# Reading for Lecture 4

***Release v10***

**Christopher Lee**

October 04, 2011

## Contents



## 1 Reading Assignment

**For Lecture 4 (Oct. 5, 2011), please read Jones & Pevzner sections 8.1 - 8.3, and the following material.**

## 2 Dependency Structure

### 2.1 Conditional Independence

As usual in conditional probability, we can extend the definition of independence by adding conditions. If

$$p(Y|X,Z) = p(Y|Z)$$

we say "$X$ and $Y$ are conditionally independent given $Z$". The intuitive meaning of this property is that $X$ contains no *extra* information about $Y$ beyond that provided by $Z$ (and the converse must also be true). Again we can put this into a symmetric form:

$$p(X,Y|Z) = p(X|Z)p(Y|Z)$$

This makes a strong statement: if any information actually does connect $X$ and $Y$, that information must be fully contained in $Z$.

# 3 Information Graph Representation

It is very useful to make a picture of these information connections, in the following form:

- An information graph represents a specific joint probability distribution of some set of random variables.
- We draw a *graph* structure: a drawing consisting of *nodes*, and *edges* connecting a pair of nodes.
- Each node represents a distinct random variable or group of variables.
- You should remind yourself that when we sample this joint probability distribution, for each random draw from the distribution we get a value for *each* random variable in the information graph. (This is just the definition of what we mean when we say "joint" probability distribution).
- We draw *directed edges* representing a particular choice of how to "traverse" the joint probability of this set of random variables. For example, if we chose to expand $p(X, Y)$ as $p(X)p(Y|X)$, then we would simply draw an edge $X \rightarrow Y$.
- Specifically, we draw edges to represent the specific conditional probability terms present in our equation for the joint probability. That is, for each term $p(X_i|X_j, X_k...X_n)$ we draw a directed edge from *each condition variable* to the subject variable. For example, for $p(Z|X, Y)$ we draw two incoming edges to $Z$: $X \rightarrow Z, Y \rightarrow Z$.
- Furthermore, we follow a *minimal conditions* principle: we reduce our conditional probability terms to the simplest form that is valid for our specific joint probability distribution. For example, if $Y, Z$ are conditionally independent given $X$, then the joint probability can be written $p(X, Y, Z) = p(X)p(Y|X)p(Z|X)$. So in this case the information graph would consist of just two edges, $X \rightarrow Y, X \rightarrow Z$. Note that we do *not* draw an edge from $Y \rightarrow Z$, because the connection between them is already fully captured by their connection to $X$.
- Of course, as always with the chain rule, we can choose to traverse the set of variables in any order we want. We generally choose the order that is most convenient for our goal (i.e. what we want to calculate).
- Clearly, one node may have multiple *incoming edges* (reflecting its conditioning), or multiple *outgoing* edges (representing its dependents).
- It is useful to distinguish one special case of multiple outgoing edges: if the target variables are both conditionally independent given the source variable, and all have the same probability distribution (i.e. "Independent and Identically Distributed"), then we draw a *single* outgoing edge from the source variable which then is forked to each of the target variables, the same symbol used to indicate the "one-to-many" relation in a database schema.

## 3.1 Information Graph "Dimensionality"

The information graph structure immediately tells us the "dimensionality" of the joint probability. For example, consider the first case in the example figure, with three variables and the maximum number of connecting edges $\binom{3}{2} = 3$. This simply represents the general chain rule for three variables. Since we see one variable (Z) with two incoming edges, we know that there is a probability term that connects *three* variables (X, Y, and Z). This tells us that the joint probability "table" is three-dimensional and cannot be reduced to a lower dimensionality without information loss. This simply reflects the general chain rule for three variables – it is of course three-dimensional. For example, if each variable had $N$ states, summing over the joint probability would take $O(N^3)$ time.

To determine this dimensionality from an information graph, we simply look for the node with the largest number of incoming edges.

## 3.2 Factoring of Summation

This is closely related to another important aspect of modeling an inference problem, namely whether it is possible to simplify a summation over the possible values of hidden variables, by factoring the summation into *separate*

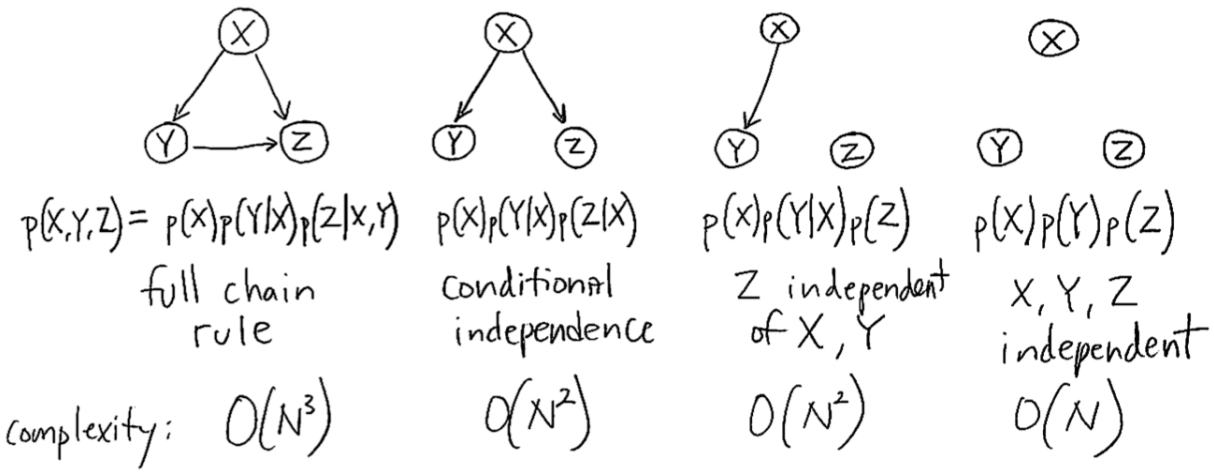


Figure 1: **The information graph, probability expression, standard description, and computational complexity for all possible levels of connectivity of three variables.**

summations.

- The key principle to remember is that for a summation over the values of a variable $X$, any multiplicative factor that does *not* depend on $X$ can be factored outside of the summation. And the information graph shows you at a glance exactly what depends on what (or more importantly, what does *not* depend on what).
- In the fully general chain-rule case, indicated by a "fully-connected information graph" structure, this is clearly not possible, because we have one variable that depends on *all* the other variables. This conditional probability term cannot be factored out of the summations of any of the variables, and "ties" them all together in an unfactorable lump.
- When some edges are missing, this implies at we can indeed factor the summation into separate factors representing the separate "branches" of the information graph. For example, for the three variable conditional-independence case, the summation is

$$\sum_{X,Y,Z} p(X,Y,Z) = \sum_{X,Y,Z} p(X)p(Y|X)p(Z|X) = \sum_{X} \left( p(X) \left( \sum_{Y} p(Y|X) \right) \left( \sum_{Y} p(Z|X) \right) \right)$$

  The three-dimensional summation is thus broken into two two-dimensional summations, far more tractable.

Note that there is a one-to-one correspondance between a specific information graph structure and particular summation factoring. The information graph gives you a very intuitive way of seeing exactly what pieces (if any) will factor.

## 3.3 Using Information Graphs for Inference

In Bayesian terms, inference means reversing a conditional probability relation, by applying Bayes Law. On the information graph, this translates to reversing the direction of arrows in our information graph: if the graph $\theta \rightarrow O$ represents the likelihood of an observation $O$ given a hidden state $\theta$, then inference simply reverses this: $\theta \leftarrow O$. Often our Bayes Law calculation also requires summing over all possible values of the hidden variable (i.e. projection, as described above). Let's consider several important cases:

- *independent observations*: say a hidden variable $\theta$ emits two observations $X, Y$ independently. We can use $X, Y$ to infer $\theta$, which simply corresponds to reversing the direction of both edges. The independent contributions of $X$ and $Y$ to $\theta$ in the information graph implies that they can be calculated separately, and this is in fact the case.

  Specifically, when observations make independent contributions to the likelihood model, we can divide the observations in any way we wish, compute the posterior using a given set of observations, and simply use the posterior as a *prior* for working with a separate set of observations. Thus, we can simply use the posterior computed using observation $X$ as a prior for computing a posterior using observation $Y$.

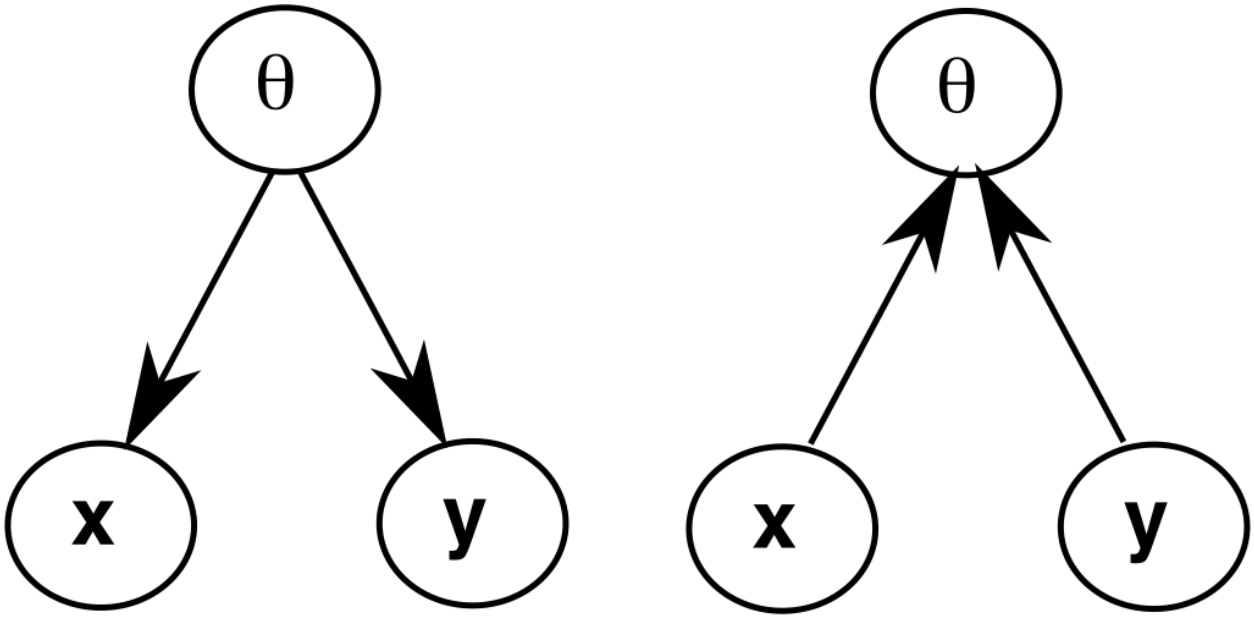


Figure 2: **Inference on $\theta$ given two conditionally independent observations** $X, Y$

- *independent hidden variables*: say two hidden variables $\theta, \lambda$ independently affect a set of observations $obs$, i.e. their contributions to the likelihood factor into separate terms. Note that this could take different forms:
  - The $obs$ could consist of two distinct subsets of observations $X_1, X_2, ...$ and $Y_1, Y_2, ...$ such that the probability of $x$ only depends on $\theta$ and the probability of $Y$ depends only on $\lambda$:

$$p(obs|\theta, \lambda) = p(X_1, X_2, ...|\theta)p(Y_1, Y_2, ...|\lambda)$$

  - Alternatively, the probability of the observations might factor into separate terms depending on $\theta$ and $\lambda$:

$$p(obs|\theta, \lambda) = f(obs, \theta)g(obs, \lambda)$$

  Inspection of the information graph for this case immediately implies that inference of $\theta$ and $\lambda$ can be done completely separately (from the same $obs$). And indeed the equations show that this is the case.

  Specifically, when we have two or more hidden variables $\theta, \lambda, ...$, and the prior and likelihood factor into separate terms for $\theta$ and $\lambda$, then these hidden variables will be conditionally independent given the observations. In other words, if $p(\theta, \lambda) = p(\theta)p(\lambda)$ and for observations $obs$, $p(obs|\theta, \lambda) = f(obs, \theta)g(obs, \lambda)$, then

$$p(\theta, \lambda|obs) = \frac{f(obs, \theta)g(obs, \lambda)p(\theta)p(\lambda)}{\int\int f(obs, \theta)g(obs, \lambda)p(\theta)p(\lambda)d\theta d\lambda}$$

$$= \frac{f(obs,\theta)p(\theta)}{\int f(obs,\theta)p(\theta)d\theta}\frac{g(obs,\lambda)p(\lambda)}{\int g(obs,\lambda)p(\lambda)d\lambda} = p(\theta|obs)p(\lambda|obs)$$

which can be solved as completely separate problems. This is not a minor detail. It means the computational complexity of solving for the hidden variables is reduced to that of solving each individual hidden variable separately. Otherwise this would become a coupled problem, in which we would have to compute a multi-dimensional integral (one dimension for each coupled hidden variable).

By contrast, if $p(obs, \theta, \lambda)$ cannot be factored into separate terms for $\theta$ and $\lambda$, these two variables become coupled, and all posterior calculations become two-dimensional problems (i.e. we must optimize $\theta$ and $\lambda$ simultaneously).

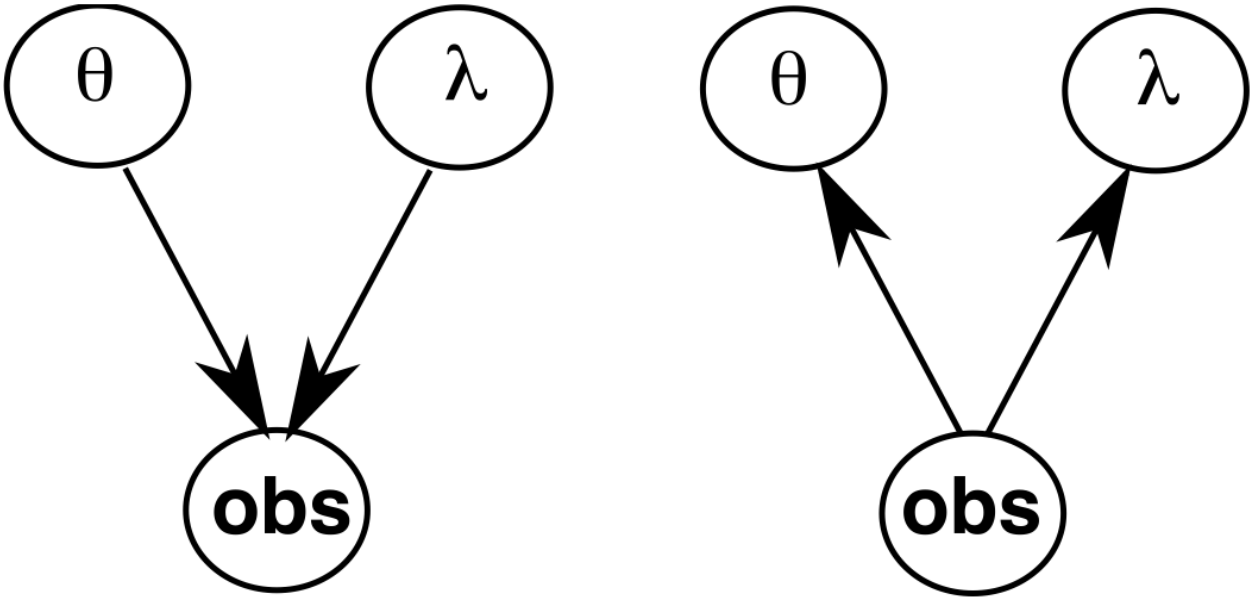


Figure 3: **Inference on $\theta, \lambda$ independently from the same set of observations** $obs$