3. Probabilistic Graphical Models

The Bayes Filter (Rep.)

$$\text{Bel}(x_t) = p(x_t \mid u_1, z_1, \dots, u_t, z_t)$$

(Bayes)
$$= \eta \, p(z_t \mid x_t, u_1, z_1, \dots, u_t) p(x_t \mid u_1, z_1, \dots, u_t)$$

(Markov)
$$= \eta \, p(z_t \mid x_t) p(x_t \mid u_1, z_1, \dots, u_t)$$

(Tot. prob.)
$$= \eta \, p(z_t \mid x_t) \int p(x_t \mid u_1, z_1, \dots, u_t, x_{t-1}) \, p(x_{t-1} \mid u_1, z_1, \dots, u_t) dx_{t-1}$$

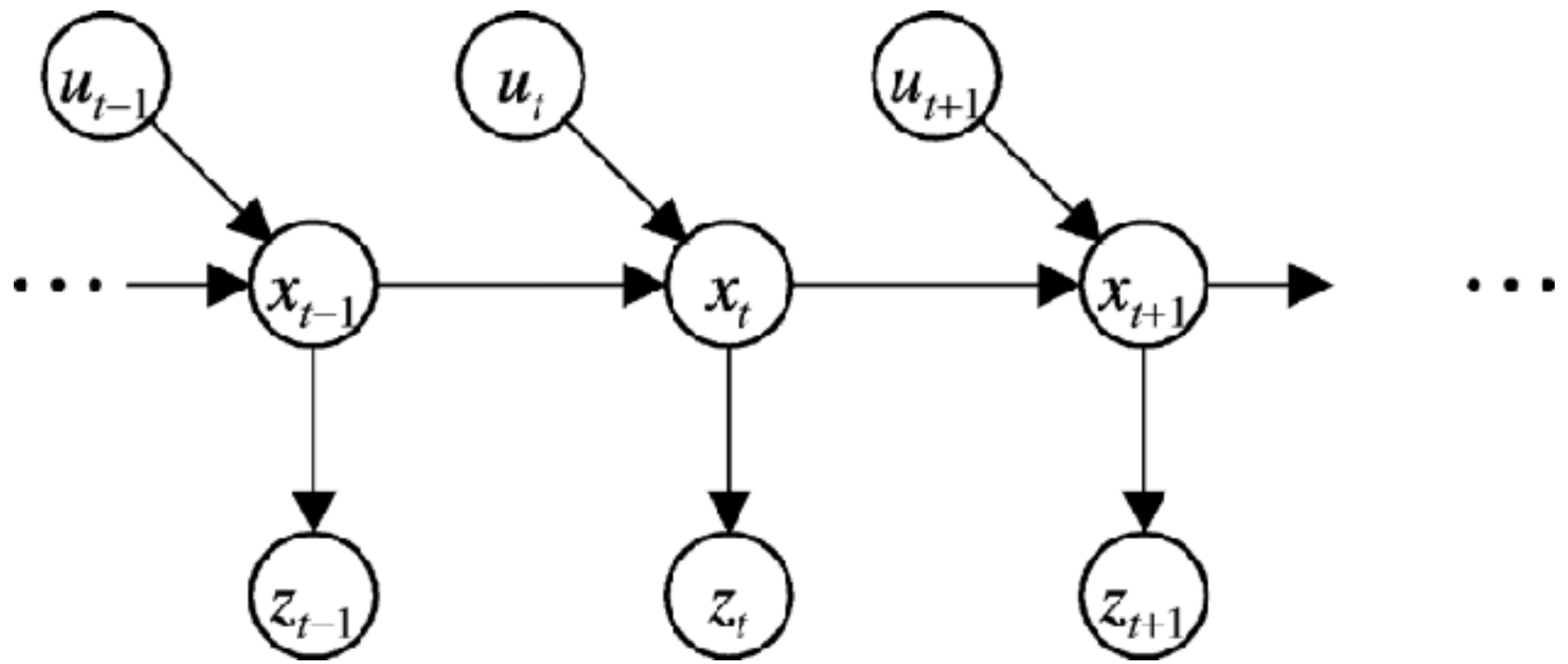
(Markov)
$$= \eta \, p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) p(x_{t-1} \mid u_1, z_1, \dots, u_t) dx_{t-1}$$

(Markov)
$$= \eta \, p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) p(x_{t-1} \mid u_1, z_1, \dots, z_{t-1}) dx_{t-1}$$

$$= \eta \, p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) \text{Bel}(x_{t-1}) dx_{t-1}$$

Graphical Representation (Rep.)

We can describe the overall process using a ***Dynamic Bayes Network***:



- This incorporates the following Markov assumptions:

$$p(z_t \mid x_{0:t}, u_{1:t}, z_{1:t}) = p(z_t \mid x_t) \quad \text{(measurement)}$$

$$p(x_t \mid x_{0:t-1}, u_{1:t}, z_{1:t}) = p(x_t \mid x_{t-1}, u_t) \quad \text{(state)}$$

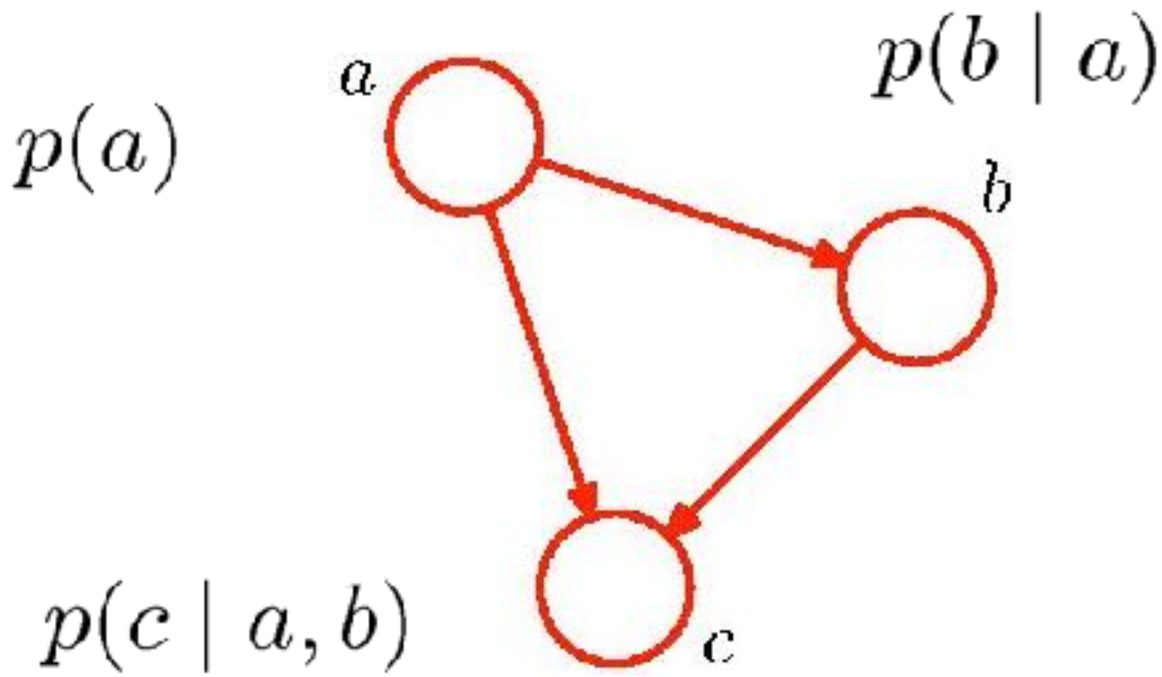
Definition

A Probabilistic Graphical Model is a diagrammatic representation of a probability distribution.

- In a Graphical Model, random variables are represented as **nodes**, and statistical dependencies are represented using **edges** between the nodes.
- The resulting graph can have the following properties:
- Cyclic / acyclic
- Directed / undirected
- The simplest graphs are Directed Acyclic Graphs (DAG).

Simple Example

- Given: 3 random variables a, b, and c
- Joint prob: $p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$



Random variables can be discrete or continuous

A Graphical Model based on a DAG is called a **Bayesian Network**

Simple Example

- In general: K random variables $x_1, x_2, \ldots, x_K$
- Joint prob:

$$p(x_1, \ldots, x_K) = p(x_K | x_1, \ldots, x_{K-1}) \ldots p(x_2 | x_1) p(x_1)$$

- This leads to a fully connected graph.
- Note: The ordering of the nodes in such a fully connected graph is **arbitrary**. They all represent the joint probability distribution:

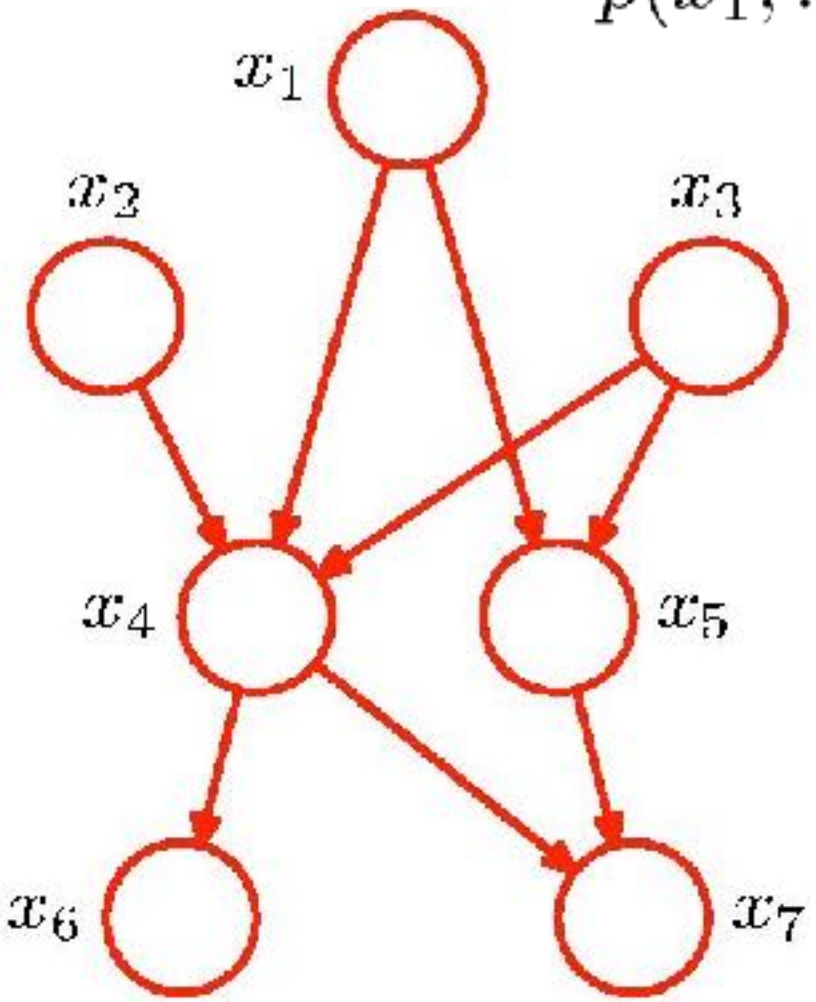
$$p(a, b, c) = p(a | b, c) p(b | c) p(c)$$

$$p(a, b, c) = p(b | a, c) p(a | c) p(c)$$

$$\vdots$$

Bayesian Networks

Statistical independence can be represented by the **absence** of edges. This makes the computation efficient.

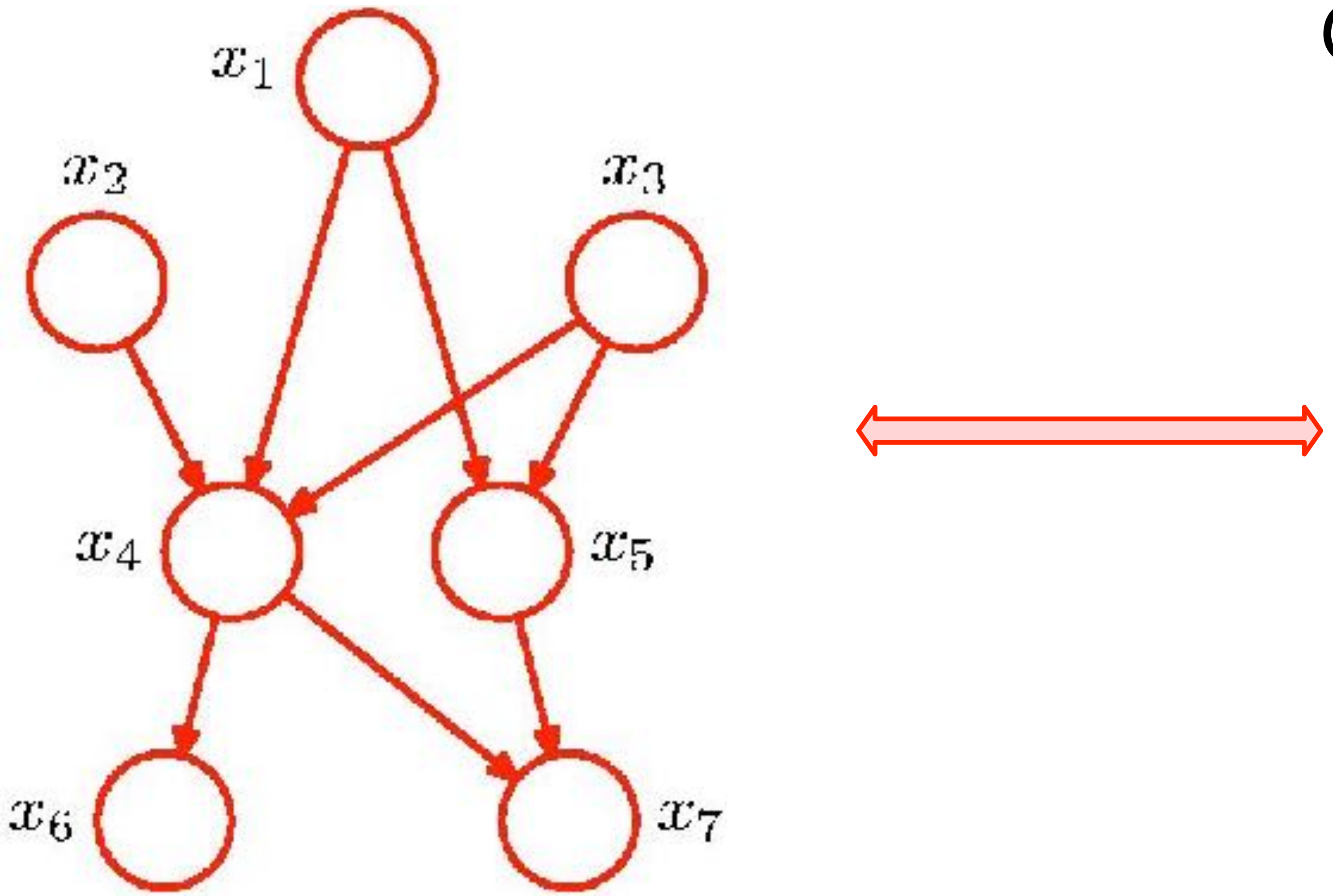


$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Intuitively: only x_1 and x_3 have an influence on x_5

Bayesian Networks

We can now define a mapping from graphical models to probabilistic formulations (factorisations) and back:



General Factorisation:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}_k)$$

where

$pa_k \triangleq$ ancestors of p_k

and

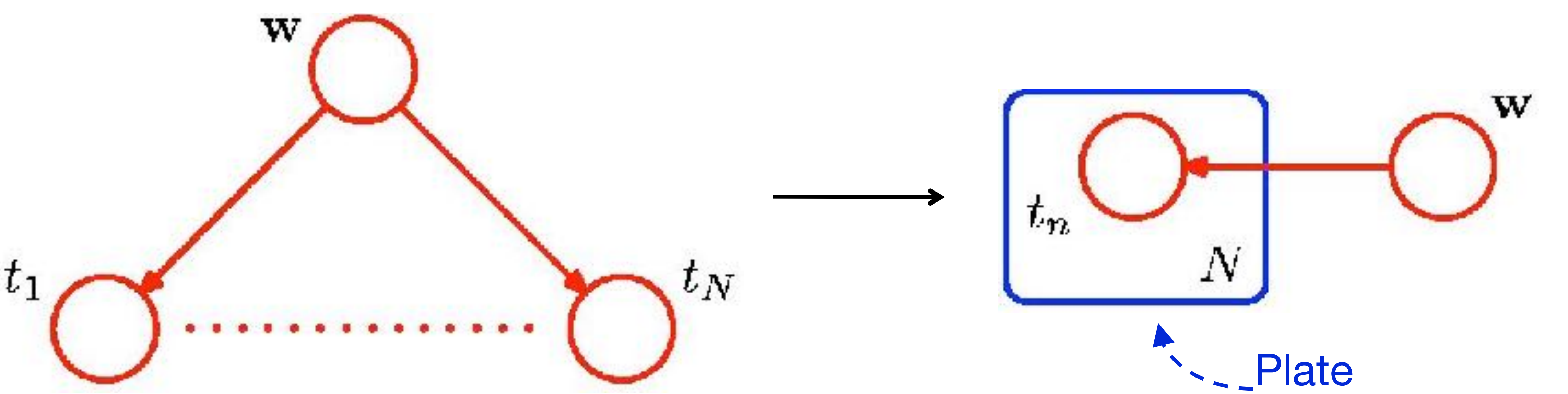
$$p(\mathbf{x}) = p(x_1, \ldots, x_K)$$

Note: Many different factorisations (and graphs) can represent the same distribution

Elements of Graphical Models

In case of a series of random variables with equal dependencies, we can subsume them using a **plate:**

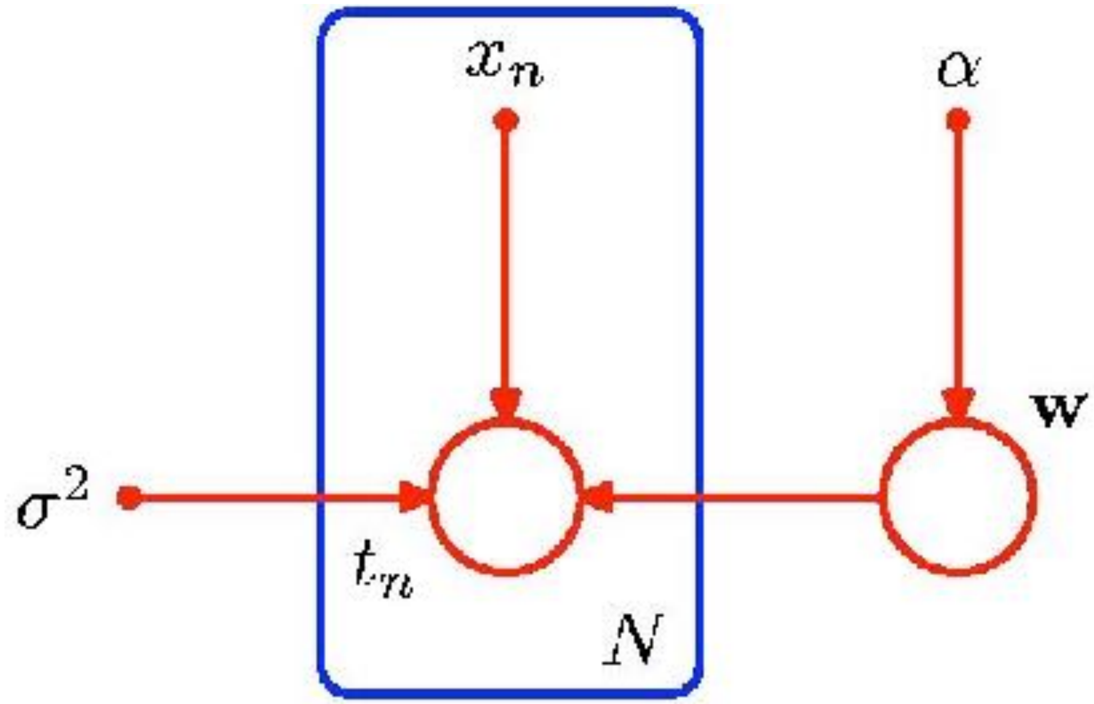
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | \mathbf{w})$$



Elements of Graphical Models (2)

We distinguish between **input** variables and explicit **hyper-parameters**:

$$p(\mathbf{t}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n | \mathbf{w}, x_n, \sigma^2).$$

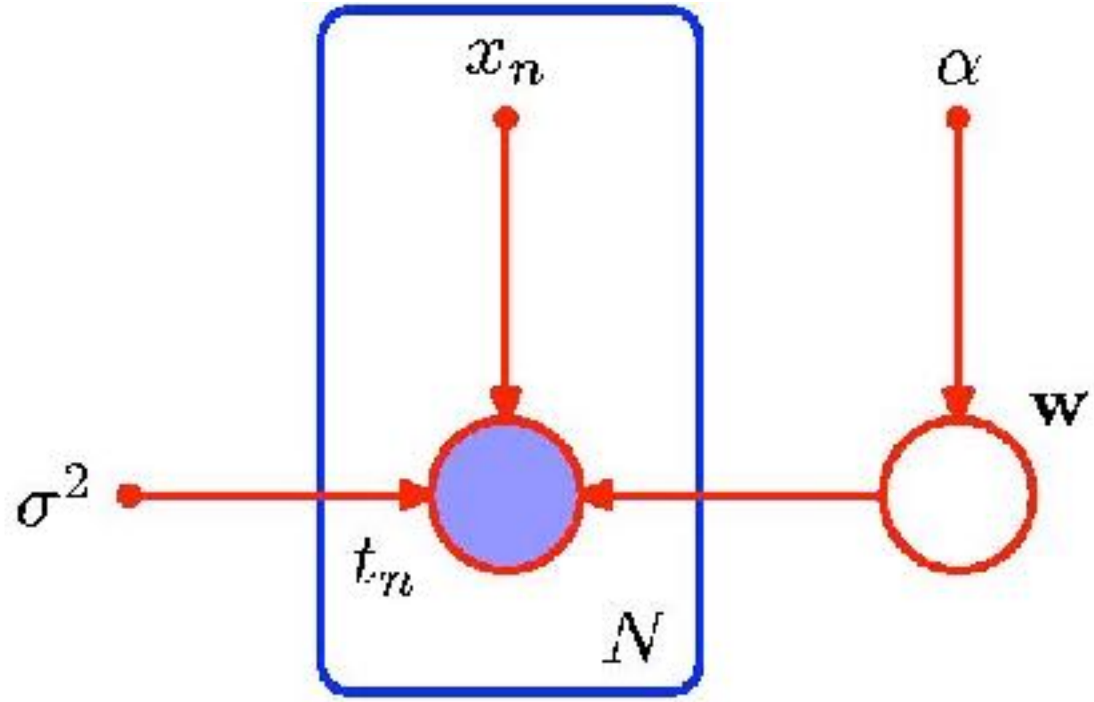


Elements of Graphical Models (3)

We distinguish between **observed** variables and **hidden** variables:

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

(deterministic parameters omitted in formula)



Example: Regression as a Graphical Model

Aim: Find a general expression to compute the **predictive distribution**: $p(\hat{t} \mid \hat{x}, \mathbf{x}, \mathbf{t})$

Notation:
$\hat{t} = t^*$

Bishop vs. Rasmussen

This expression should

- model all conditional independencies
- explicitly incorporate all parameters (also the deterministic ones)

Example: Regression as a Graphical Model

Aim: Find a general expression to compute the **predictive distribution**: $p(\hat{t} \mid \hat{x}, \mathbf{x}, \mathbf{t})$

Notation:

$\hat{t} = t^*$

Bishop vs. Rasmussen

This expression should

- model all conditional independencies
- explicitly incorporate all parameters (also the deterministic ones)

$$p(\hat{t} \mid \hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) = \int p(\hat{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) d\mathbf{w}$$

$$= \int \frac{p(\hat{t}, \mathbf{w}, \mathbf{t} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2)}{p(\mathbf{t} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2)} d\mathbf{w} \quad \propto \int p(\hat{t}, \mathbf{w}, \mathbf{t} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w}$$

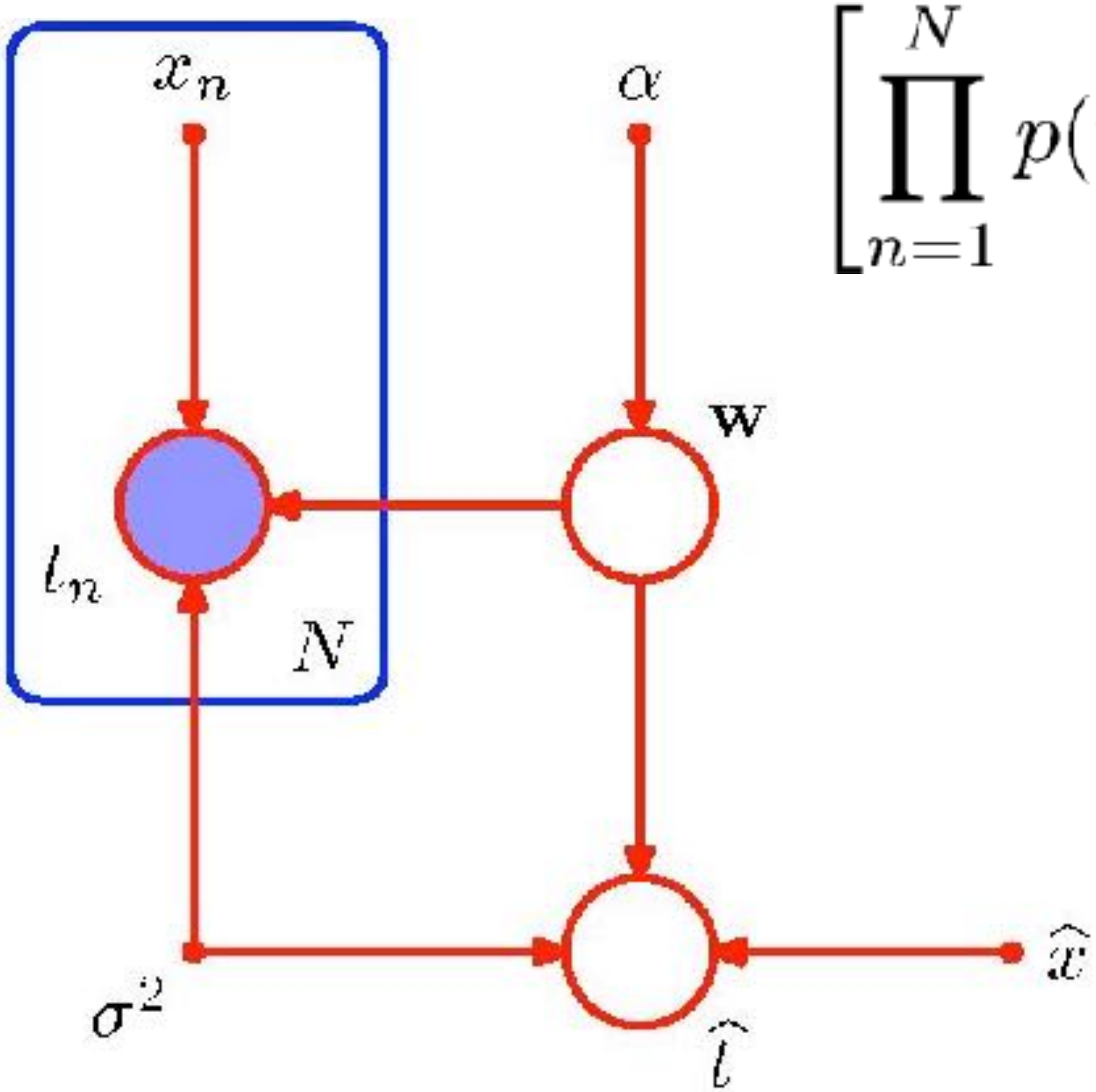
Regression as a Graphical Model

Regression: Prediction of a new target value $\hat{t}$

Notation:
$\hat{t} = t^*$

$$p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) =$$

$$\left[\prod_{n=1}^{N} p(t_n \mid x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w} \mid \alpha) p(\hat{t} | \hat{x}, \mathbf{w}, \sigma^2)$$



Here: conditioning on all deterministic parameters

Using this, we can obtain the **predictive distribution:**

$$p(\hat{t} | \hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\hat{t}, \mathbf{t}, \mathbf{w} | \hat{x}, \mathbf{x}, \alpha, \sigma^2) \, d\mathbf{w}$$

Example: Discrete Variables

- Two dependent variables: $K^2 - 1$ parameters

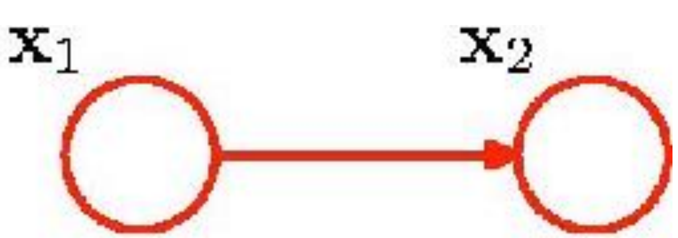
Here: $K = 2$

x_1	$p(x_1)$
1	0.2
~~2~~	~~0.8~~

$K - 1$

x_1	x_2	$p(x_2 \mid x_1)$
1	1	0.25
~~1~~	~~2~~	~~0.75~~
2	1	0.1
~~2~~	~~2~~	~~0.9~~

$K - 1$, $K - 1$, $K(K-1)$



$$K - 1 + K(K-1) = K^2 - 1$$

- Independent joint distribution: $2(K - 1)$ parameters

$$K - 1 + K - 1 = 2(K-1)$$

Discrete Variables: General Case

In a general joint distribution with M variables we need to store $K^M - 1$ parameters

If the distribution can be described by this graph:

then we have only $K - 1 + (M - 1)\,K(K - 1)$ parameters.

This graph is called a **Markov chain** with M nodes.

The number of parameters grows only **linearly** with the number of variables.

Independence (Rep.)

Definition 1.4: Two random variables X and Y are *independent* iff: $p(x, y) = p(x)p(y)$

For independent random variables X and Y we have:

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y)}{p(y)} = p(x)$$

Notation: $x \perp\!\!\!\perp y \mid \emptyset$

Independence does **not** imply conditional independence!

The same is true for the opposite case.

Conditional Independence (Rep.)

Definition 1.5: Two random variables X and Y are *conditional independent* given a third random variable Z iff:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

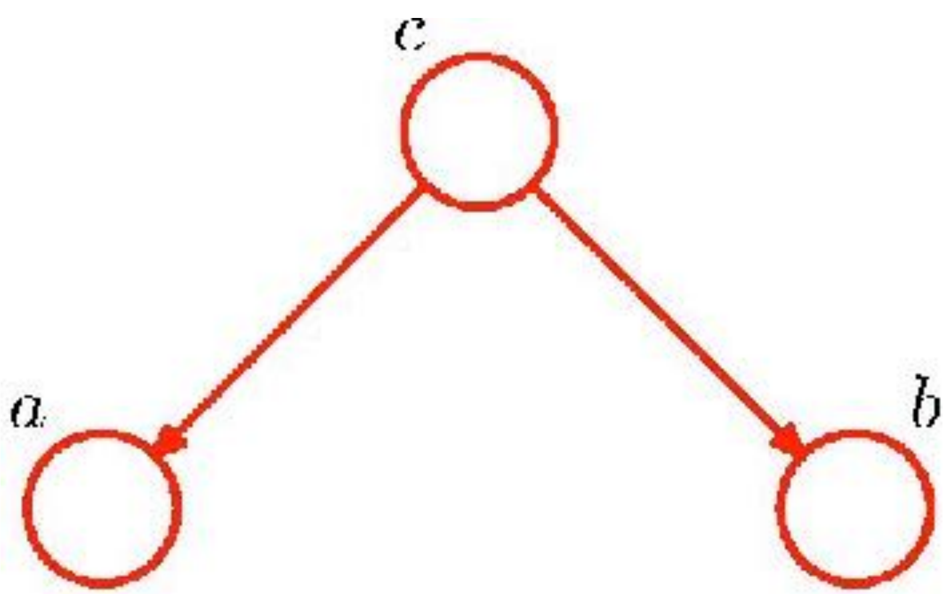
This is equivalent to:

$$p(x \mid z) = p(x \mid y, z) \quad \text{and}$$

$$p(y \mid z) = p(y \mid x, z)$$

Notation: $x \perp\!\!\!\perp y \mid z$

Conditional Independence: Example 1



This graph represents the probability distribution:

$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

Marginalizing out *c* on both sides gives

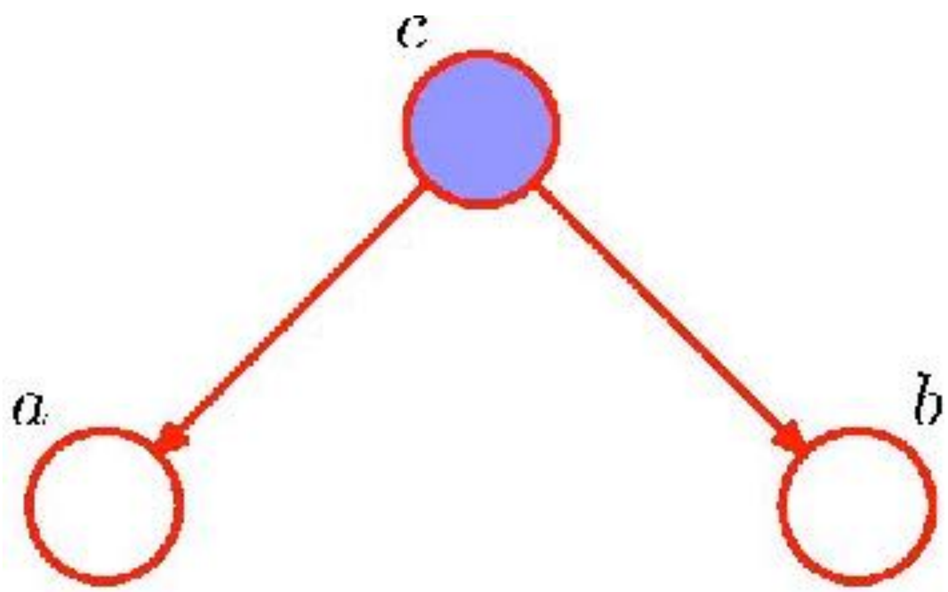
$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

This is in general not equal to $p(a)p(b)$.

Thus: a and b are not independent: $a \not\perp\!\!\!\perp b \mid \emptyset$

Conditional Independence: Example 1

Now, we condition on c (it is assumed to be known):

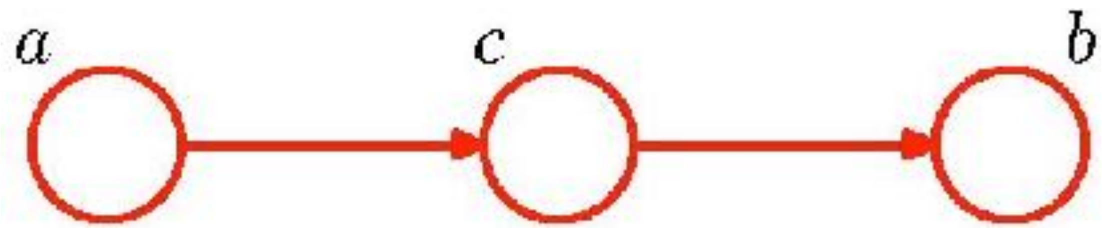


$$
\begin{aligned}
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

Thus: a and b are conditionally independent given c: $a \perp\!\!\!\perp b \mid c$

We say that the node at c is a **tail-to-tail node** on the path between a and b

Conditional Independence: Example 2



This graph represents the distribution:

$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

Again, we marginalize over c:

$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a) \sum_c p(c|a)p(b|c, a)$$

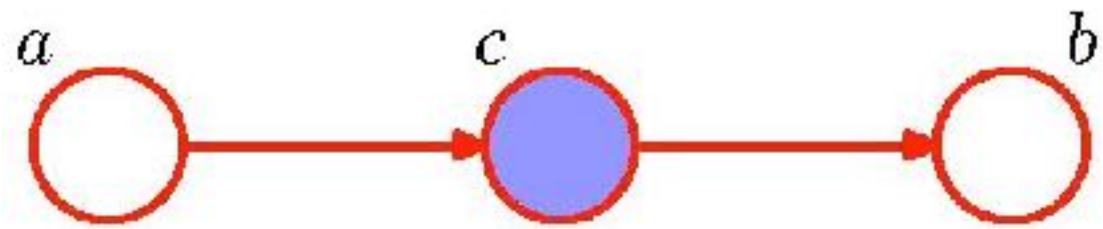
$$= p(a) \sum_c \frac{p(c, a)p(b, c, a)}{p(a)p(c, a)} = p(a) \sum_c p(b, c \mid a)$$

$$= p(a)p(b|a)$$

And we obtain: $a \not\perp\!\!\!\perp b \mid \emptyset$

Conditional Independence: Example 2

As before, now we condition on c :



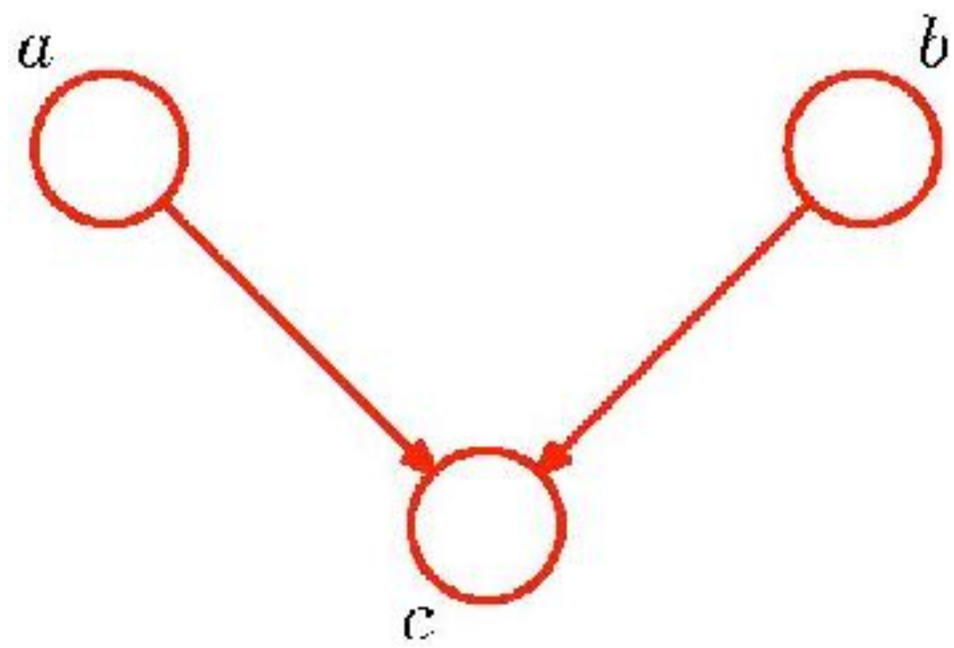
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

And we obtain: $a \perp\!\!\!\perp b \mid c$

We say that the node at c is a **head-to-tail node** on the path between a and b.

Conditional Independence: Example 3

Now consider this graph:



$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

using:

$$\sum_c p(a,b,c) = p(a)p(b)\sum_c p(c \mid a,b)$$

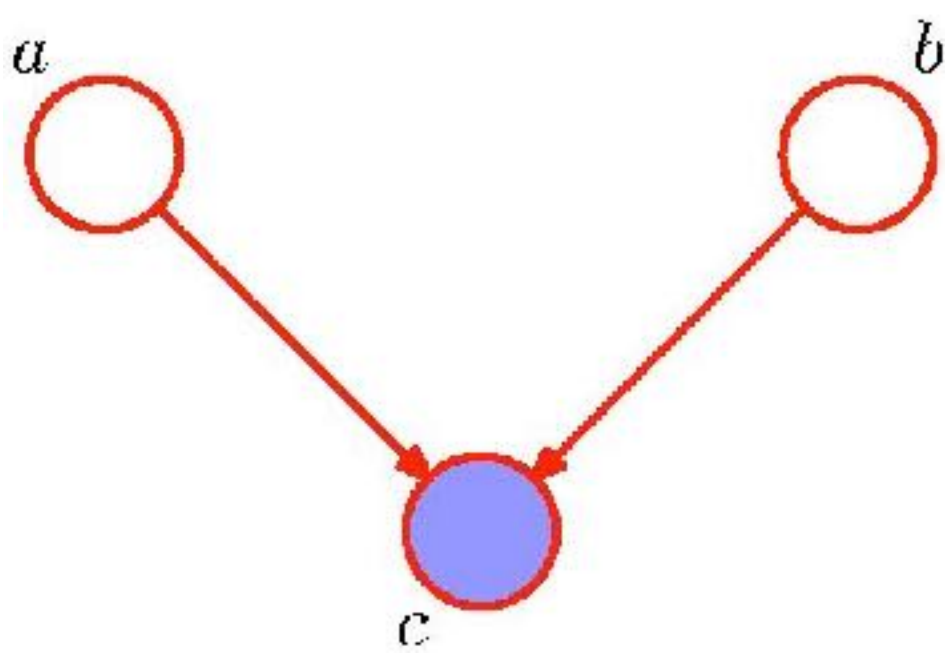
we obtain:

$$p(a,b) = p(a)p(b)$$

And the result is: $a \perp\!\!\!\perp b \mid \emptyset$

Conditional Independence: Example 3

Again, we condition on c



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

This results in: $a \not\perp b \mid c$

We say that the node at c is a **head-to-head node** on the path between a and b.

To Summarize

- When does the graph represent (conditional) independence?

  Tail-to-tail case: if we condition on the tail-to-tail node

  Head-to-tail case: if we cond. on the head-to-tail node

  Head-to-head case: if we do **not** condition on the head-to-head node (and neither on any of its descendants)

In general, this leads to the notion of **D-separation** for directed graphical models.