Summary: MAP Estimation

To summarize, we have the following optimization problem:

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \qquad \phi(\mathbf{x}_n) \in \mathbb{R}^M$$

The same in vector notation:

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \mathbf{w}\Phi^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\mathbf{t} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \quad \mathbf{t} \in \mathbb{R}^N$$

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_{M-1}(x_N) \end{pmatrix} \in \mathbb{R}^{N \times M}$$

“Feature Matrix”

Summary: MAP Estimation

To summarize, we have the following optimization problem:

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}_n) - t_n)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \qquad \phi(\mathbf{x}_n) \in \mathbb{R}^M$$

The same in vector notation:

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \mathbf{w}\Phi^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\mathbf{t} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \quad \mathbf{t} \in \mathbb{R}^N$$

And the solution is

$$\mathbf{w}^* = (\lambda I_M + \Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

Identity matrix of size M by M

MLE And MAP

- The benefit of MAP over MLE is that prediction is less sensitive to **overfitting,** i.e. even if there is only little data the model predicts well.
- This is achieved by using **prior information,** i.e. model assumptions that are not based on any observations (= data)
- But: both methods only give the **most likely** model, there is no notion of **uncertainty** yet

Idea 1: Find a **distribution** over model parameters ("parameter posterior")

MLE And MAP

- The benefit of MAP over MLE is that prediction is less sensitive to **overfitting,** i.e. even if there is only little data the model predicts well.
- This is achieved by using **prior information,** i.e. model assumptions that are not based on any observations (= data)
- But: both methods only give the **most likely** model, there is no notion of **uncertainty** yet

Idea 1: Find a **distribution** over model parameters

Idea 2: Use that distribution to estimate **prediction uncertainty** (“predictive distribution”)

When Bayes Meets Gauß

Theorem: If we are given this:

I. $$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \mu, \Sigma_1)$$

II. $$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid A\mathbf{x} + \mathbf{b}, \Sigma_2)$$

linear dependency on $\mathbf{x}$

Then it follows (properties of Gaussians):

III. $$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid A\mu + \mathbf{b}, \Sigma_2 + A\Sigma_1 A^T)$$

IV. $$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \Sigma(A^T\Sigma_2^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_1^{-1}\mu), \Sigma)$$

where

$$\Sigma = (\Sigma_1^{-1} + A^T\Sigma_2^{-1}A)^{-1}$$

See Bishop's book for the proof!

"Linear Gaussian Model"

When Bayes Meets Gauß

Thus: When using the Bayesian approach, we can do even more than MLE and MAP by using these formulae.

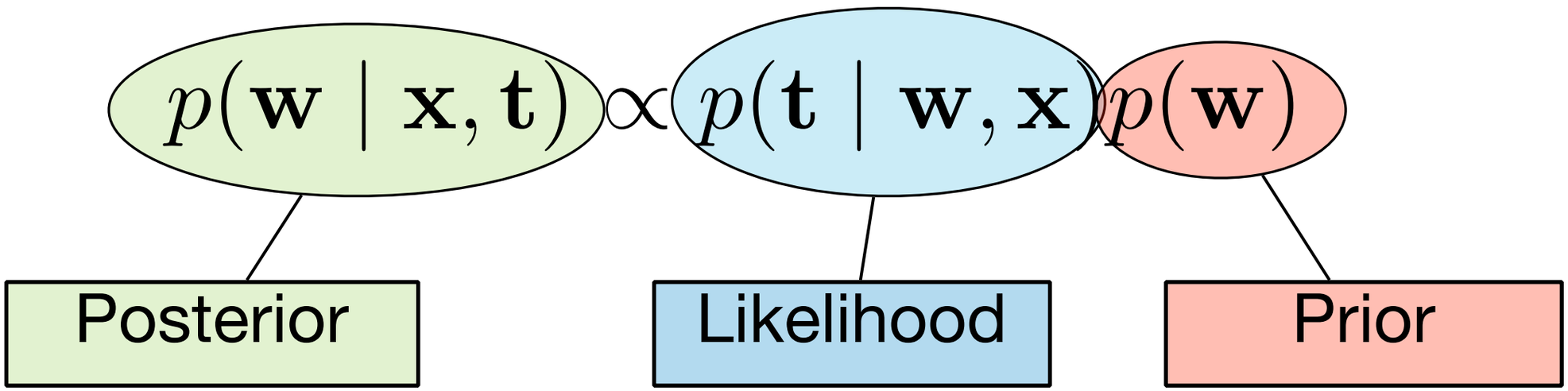
This means:

If the prior and the likelihood are Gaussian then the **posterior** and the **normalizer** are also Gaussian and we can compute them in closed form.

This gives us a natural way to compute uncertainty!

The Posterior Distribution

Remember Bayes Rule:



With our theorem, we can compute the posterior in **closed form** (and not just its maximum)!

The posterior is also a Gaussian and its **mean** is the MAP solution.

The Posterior Distribution

We have $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \sigma_2^2 I_M)$

and $p(\mathbf{t} \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(\mathbf{t}; \Phi\mathbf{w}, \sigma_1^2 I_N)$

From this and IV. we get the **posterior covariance**:

$$\Sigma = (\sigma_2^{-2} I_M + \sigma_1^{-2} \Phi^T \Phi)^{-1}$$

$$= \sigma_1^2 (\frac{\sigma_1^2}{\sigma_2^2} I_M + \Phi^T \Phi)^{-1}$$

and the **mean**: $\boldsymbol{\mu} = \sigma_1^{-2} \Sigma \Phi^T \mathbf{t}$

So the entire posterior distribution is

$$p(\mathbf{w} \mid \mathbf{t}, \mathbf{x}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \Sigma)$$

Note: So far we only used the **training** data!

$(\mathbf{x}, \mathbf{t})$

The Predictive Distribution

We obtain the **predictive distribution** by integrating over all possible model parameters ("inference"):

$$p(t^*|\, x^*, \mathbf{t}, \mathbf{x}) = \int p(t^*|\, x^*, \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

Test data

Test data likelihood

Parameter posterior

This distribution can be computed in closed form, because both terms on the RHS are Gaussian.

From above we have $p(\mathbf{w} \mid \mathbf{t}, \mathbf{x}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \Sigma)$

where $\boldsymbol{\mu} = \sigma_1^{-2} \Sigma \Phi^T \mathbf{t}$

and $\Sigma = \sigma_1^2 (\frac{\sigma_1^2}{\sigma_2^2} I_M + \Phi^T \Phi)^{-1}$

The Predictive Distribution

We obtain the **predictive distribution** by integrating over all possible model parameters ("inference"):

$$p(t^*|\, x^*, \mathbf{t}, \mathbf{x}) = \int p(t^*|\, x^*, \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

Test data

Test data likelihood

Parameter posterior

This distribution can be computed in closed form, because both terms on the RHS are Gaussian.

From above we have $p(\mathbf{w} \mid \mathbf{t}, \mathbf{x}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \Sigma)$

where $\boldsymbol{\mu} = \sigma_1^{-2} \Sigma \Phi^T \mathbf{t}$

and $\Sigma = \sigma_1^2 (\frac{\sigma_1^2}{\sigma_2^2} I_M + \Phi^T \Phi)^{-1}$ $\Rightarrow \boldsymbol{\mu} = (\lambda I_M + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$

MAP solution

The Predictive Distribution

Using formula III. from above (linear Gaussian),

$$p(t^*|\, x^*, \mathbf{t}, \mathbf{x}) = \int p(t^*|\, x^*, \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

$$= \int \mathcal{N}(t^*; \phi(x^*)^T \mathbf{w}, \sigma) \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \Sigma) d\mathbf{w}$$

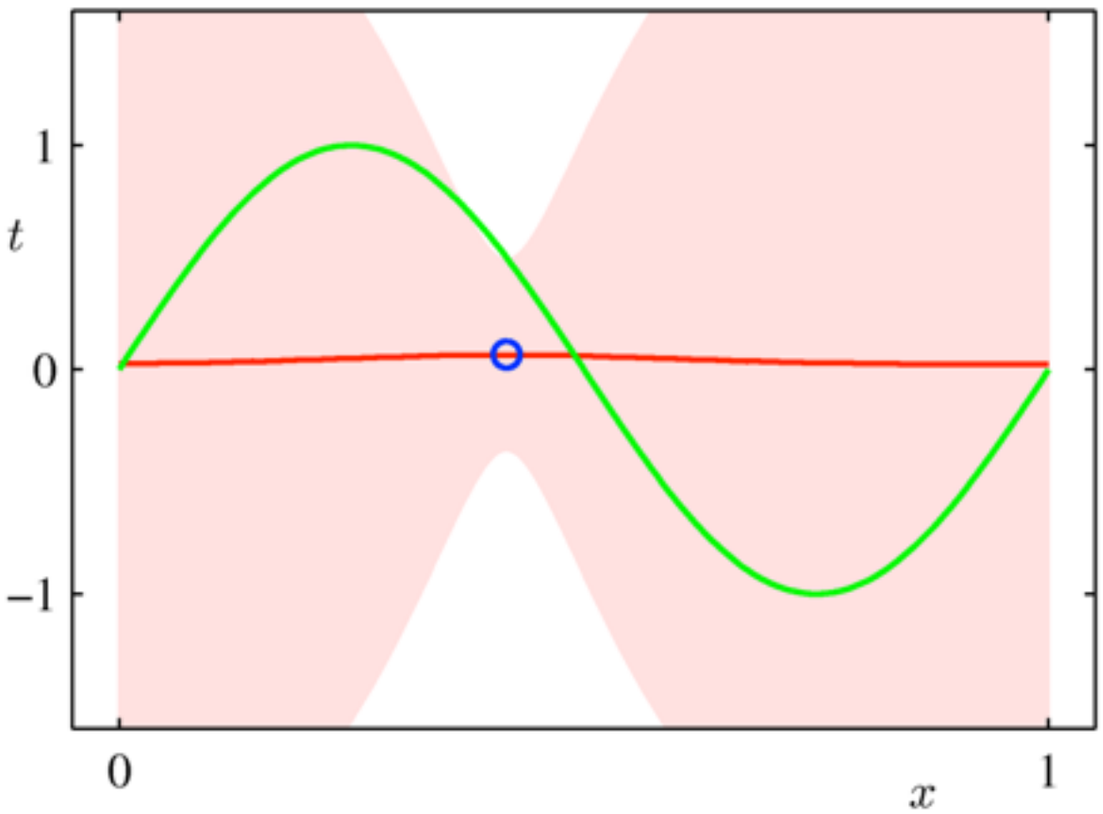
$$= \mathcal{N}(t^*; \phi(x^*)^T \boldsymbol{\mu}, \sigma_N^2(x^*))$$

where

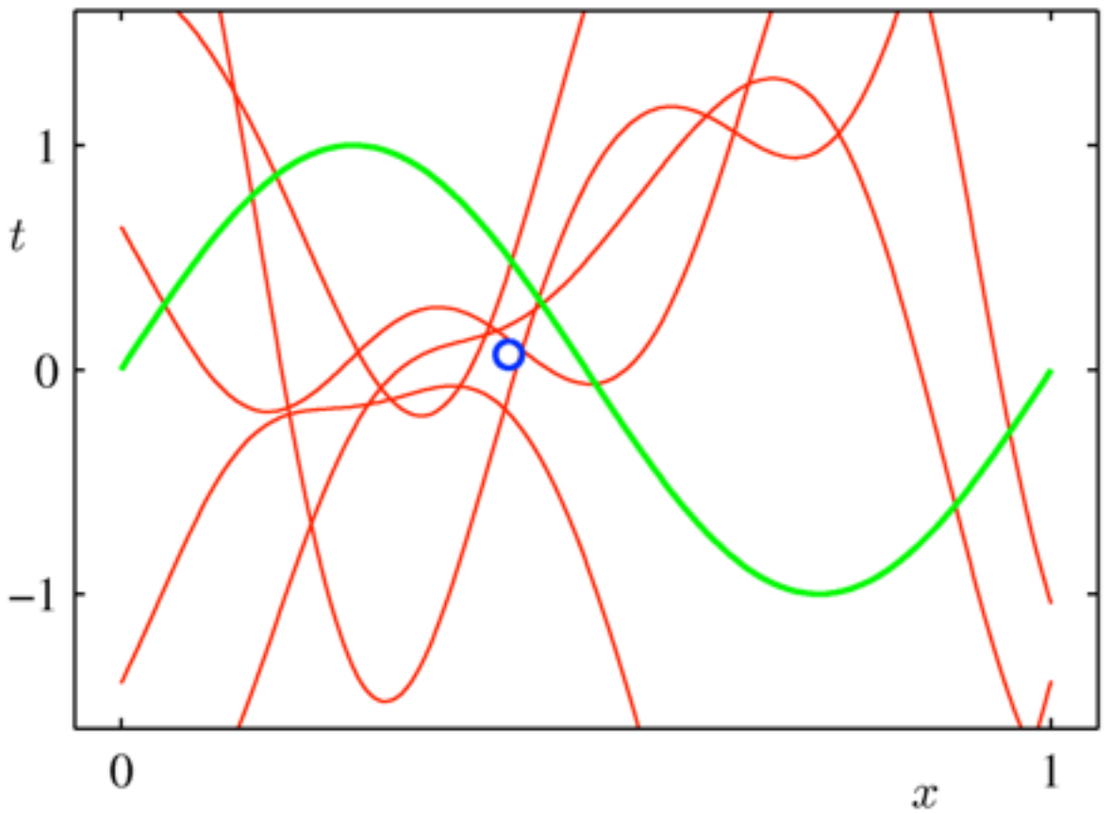
$$\sigma_N^2(x) = \sigma^2 + \phi(x)^T \Sigma \phi(x)$$

The Predictive Distribution (2)

- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point



The predictive distribution

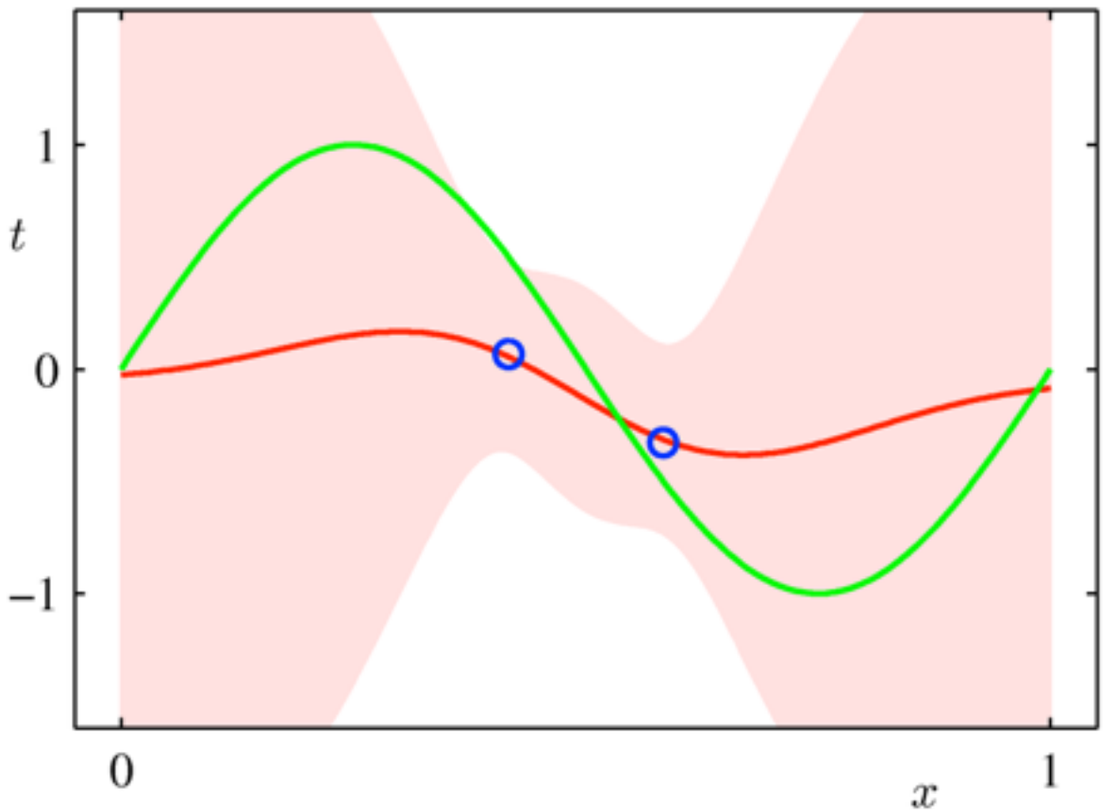


Some samples from the posterior

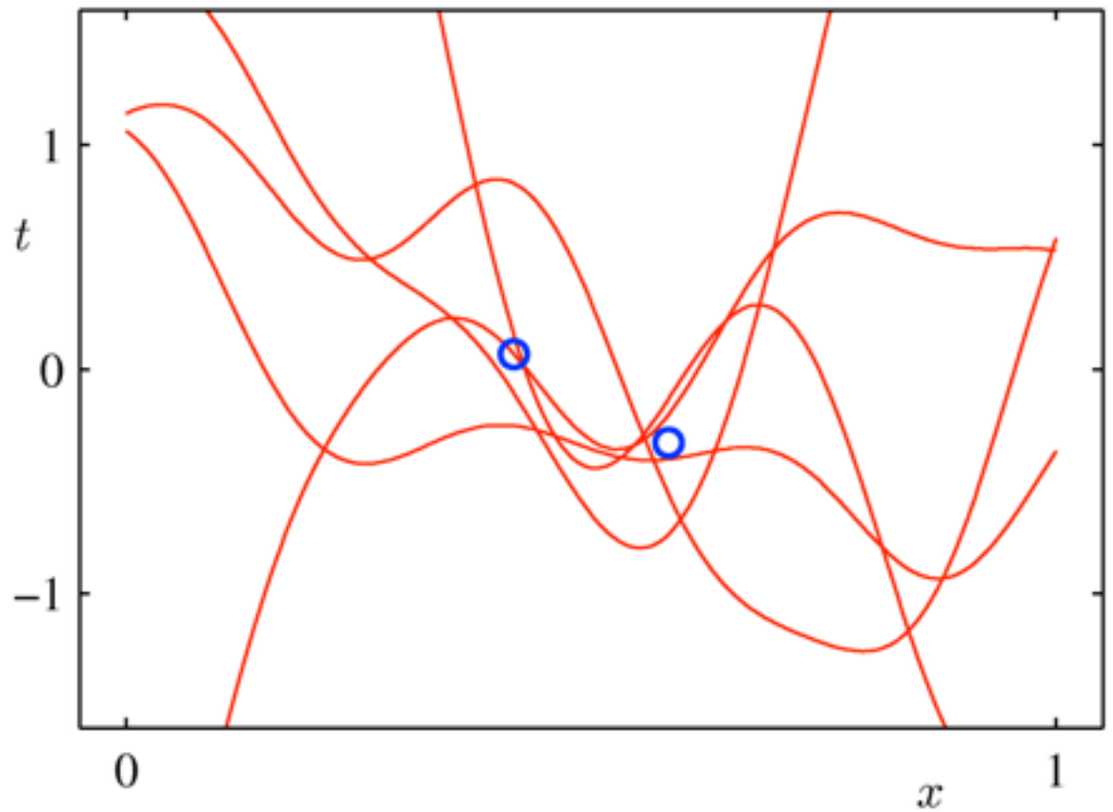
From: C.M. Bishop

Predictive Distribution (3)

- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points



The predictive distribution

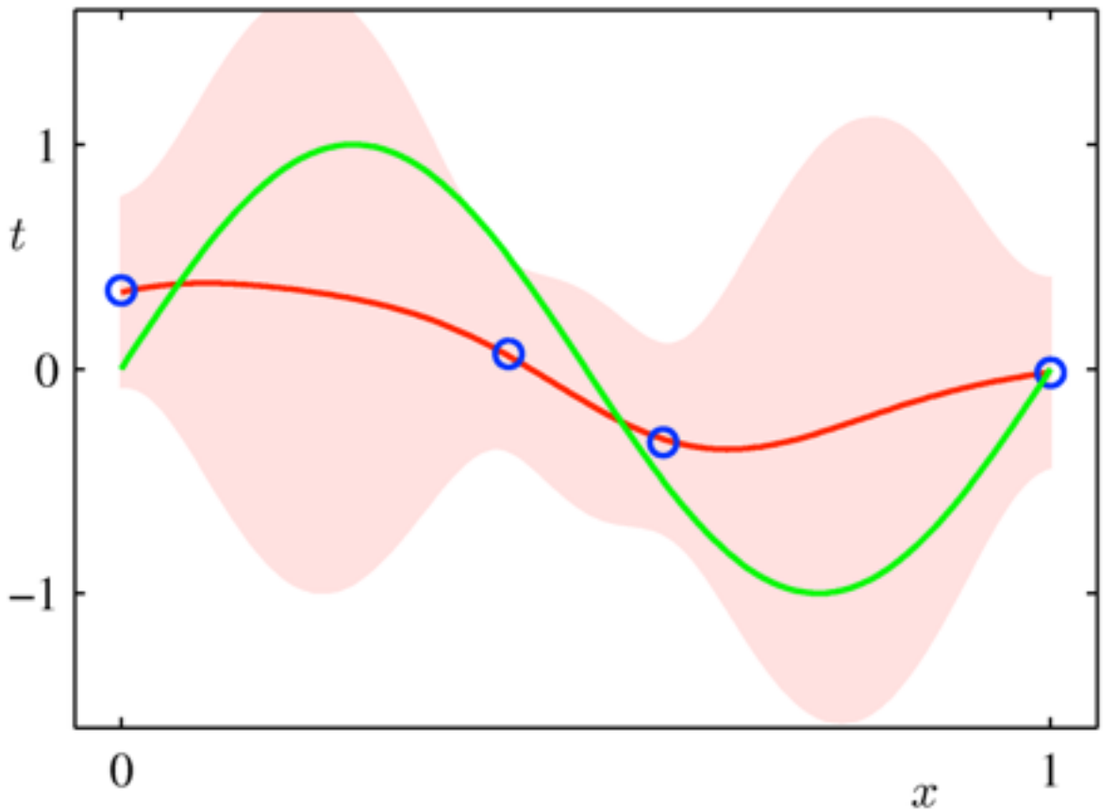


Some samples from the posterior

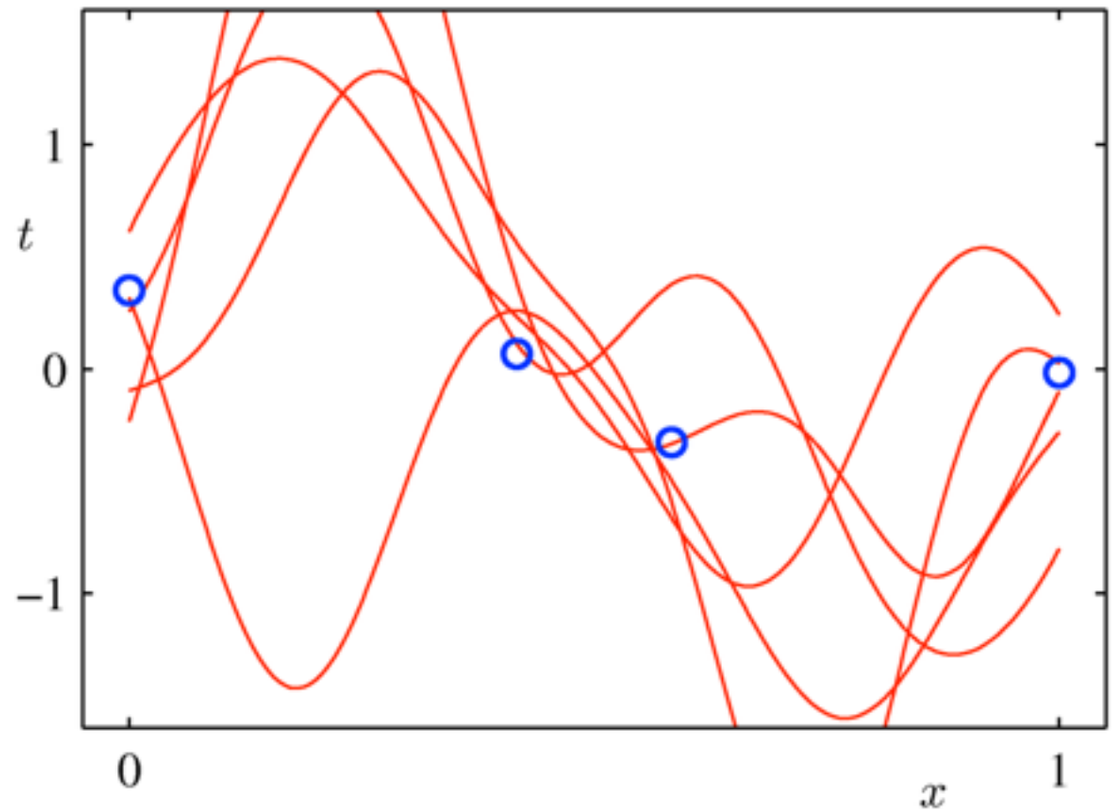
From: C.M. Bishop

Predictive Distribution (4)

- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points



The predictive distribution

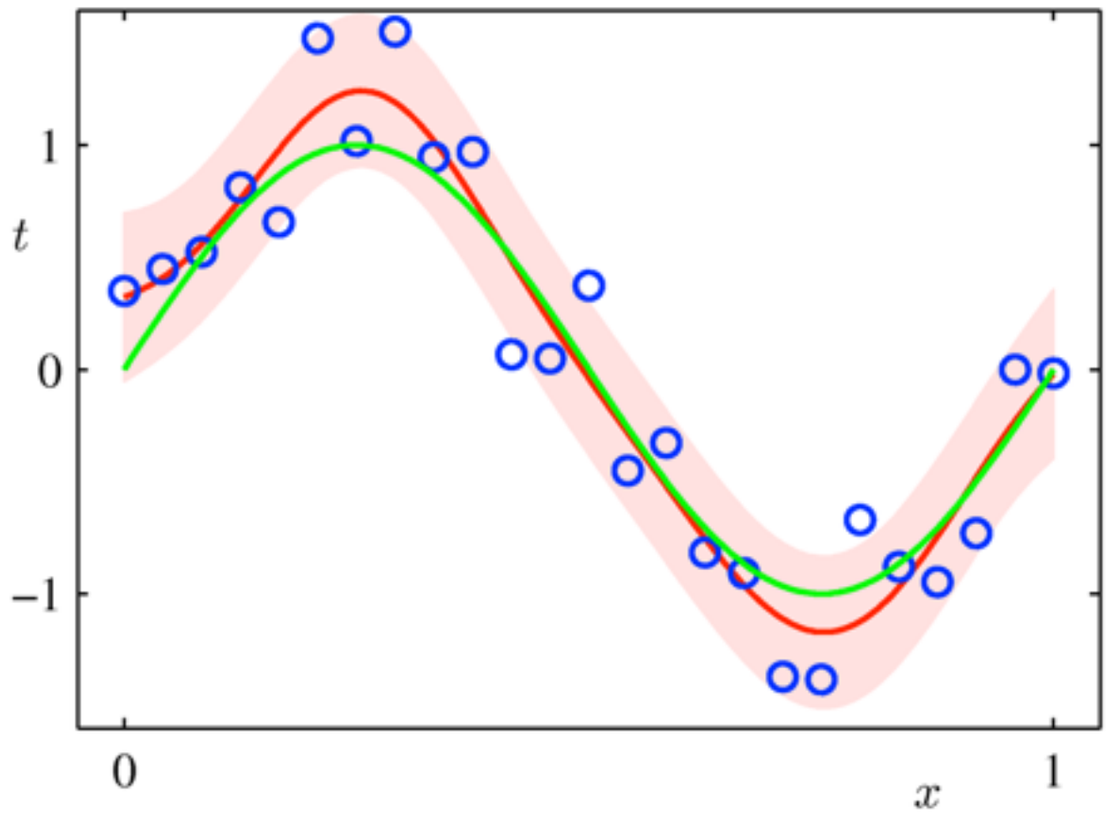


Some samples from the posterior

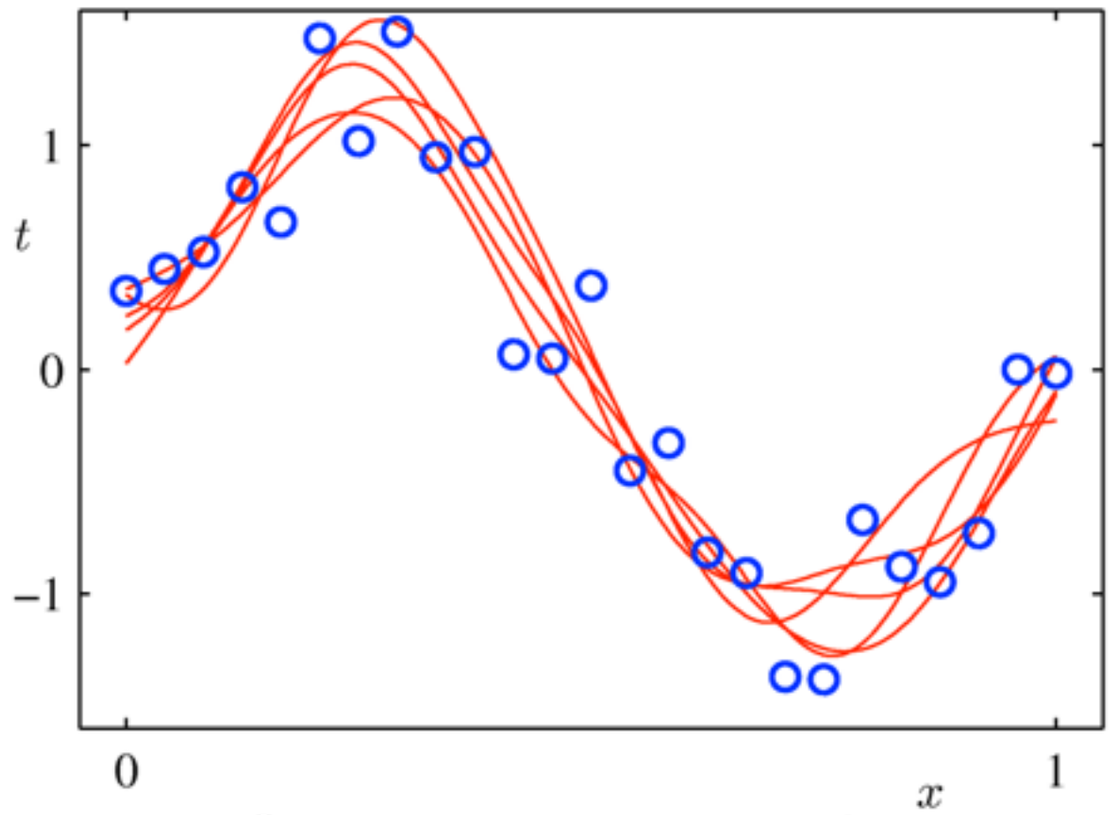
From: C.M. Bishop

Predictive Distribution (5)

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points



The predictive distribution



Some samples from the posterior

From: C.M. Bishop

Summary

- Regression can be expressed as a **least-squares** problem
- To avoid overfitting, we need to introduce a **regularisation term** with an additional parameter λ
- Regression **without** regularisation is equivalent to Maximum Likelihood Estimation
- Regression **with** regularisation is Maximum A-Posteriori
- When using Gaussian priors (and Gaussian noise), all computations can be done **analytically**
- This gives a closed form of the **parameter posterior** and the **predictive distribution**

3. Probabilistic Graphical Models

Directed Models

The Bayes Filter (Rep.)

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

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

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

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

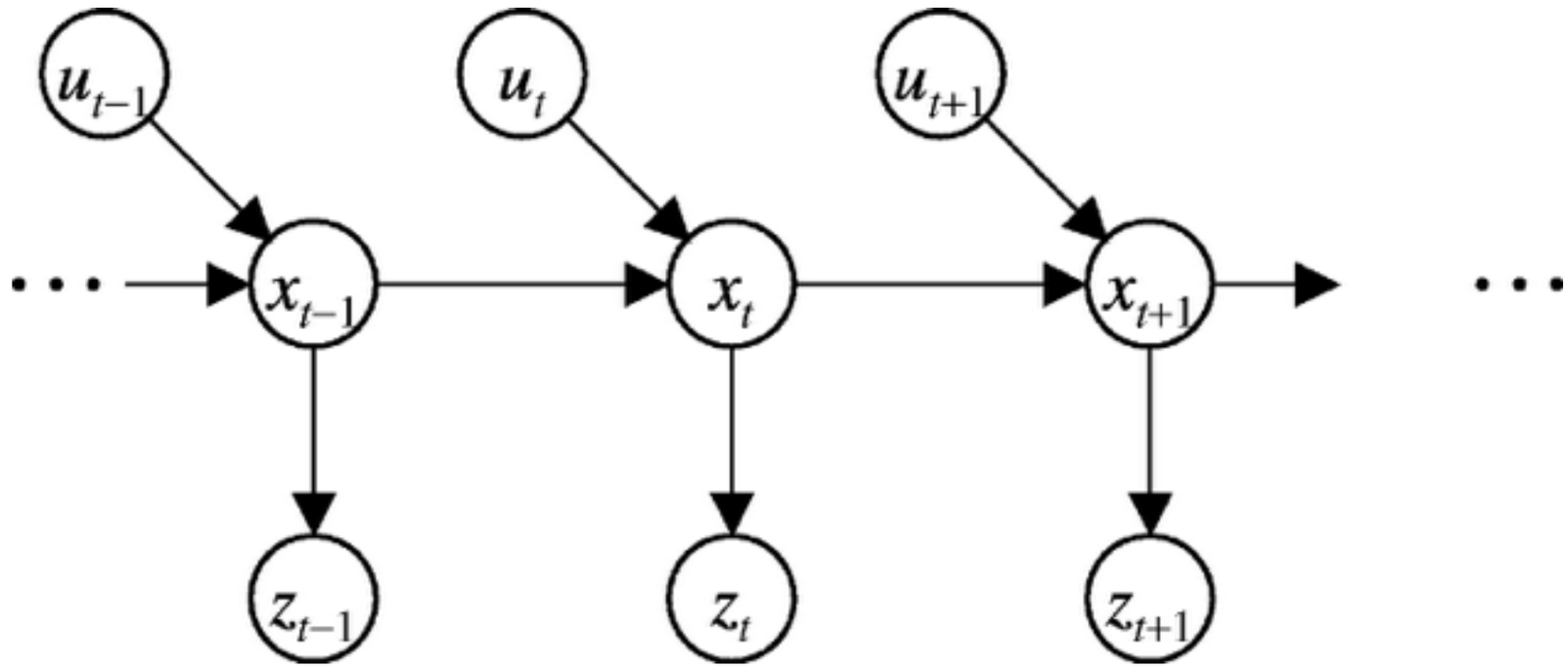
(Markov) $$= \eta\, p(z_t \mid x_t) \int p(x_t \mid u_t, x_{t-1}) p(x_{t-1} \mid u_1, z_1, \ldots, u_t) dx_{t-1}$$

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

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

Graphical Representation (Rep.)

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



- This incorporates the following Markov assumptions:

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

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

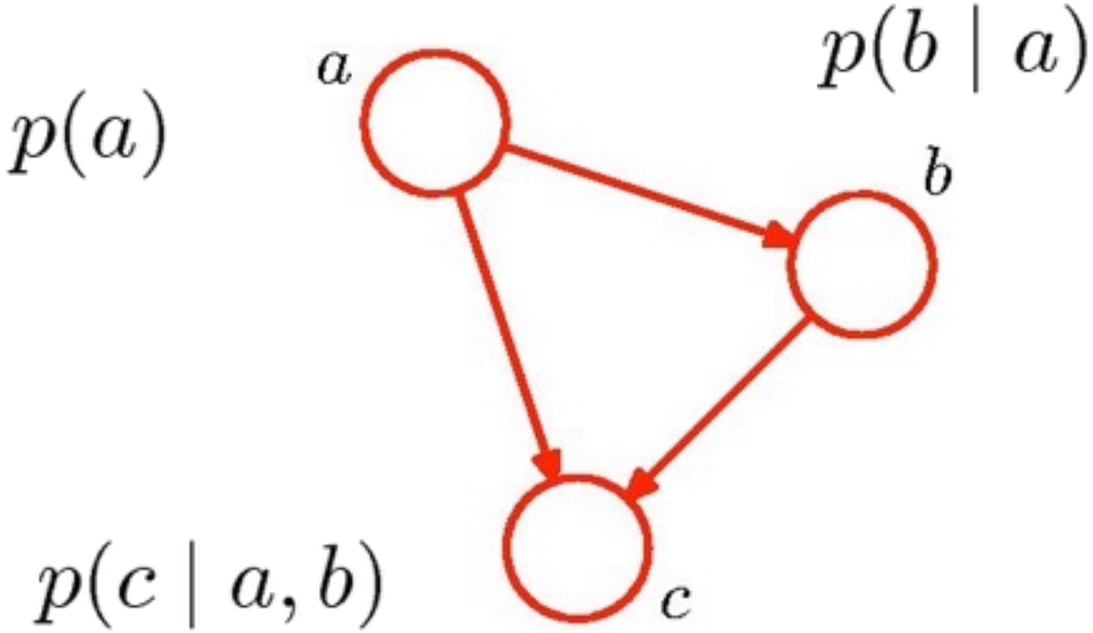
Definition

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 Acyclig 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:

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

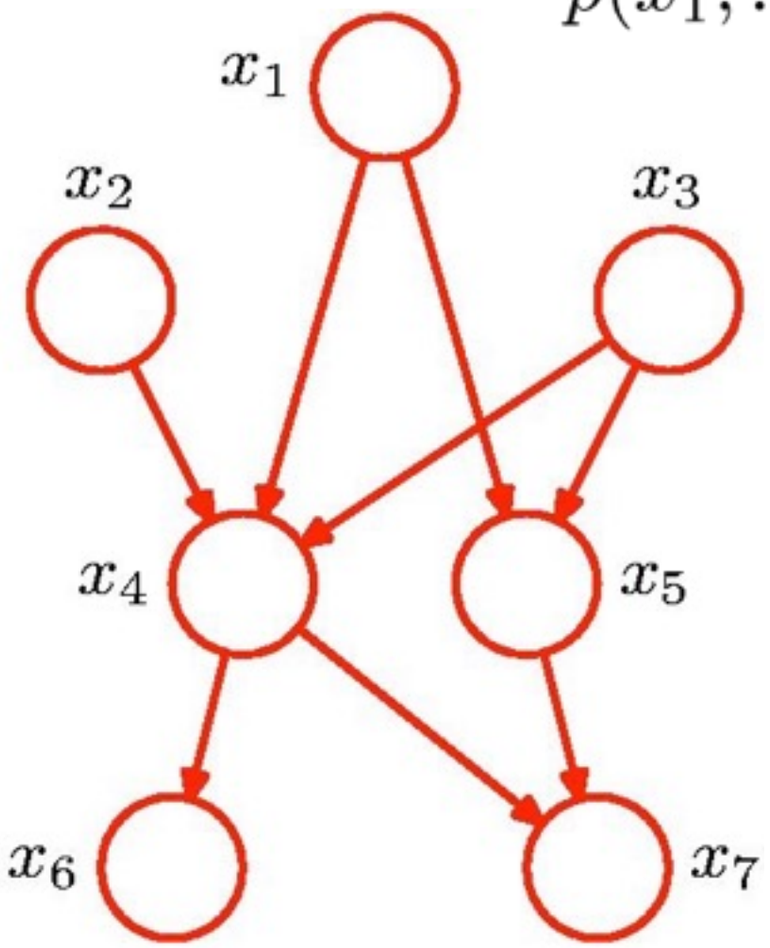
$$p(a, b, c) = p(b|a, c)p(a|c)p(c)$$

$$\vdots$$

Bayesian Networks

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

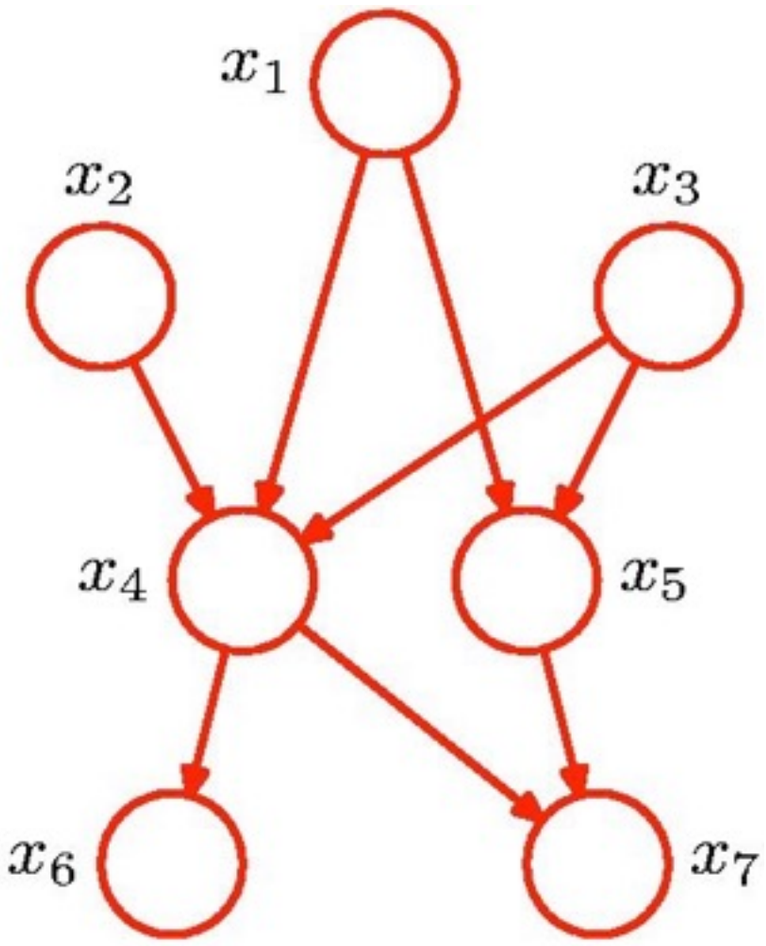
$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$
$$p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



Intuitively: only x_1 and x_3 have an influence on x_5

Bayesian Networks

We can now define a one-to-one mapping from graphical models to probabilistic formulations:



General Factorization:

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

where

$pa_k \triangleq$ ancestors of p_k

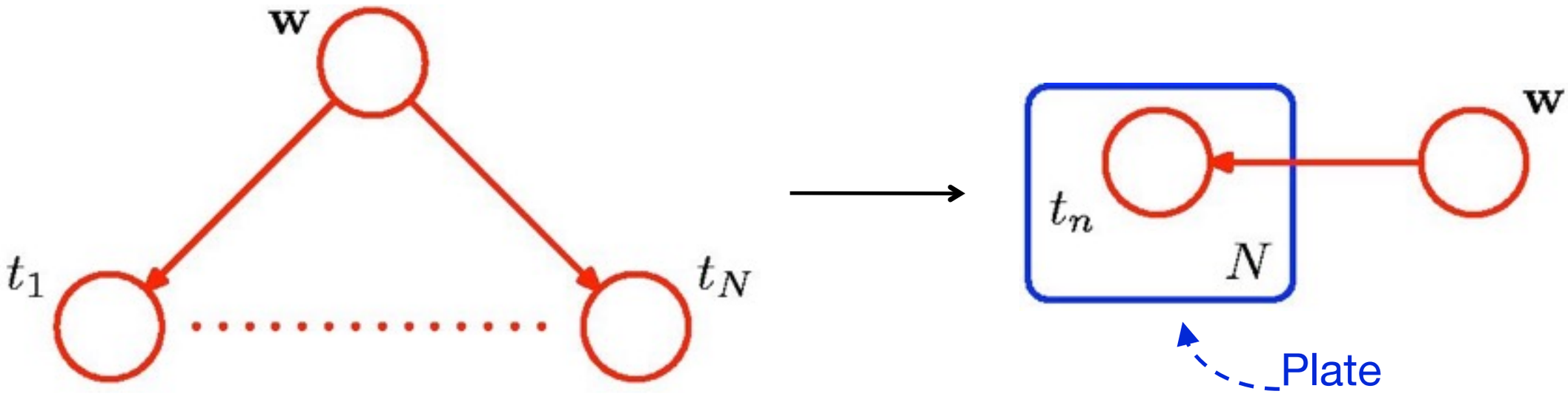
and

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

Elements of Graphical Models

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

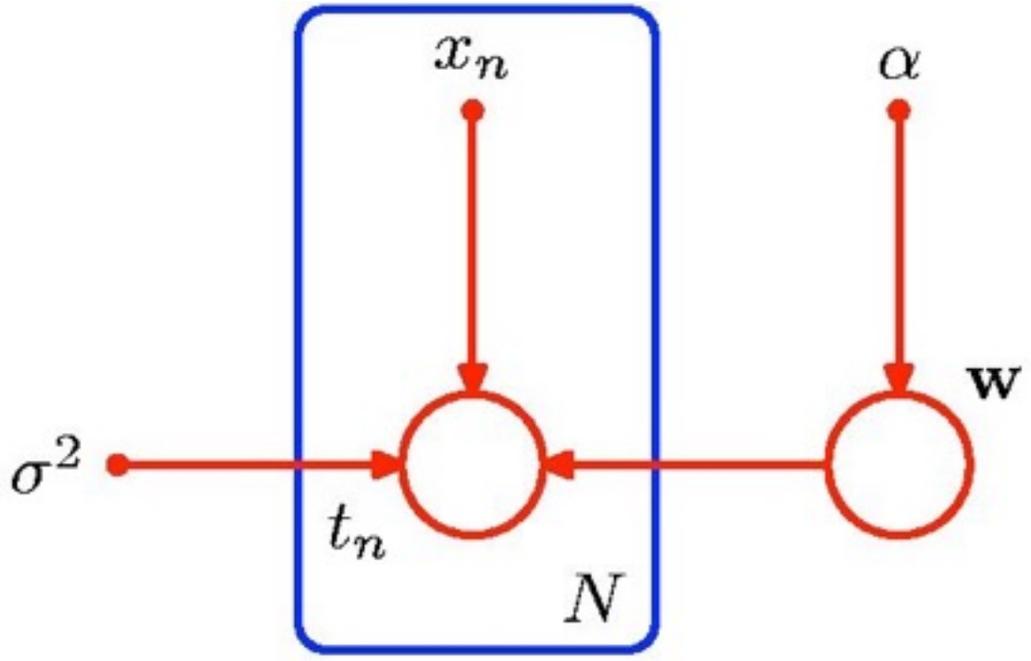
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



Elements of Graphical Models (2)

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

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

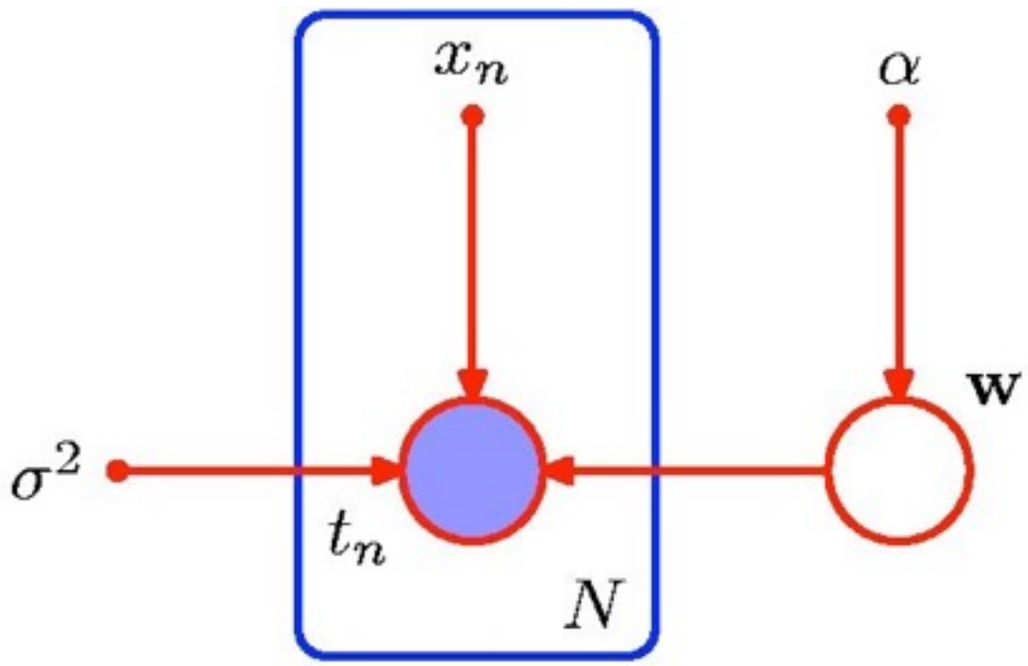


Elements of Graphical Models (3)

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)

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

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

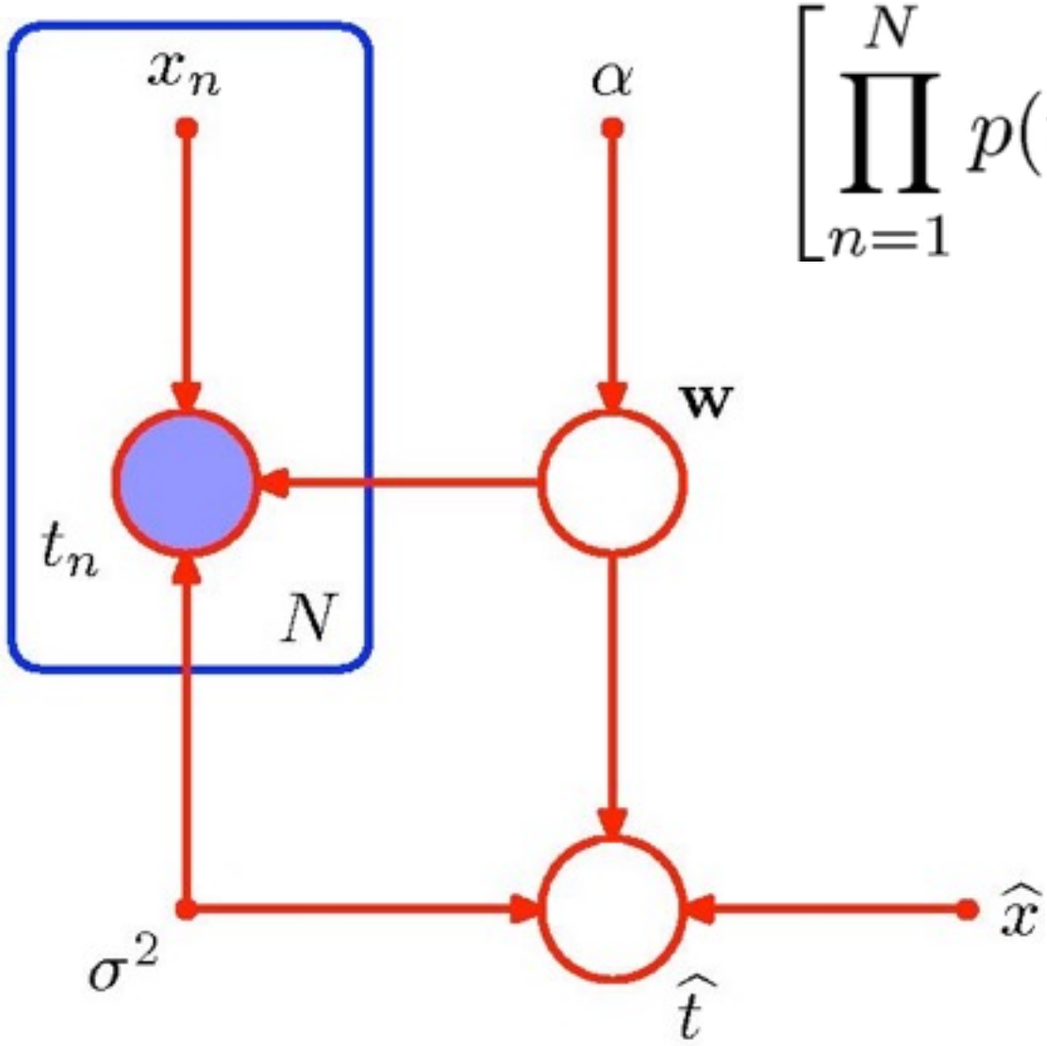
Regression as a Graphical Model

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

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

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

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



Here: conditioning on all deterministic parameters

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

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

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:

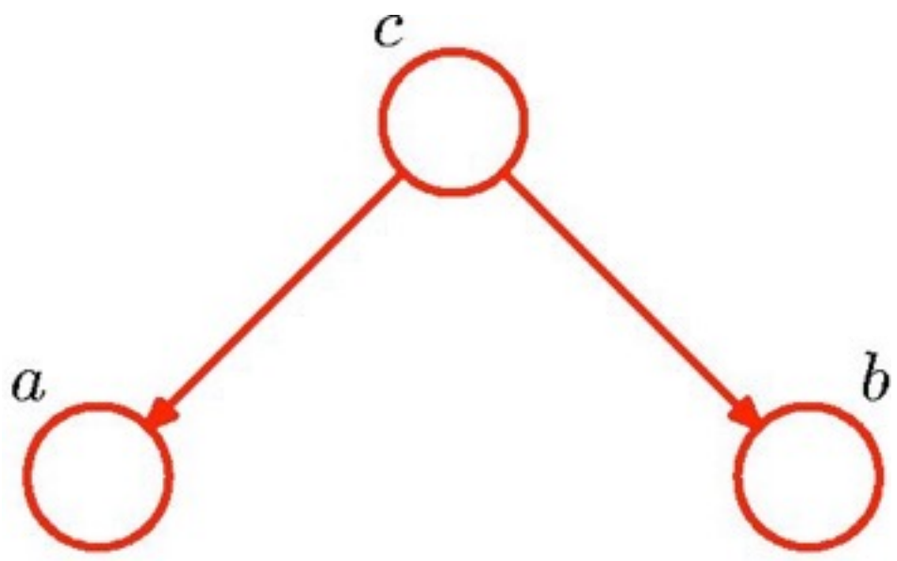
$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

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

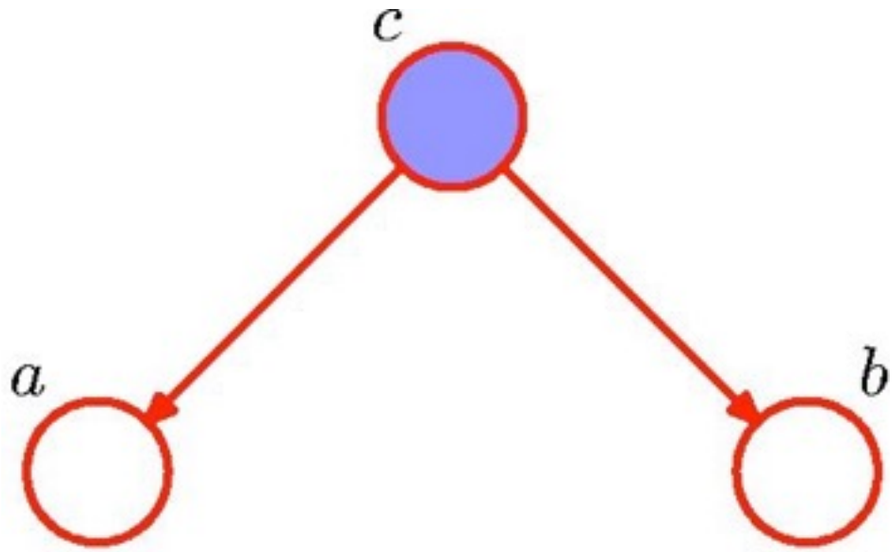
$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

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):

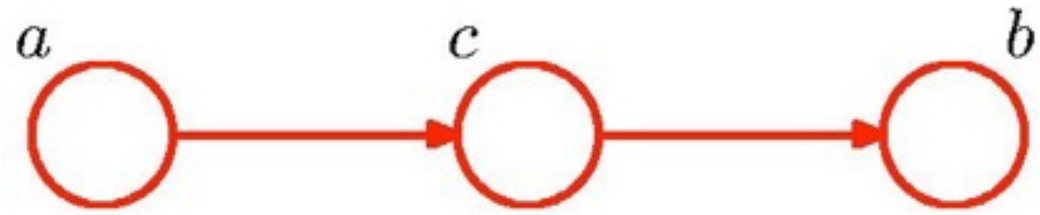


$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$
$$= p(a|c)p(b|c)$$

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:

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

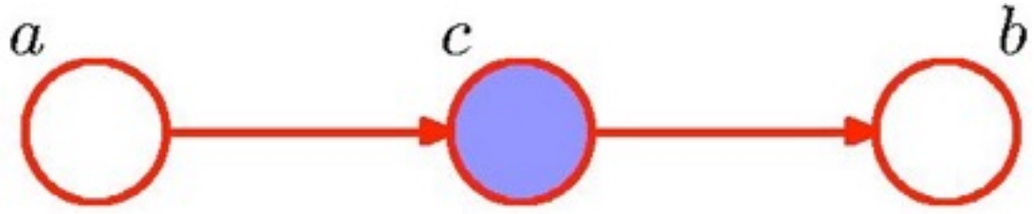
$$= p(a) \sum_c \frac{p(c, a)p(b, c, a)}{p(a)p(c, a)} = p(a) \sum_c p(b, c \mid a)$$

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

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

Conditional Independence: Example 2

As before, now we condition on c :



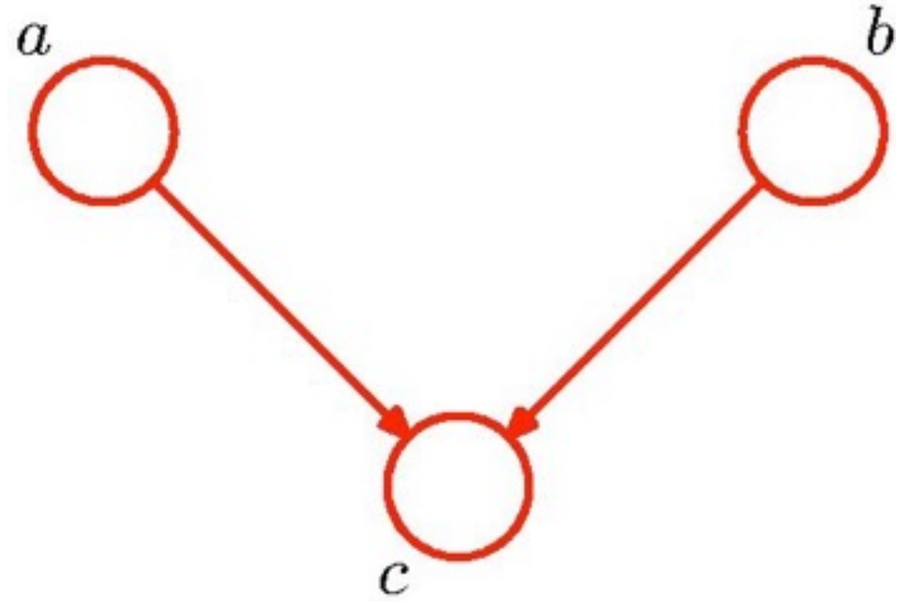
$$
\begin{aligned}
p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

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:

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

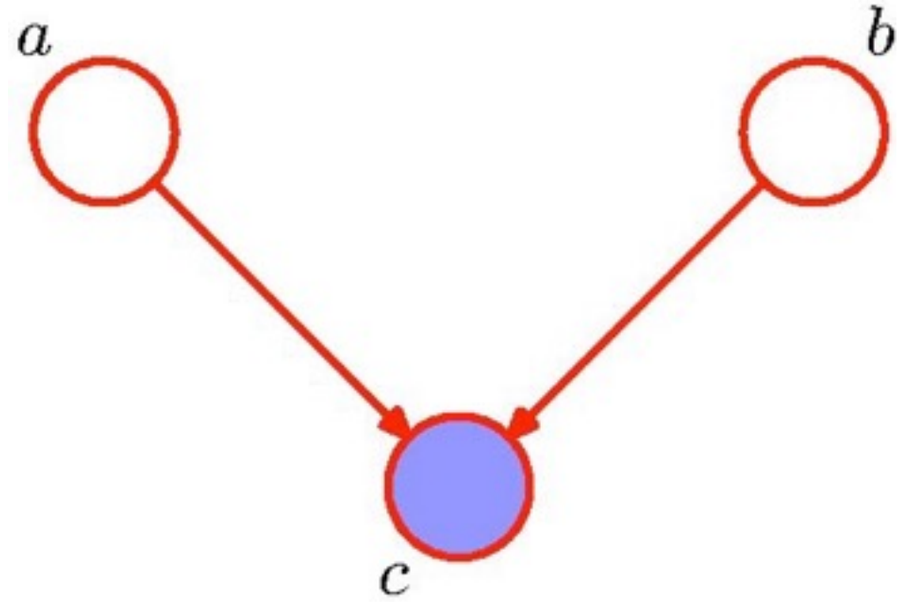
we obtain:

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

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: $a \not\perp\!\!\!\perp b \mid c$

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.

D-Separation

Say: A, B, and C are non-intersecting subsets of nodes in a directed graph.

A path from A to B is **blocked** by C if it contains a node such that either

a) the arrows on the path meet either **head-to-tail** or **tail-to-tail** at the node, and the node is **in** the set C, or

b) the arrows meet **head-to-head** at the node, and neither the node, nor any of its descendants, are in the set C.

If all paths from A to B are blocked, A is said to be **d-separated** from B by C.

Notation: $\mathrm{dsep}(A, B|C)$

D-Separation

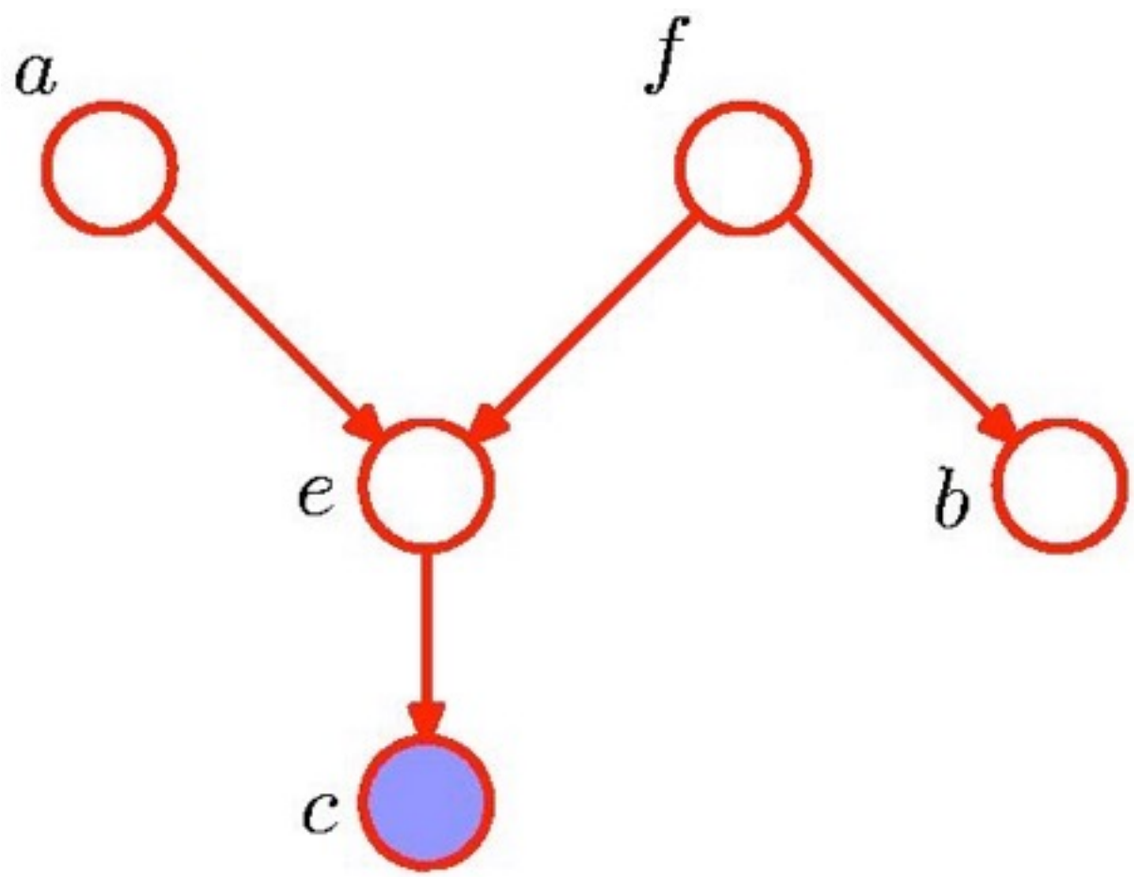
Say: A, B, and C are non-intersecting subsets of nodes in a directed graph.

- A path ... ntains a node ...
  a) the arrows ... **tail-to-tail** at the node ...
  b) the arrows ... neither the node ... C.
- If all paths ... said to be **d-separated** from B by C.

Notation: $\mathrm{dsep}(A, B|C)$

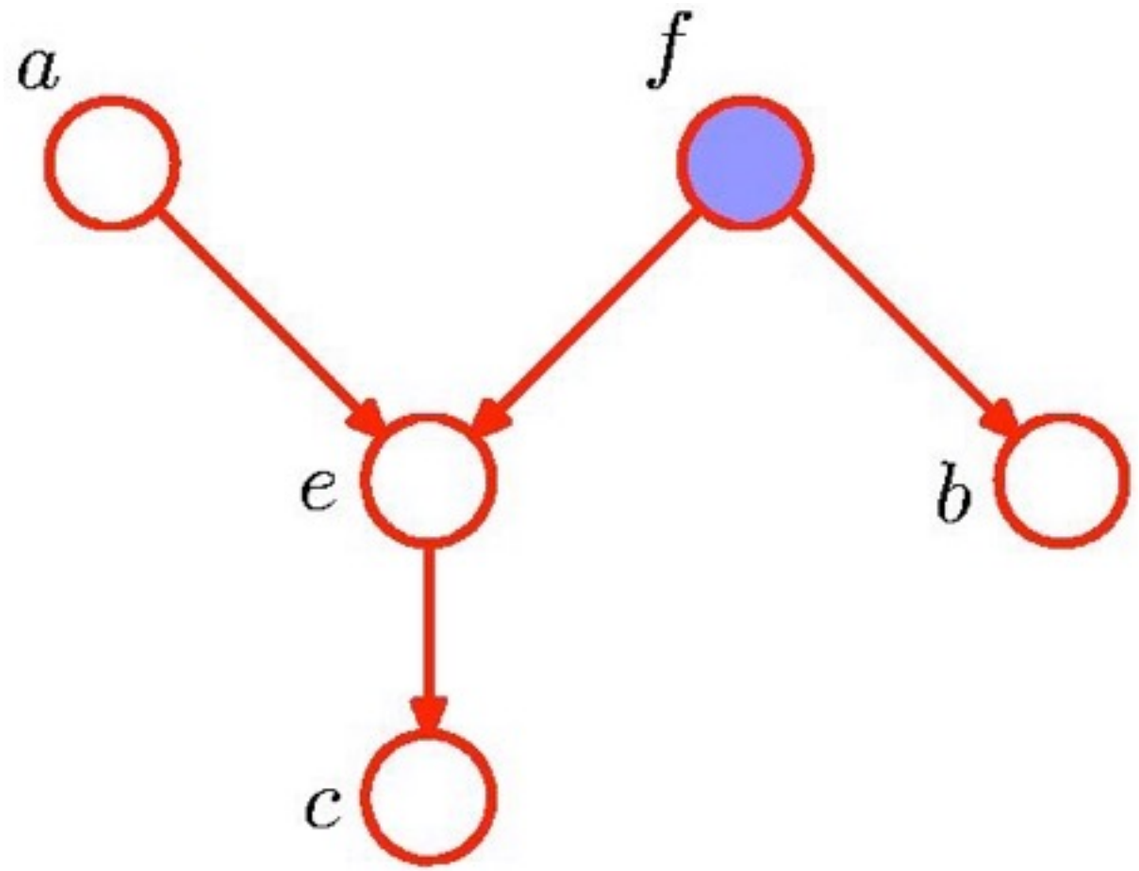
D-Separation is a property of graphs and not of probability distributions

D-Separation: Example



$$\neg \text{dsep}(a, b|c)$$

We condition on a descendant of e, i.e. it does not block the path from a to b.



$$\text{dsep}(a, b|f)$$

We condition on a tail-to-tail node on the only path from a to b, i.e f blocks the path.