iff $I(G) \subseteq I(p)$, where $I(p)$ is the set of all CI statements that hold for distribution $p$.

Note that fully connected graph is an I-map of all distributions, since it makes no CI assertions at all (since it is not missing any edges). We therefore say $G$ is a ***minimal I-map*** of $p$ if $G$ is an I-map of $p$, and if there is no $\acute{G} \subseteq G$ which is an I-map of $p$.

# CI properties of DGMs (2/2)

We say an undirected path $P$ is ***d-separated*** by a set of nodes $E$ (containing the evidence) iff at least one of the following conditions hold:

1. $P$ contains a chain, $s \rightarrow m \rightarrow t$ or $s \leftarrow m \leftarrow t$, where $m \in E$;
2. $P$ contains a tent or fork, $s \swarrow^{m} \searrow t$, where $m \in E$;
3. $P$ contains a ***collider*** or ***v-structure***, $s \searrow_{m} \swarrow t$, where $m$ is not in $E$ and nor 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 $E$ iff each undirected path from every node $a \in A$ to every node $b \in B$ is d-separated by $E$.

$$\boxed{\mathbf{x}_A \perp_G \mathbf{x}_B | \mathbf{x}_E \Leftrightarrow A\ is\ d - separated\ from\ B\ given\ E}$$

# Conclusion

- Chain rule and conditional independence were considered;
- Terminology for the graphical models was presented;
- Methods for inference in Directed graphical models as well as Markov and hidden Markov models was shown;
- CI properties of DGMs was considered.