

FU master's course – May 9th 2011

Machine learning

Julia Lasserre

Max Planck Institute for Molecular Genetics - Berlin

Notations

\mathcal{S} : space of features

$\{\mathbf{x}_n, t_n\}$: n^{th} training data point and its target (label or output)

\mathbf{X} : matrix of training data

\mathbf{t} : vector of training targets (labels or outputs)

\mathcal{D} : global data, can be $\{\mathbf{X}, \mathbf{t}\}$ or \mathbf{X}

$\{\hat{\mathbf{x}}, \hat{t}\}$: a new data point and its true label

$\{\mathbf{x}, \mathbf{x}'\}$: data points

$\{t, t'\}$: targets (labels or outputs)

\mathcal{M} : a particular choice of model, ex: a 2-GMM

θ : parameters (given a particular model \mathcal{M})

Bayesian networks

Plan

- Probability theory
- Graphical models
- Bayesian networks
- Inference
- Structural learning
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Plan

- *Probability theory*
- Graphical models
- Bayesian networks
- Inference
- Structural learning
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Probability theory

Probability

- a probability is a complicated mathematical object but it has a very intuitive interpretation
- *the probability of an outcome is the limiting frequency of this outcome when the corresponding experiment is repeated infinitely many times*

Probability theory

Probability distribution – discrete case

$$0 \leq p(A = a) \leq 1$$

$$\sum_{a \in \mathcal{A}} p(A = a) = 1$$

$$p(A \in [a_1, a_2]) = \sum_{a=a_1}^{a_2} p(A = a)$$

Probability theory

Probability distribution – discrete case – shortcuts

- if we do not have any particular interest in the value a

$$p(A = a) \quad \text{will be denoted} \quad p(A)$$

- $\sum_{a \in \mathcal{A}} p(A = a)$ will be denoted $\sum_A p(A)$

Probability theory

Probability density – continuous case

from the measure theory point of view: if the probability of a real-valued variable a falling in the interval $[a, a+da]$ is given by $p(a)da$ for $da \rightarrow 0$, then $p(a)$ is called the probability density over a

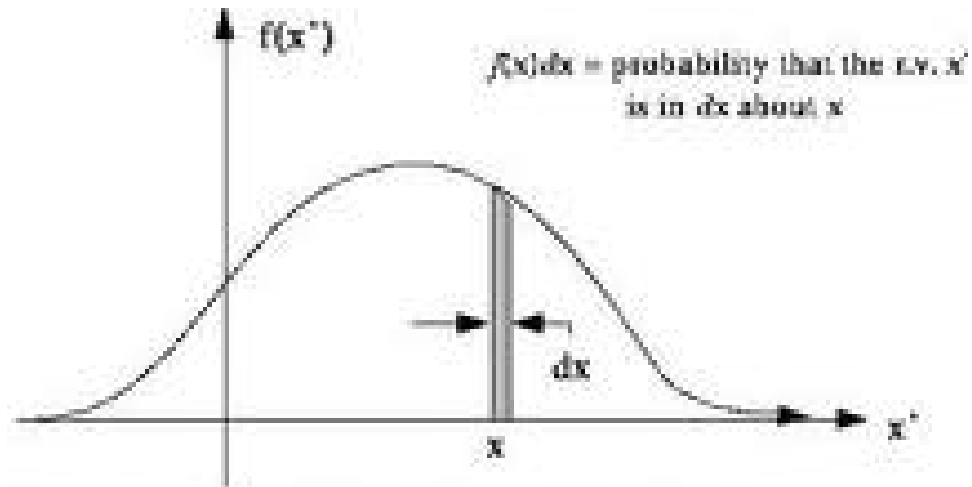


Figure 4. Typical Probability Distribution Function (pdf)

Probability theory

Probability density – continuous case

$$p(a) \geq 0$$

$$\int_{\mathcal{A}} p(a) da = 1$$

$$p(a \in [a_1, a_2]) = \int_{a_1}^{a_2} p(a) da$$

Probability theory

- Sum rule

$$p(A) = \sum_B p(A, B) \quad \text{or} \quad p(a) = \int_{\mathcal{B}} p(a, b) db$$

- Product rule

$$p(A, B) = p(A | B) p(B) \quad \text{or} \quad p(a, b) = p(a | b) p(b)$$

Probability theory

Bayes theorem

$$p(B | A) \equiv \frac{p(A, B)}{p(A)} = \frac{p(A | B) p(B)}{p(A)}$$

Probability theory

Bayes theorem

$$p(B | A) = \frac{p(A, B)}{p(A)} = \frac{p(A | B) p(B)}{p(A)}$$

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

Probability theory

Independence

- marginal:

$$A \perp B \Leftrightarrow p(A, B) = p(A)p(B)$$

- conditional:

$$A \perp B | C \Leftrightarrow p(A, B | C) = p(A | C)p(B | C)$$

Probability theory

Expectation (mean, average)

$$\mathbb{E}_A [f(A)] = \sum_A f(A) p(A) \approx \frac{1}{N} \sum_{n=1}^N f(A_n)$$

$$\mathbb{E}_a [f(a)] = \int_A f(a) p(a) da \approx \frac{1}{N} \sum_{n=1}^N f(a_n)$$

Probability theory

Variance

$$\begin{aligned}\text{var}_a [f(a)] &= \mathbb{E}_a \left[\left(f(a) - \mathbb{E}_a [f(a)] \right)^2 \right] \\ &= \mathbb{E}_a [f(a)^2] - \mathbb{E}_a [f(a)]^2 \\ &\approx \frac{1}{N} \sum_{n=1}^N \left(f(a_n) - \frac{1}{N} \sum_{n=1}^N f(a_n) \right)^2\end{aligned}$$

Standard deviation

$$\sigma_a(f(a)) = \sqrt{\text{var}_a(f(a))}$$

Probability theory

Covariance

$$\begin{aligned}\text{cov}_{a,b} [f(a), g(b)] &= \mathbb{E}_{a,b} \left[\left(f(a) - \mathbb{E}_a [f(a)] \right) \left(g(b) - \mathbb{E}_b [g(b)] \right) \right] \\ &= \mathbb{E}_{a,b} [f(a)g(b)] - \mathbb{E}_a [f(a)] \mathbb{E}_b [g(b)] \\ &\approx \frac{1}{N} \sum_{n=1}^N \left(f(a_n) - \frac{1}{N} \sum_{n=1}^N f(a_n) \right) \left(g(b_n) - \frac{1}{N} \sum_{n=1}^N g(b_n) \right)\end{aligned}$$

$$\text{cov}_a [f(a)] \equiv \text{cov}_{a,a} [f(a), f(a)] = \text{var}_a [f(a)]$$

Probability theory

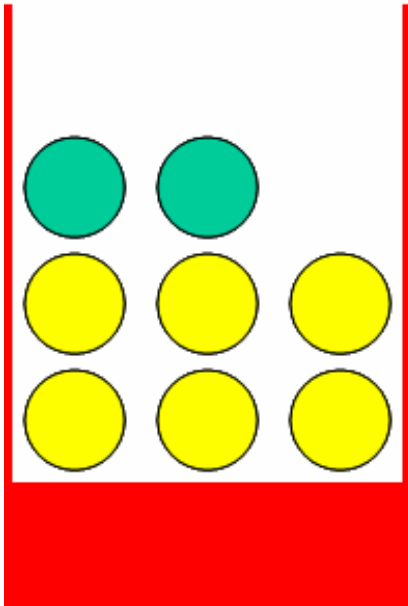
Correlation

$$\rho_{a,b} = \frac{\text{cov}(a,b)}{\sigma_a \sigma_b}$$

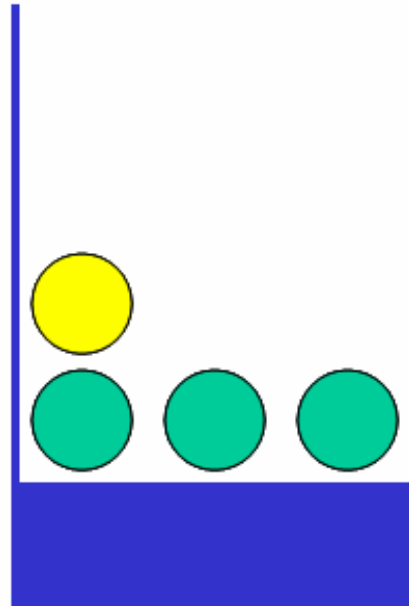
$$E_{b|a}[b] = E_b[b] + \rho_{a,b} \frac{a - E_a[a]}{\sigma_a}$$

Probability theory

Example



$$p(B = \text{red}) = 0.4$$



C : color of the ball we pick

B : color of the box we use

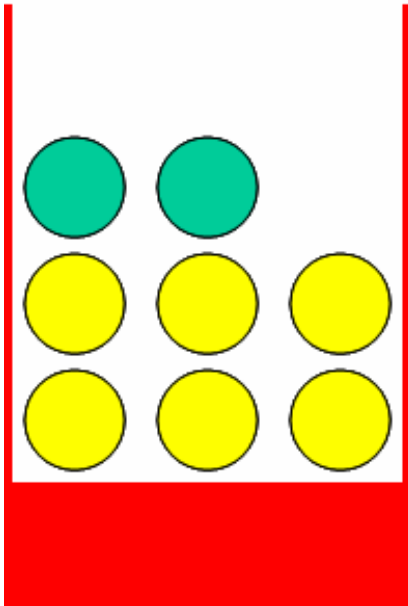
$$p(B = \text{red}) = 0.4$$

What is the probability that we pick a green ball?

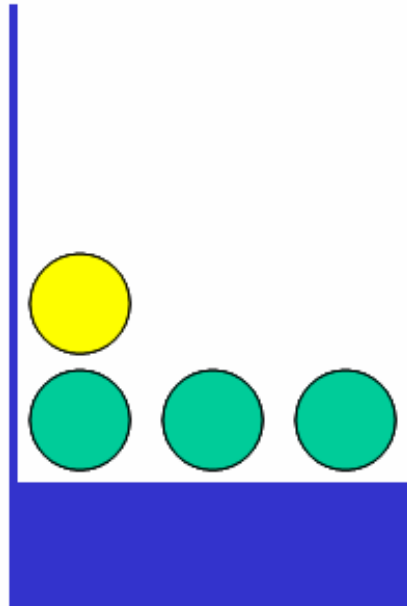
Given that we picked a yellow ball, what is the probability that we used the blue box?

Probability theory

Example



$$p(B = \text{red}) = 0.4$$

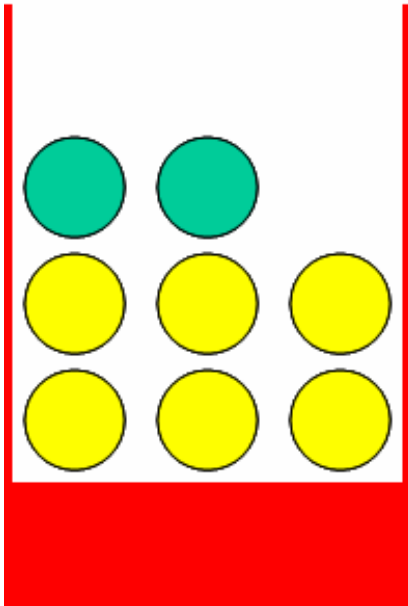


$$\begin{aligned} p(C = g) &= \sum_B p(C = g, B) \\ &= \sum_B p(C = g | B) p(B) \\ &= p(C = g | B = r) p(B = r) + \\ &\quad p(C = g | B = b) p(B = b) \\ &= \frac{1}{4} \frac{4}{10} + \frac{3}{4} \frac{6}{10} \end{aligned}$$

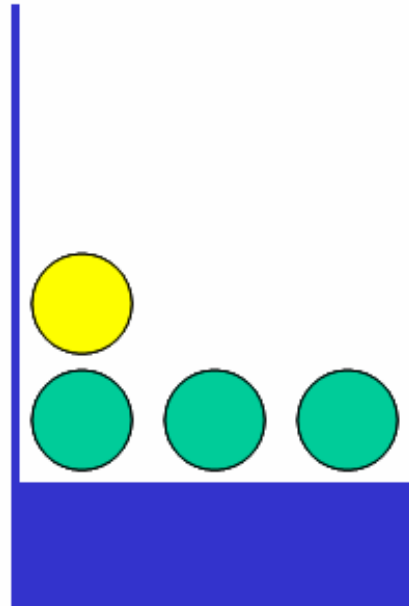
$$p(C = \text{green}) = \frac{11}{20}$$

Probability theory

Example



$$p(B = \text{red}) = 0.4$$



C : color of the ball we pick

B : color of the box we use

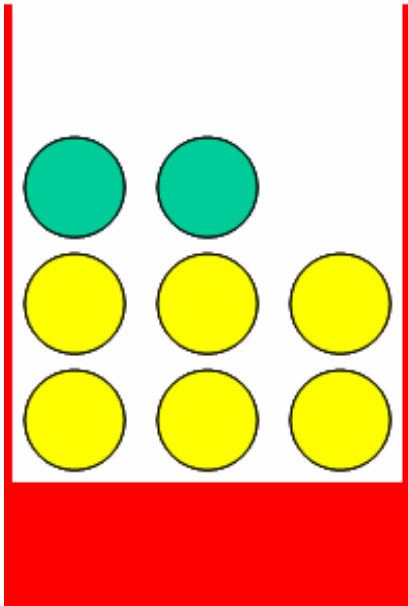
$$p(B = \text{red}) = 0.4$$

What is the probability that we pick a green ball?

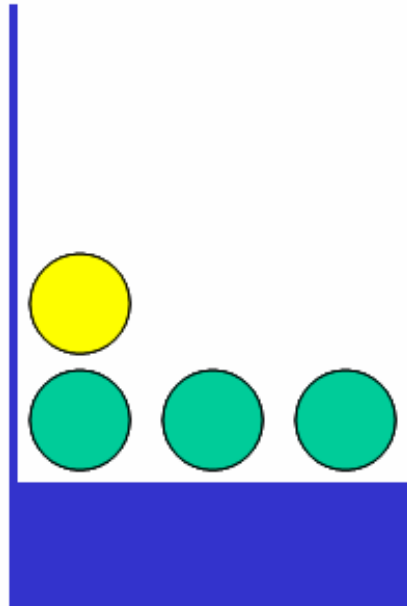
Given that we picked a yellow ball, what is the probability that we used the blue box?

Probability theory

Example



$$p(B = \text{red}) = 0.4$$



$$\begin{aligned} p(B = b | C = y) &= \frac{p(C = y | B = b)p(B = b)}{p(C = y)} \\ &= \frac{1}{4} \frac{6}{10} \times \left(1 - \frac{11}{20}\right)^{-1} \\ &= \frac{3}{20} \frac{20}{9} \end{aligned}$$

$$p(B = \text{blue} | C = \text{yellow}) = \frac{1}{3}$$

Probability theory

Bayesian thinking

- $p(B = \text{blue})$ is the **prior** probability of using the blue box
"before" we can observe which type of ball we have picked
- we can now see the color of the ball
- $p(B = \text{blue} | C)$ is the **posterior** probability of using the blue box
"after" we have observed which type of ball we have picked
- note that we compute $p(B | C)$ even though C happens after B
 $\Rightarrow p(y | x)$ does not imply a causal link $x \rightarrow y$

Plan

- *Probability theory*
- *Graphical models*
- Bayesian networks
- Inference
- Structural learning
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Graphical models

What are they?

- a cross between probability theory and graph theory
- the nodes represent random variables
- the arcs and the lack of arcs show the conditional independencies between the random variables
- leads to a compact representation of the joint distribution

Graphical models

What are they?

- undirected graphs
 - Markov networks / Markov random fields
 - popular in physics and vision

- directed graphs
 - Bayesian networks / belief networks
 - popular in machine learning

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- Inference
- Structural learning
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

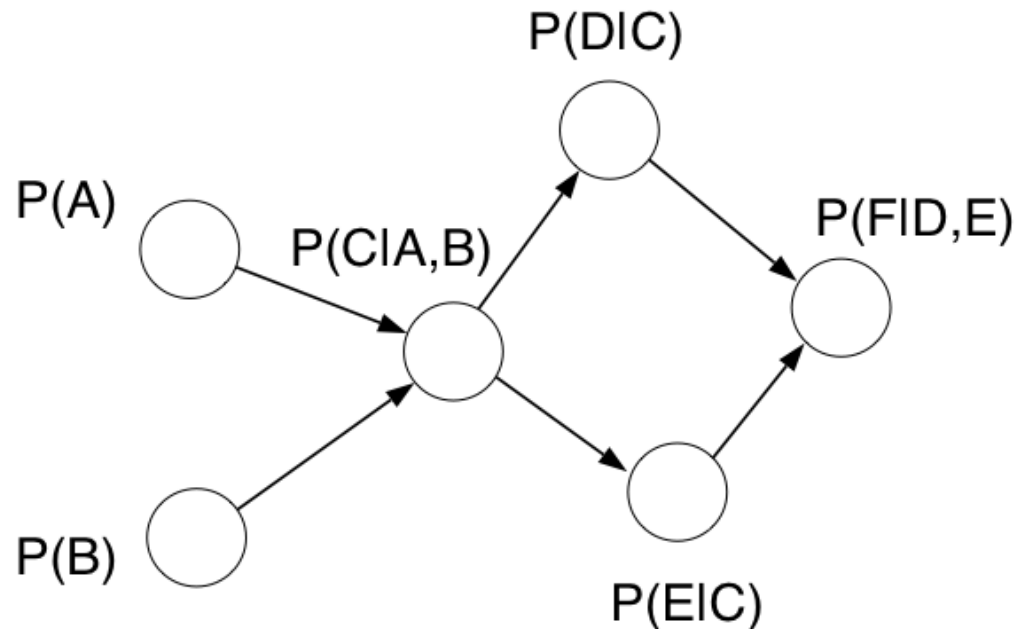
Bayesian networks

- Directed acyclic graphs
- A few myths
 - they require a causal semantics for the edges
 - they are necessarily Bayesian
 - they are intractable

Bayesian networks

Quantitative specification

- how do we specify a joint distribution for the nodes in the graph?
- simply define local distributions and multiply them



Bayesian networks

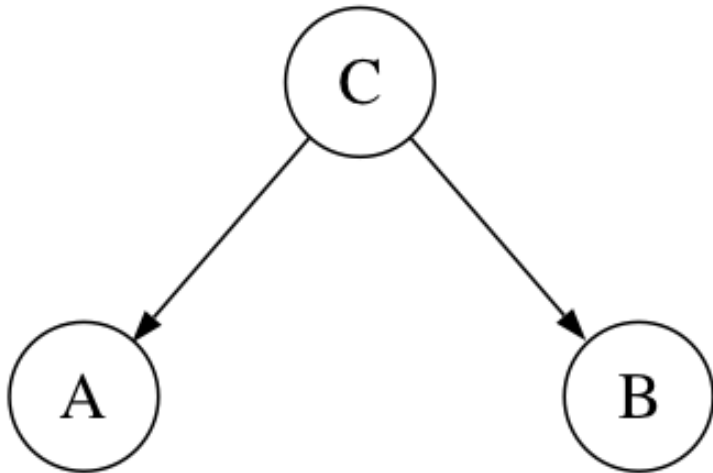
Quantitative specification

- how do we specify a joint distribution for the nodes in the graph?
- simply define local distributions and multiply them

$$p(S_1, \dots, S_K) = \prod_{i=1}^K p(S_i | \text{pa}(S_i))$$

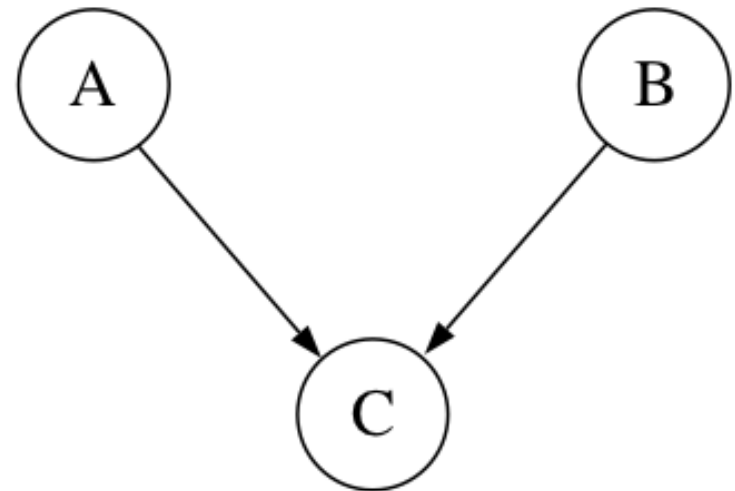
Bayesian networks

Qualitative specification



$$p(A, B, C) = p(C)p(A | C)p(B | C)$$

A and B are marginally dependent
A and B are conditionally independent

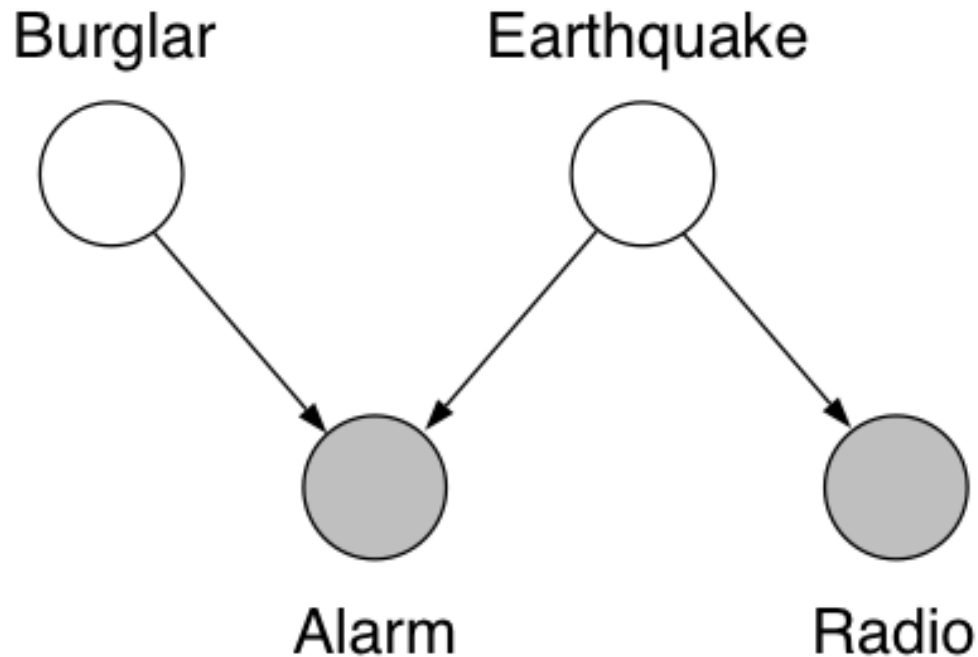


$$p(A, B, C) = p(A)p(B)p(C | A, B)$$

A and B are marginally independent
A and B are conditionally dependent

Bayesian networks

Qualitative specification



All connections are excitatory (in both directions). But an increase in activation of Earthquake leads to a decrease in activation of Burglar.

Bayesian networks

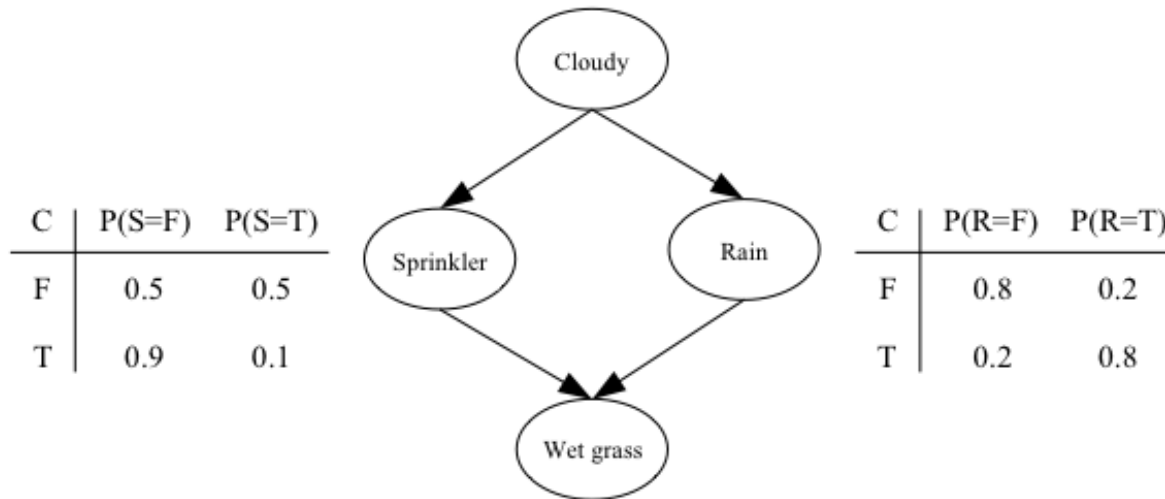
Compact representation of the likelihood

- a node is independent of its siblings given its parents
- a node is independent of its ancestors given its parents
- these 2 rules lead to a reduction of the number of parameters
 - suppose every node encodes a random binary variable
 - with the chain rule, the joint distribution has in the order of 2^N parameters, where N is the number of nodes
 - with a graphical model, only around $N(2^k)$ parameters are required, where k is the maximum number of parents

Bayesian networks

Compact representation of the likelihood – example

$P(C=F)$	$P(C=T)$
0.5	0.5



C	$P(S=F)$	$P(S=T)$
F	0.5	0.5
T	0.9	0.1

C	$P(R=F)$	$P(R=T)$
F	0.8	0.2
T	0.2	0.8

S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Bayesian networks

Compact representation of the likelihood – example

- with the chain rule

$$p(C, S, R, W) = p(C)p(S | C)p(R | S, C)p(W | R, S, C)$$

- using the graphical model

$$p(C, S, R, W) = p(C)p(S | C)p(R | C)p(W | R, S)$$

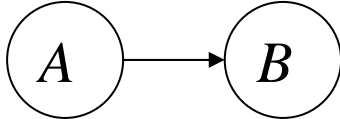
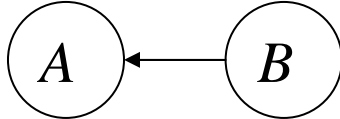
Bayesian networks

What do they show?

- how to break the joint distribution
- conditional independences
- directionality does NOT imply CAUSALITY

Bayesian networks

Directionality does not always matter

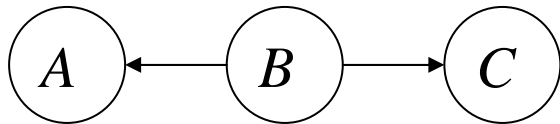
 is equivalent to 

$$p(A, B) = p(B)p(A | B) \qquad p(A, B) = p(A)p(B | A)$$

- the earthquake causes the radio report so in causal terms $E \rightarrow R$
- however, we can still compute $p(E|R)$ and find that $p(E|R) \neq p(E)$
 - R carries information about the state of E
 - in probability terms, $E \rightarrow R$ and $R \rightarrow E$

Bayesian networks

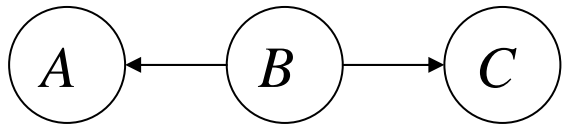
Directionality indicates independencies



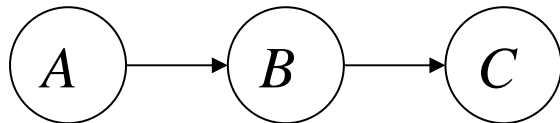
$$p(A, B, C) = p(B)p(A | B)p(C | B) \Leftrightarrow A \perp C | B$$

Bayesian networks

Directionality indicates independencies



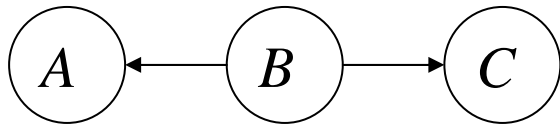
$$p(A, B, C) = p(B)p(A | B)p(C | B) \Leftrightarrow A \perp C | B$$



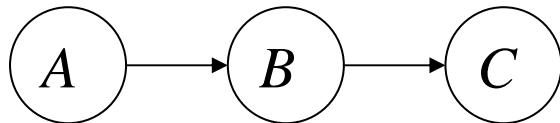
$$\begin{aligned} p(A, B, C) &= p(A)p(B | A)p(C | B) \\ &= p(B)p(A | B)p(C | B) \Leftrightarrow A \perp C | B \end{aligned}$$

Bayesian networks

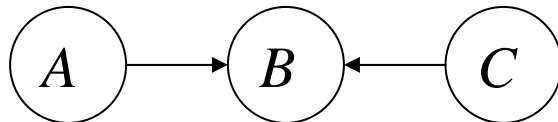
Directionality indicates independencies



$$p(A, B, C) = p(B)p(A | B)p(C | B) \Leftrightarrow A \perp C | B$$



$$\begin{aligned} p(A, B, C) &= p(A)p(B | A)p(C | B) \\ &= p(B)p(A | B)p(C | B) \Leftrightarrow A \perp C | B \end{aligned}$$



$$p(A, B, C) = p(A)p(C)p(B | A, C) \not\Leftrightarrow A \perp C | B$$

Bayesian networks

What do they show?

- how to break the joint distribution
- conditional independences
- directionality does NOT imply CAUSALITY
- equivalence classes of Bayesian networks
 - programs return one particular instance
 - most directions do not imply causality
 - do not make conclusions on biology semantics

Bayesian networks

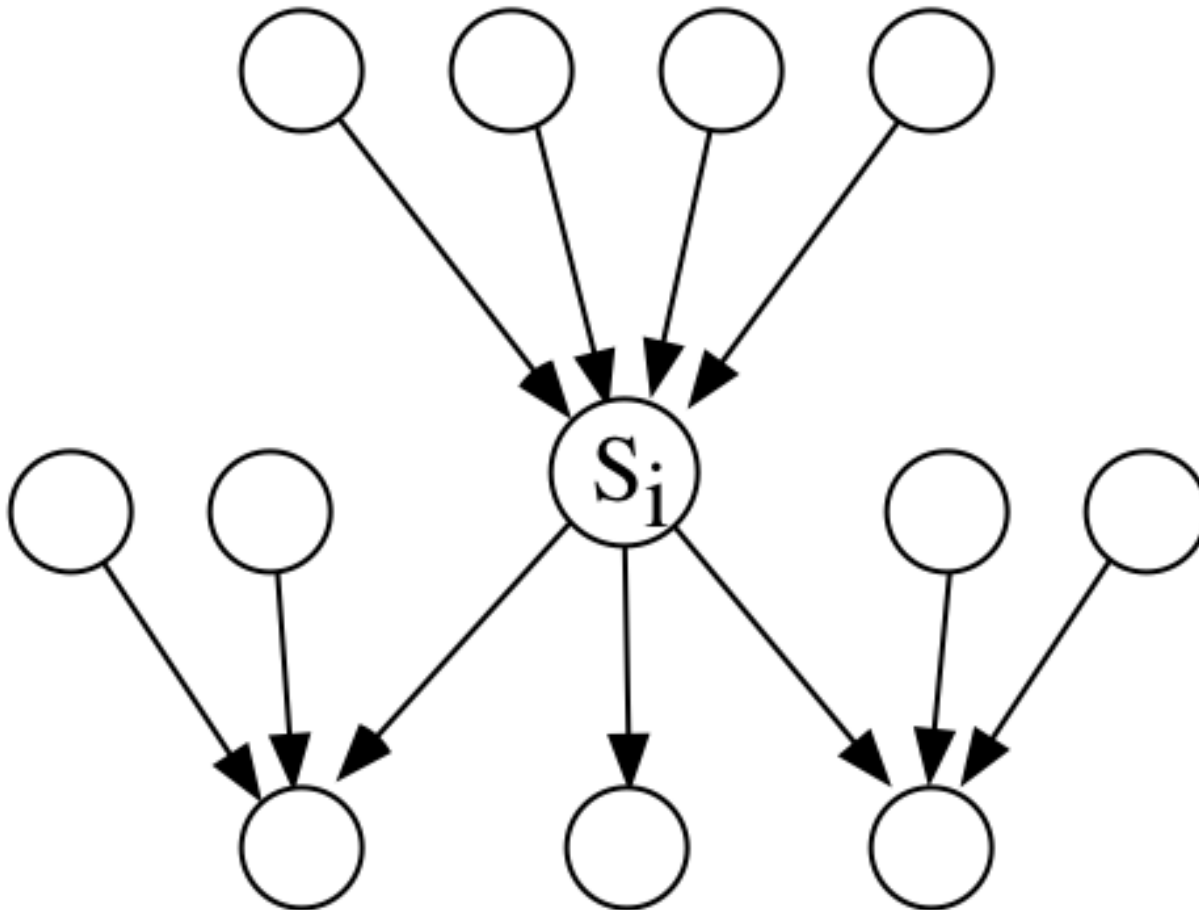
Markov blanket

- the minimal set of nodes that make the node S_i conditionally independent of all other nodes
- for Bayesian networks, it is the set of parents, children and co-parents

$$p(S_i | S) = p(S_i | \text{pa}(S_i), \text{ch}(S_i), \text{copa}(S_i))$$

Bayesian networks

Markov blanket



Bayesian networks

Evidence

- absorbing evidence means observing the values of certain of the nodes
- this process divides the nodes of the networks into 2 groups: the observed nodes O and the hidden nodes H

Bayesian networks

Inference

- doing inference means computing $p(H|O)$, the distribution of the hidden variables conditioned on the observed variables
- using Bayes theorem

$$p(H | O) = \frac{p(H, O)}{p(O)} = \frac{p(H, O)}{\sum_H p(H, O)}$$

- predictions and diagnoses are special cases of inference
 - predictions: roots are observed
 - diagnoses: leaves are observed

Bayesian networks

Inference

- inference is often intractable because of the summation
- speed-up methods such as variable elimination or dynamic programming
- approximate inference with methods such as MCMC or variational inference

Bayesian networks

Learning

- the learning process is the estimation of the parameters (and/or of the structure) of the graph
- different methods depending on the information available

structure	Non-Bayesian		Bayesian
	full observability	partial observability	
known	closed form	EM	inference
unknown	local search	structural EM	sampling methods

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- Structural learning
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Inference

- Elegant probabilistic reasoning
- In theory, we can infer the state of any variable

$$p(S_i | S_j) = \frac{p(S_i, S_j)}{p(S_j)} = \frac{\sum_{S_k, k \neq i, k \neq j} p(S_1, \dots, S_K)}{\sum_{S_k, k \neq j} p(S_1, \dots, S_K)}$$

- But it is tricky (marginalising is expensive)
approximations: variational inference, MCMC, etc

Inference

Example: burglar, alarm and earthquake

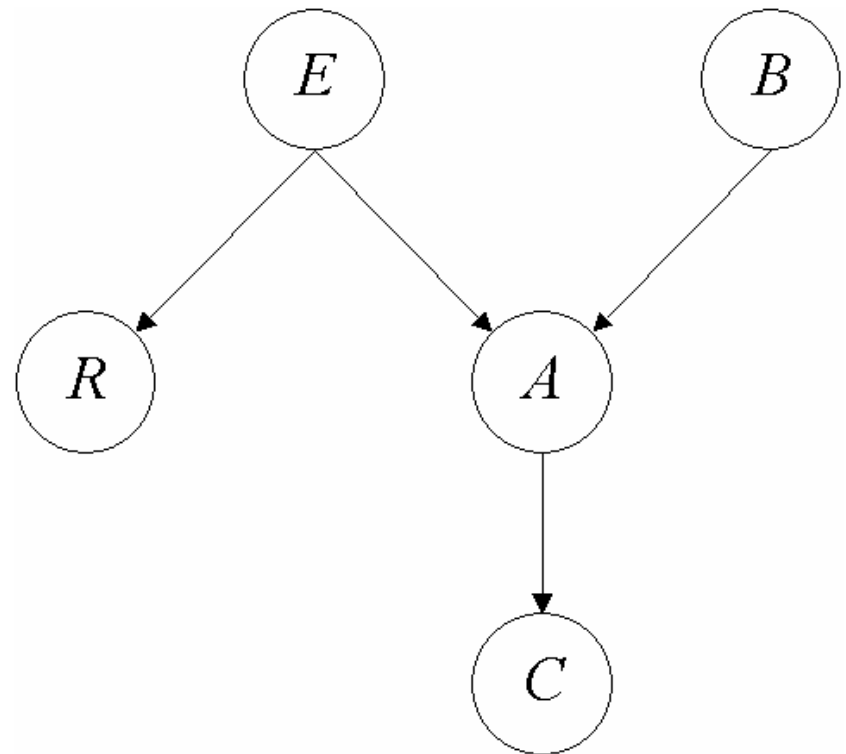
E : an earthquake happened

B : a burglar was in the house

A : the alarm is ringing

C : Fred receives a phone call reporting the alarm

R : Fred hears the radio report



$E \perp B$, i.e. $p(E, B) = p(E)p(B)$

~~$R \perp A$~~ but $R \perp A | E$

i.e. $p(R, A) \neq p(R)p(A)$ but $p(R, A | E) = p(R | E)p(A | E)$

Inference

Example: burglar, alarm and earthquake

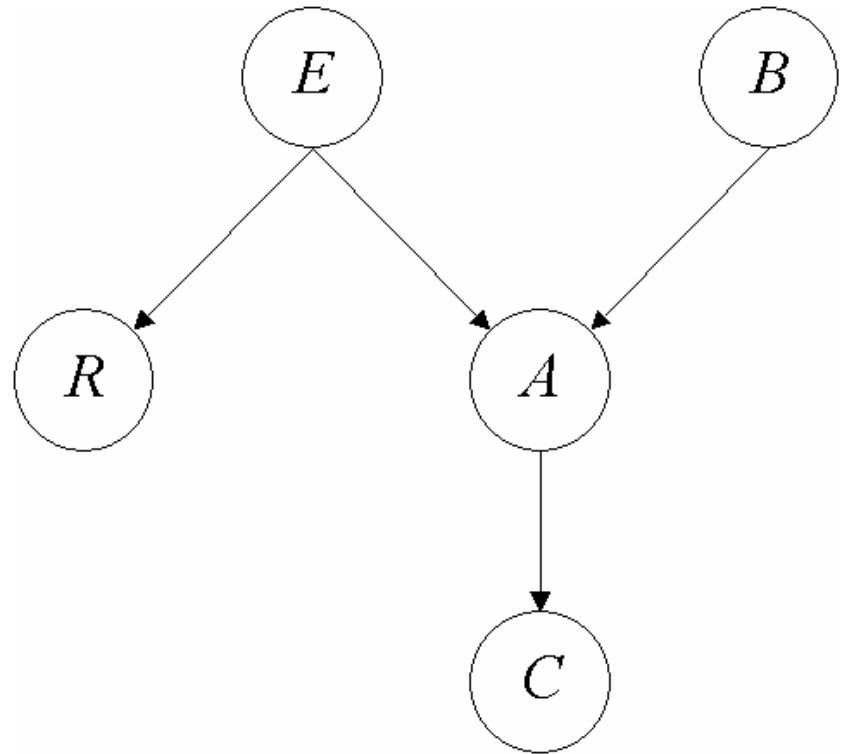
E : an earthquake happened

B : a burglar was in the house

A : the alarm is ringing

C : Fred receives a phone call
reporting the alarm

R : Fred hears the radio report



$$p(E, B, R, A, C) = p(E)p(B)p(R | E)p(A | E, B)p(C | A)$$

Inference

Example: burglar, alarm and earthquake

- while at work, Fred receives a phone call from his neighbour saying that his alarm is ringing. What is the probability that there was a burglar in his house?
- on his way home, Fred hears on the radio that there was an earthquake that day near his house. Relieved, he thinks the alarm was probably set off by the earthquake. What is the probability that there was a burglar in his house?

Inference

Fred received a phone call saying that his alarm was ringing. What is the probability there was a burglar?

$$p(B = 1 | C = 1) = \frac{p(B = 1, C = 1)}{p(C = 1)} = \frac{\sum_{E, R, A} p(E, B = 1, R, A, C = 1)}{\sum_{E, B, R, A} p(E, B, R, A, C = 1)}$$

$$p(B = 1 | C = 1) = \frac{\sum_{E, R, A} p(E) p(B = 1) p(\cancel{R | E}) p(A | E, B = 1) p(C = 1 | A)}{\sum_{E, B, R, A} p(E) p(B) p(\cancel{R | E}) p(A | E, B) p(C = 1 | A)}$$

Inference

Fred is now listening to the radio report of the earth-quake.
What is the probability there was a burglar?

$$p(B = 1 | C = 1, R = 1) = \frac{p(B = 1, R = 1, C = 1)}{p(R = 1, C = 1)} = \frac{\sum_{E,A} p(E, B = 1, R = 1, A, C = 1)}{\sum_{E,B,A} p(E, B, R = 1, A, C = 1)}$$

$$p(B = 1 | C = 1, R = 1) = \frac{\sum_{E,A} p(E)p(B = 1)p(R = 1 | E)p(A | E, B = 1)p(C = 1 | A)}{\sum_{E,B,A} p(E)p(B)p(R = 1 | E)p(A | E, B)p(C = 1 | A)}$$

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- *Structural learning*
- Known structure
 - ex: linear regression
 - ex: Gaussian mixture models
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Structural learning

- Every network fits the data differently
 - => use it as a score and pick the best network
- Pb of maximum likelihood: the more edges the better
- We need a score that penalises the number of edges
- BIC criterion: function of
 - the likelihood
 - the number of parameters

Structural learning

- How to explore the space?
- Go from one network to the next using simple alterations
- Add edge, remove edge, swap nodes
- The distributions must be relearnt everytime
usually assumes that the data is discretised

Structural learning

- In biology: gene networks or protein networks or gene-protein networks
- Many genes (many nodes)
- Few conditions (few samples)
- Bad...

Plan

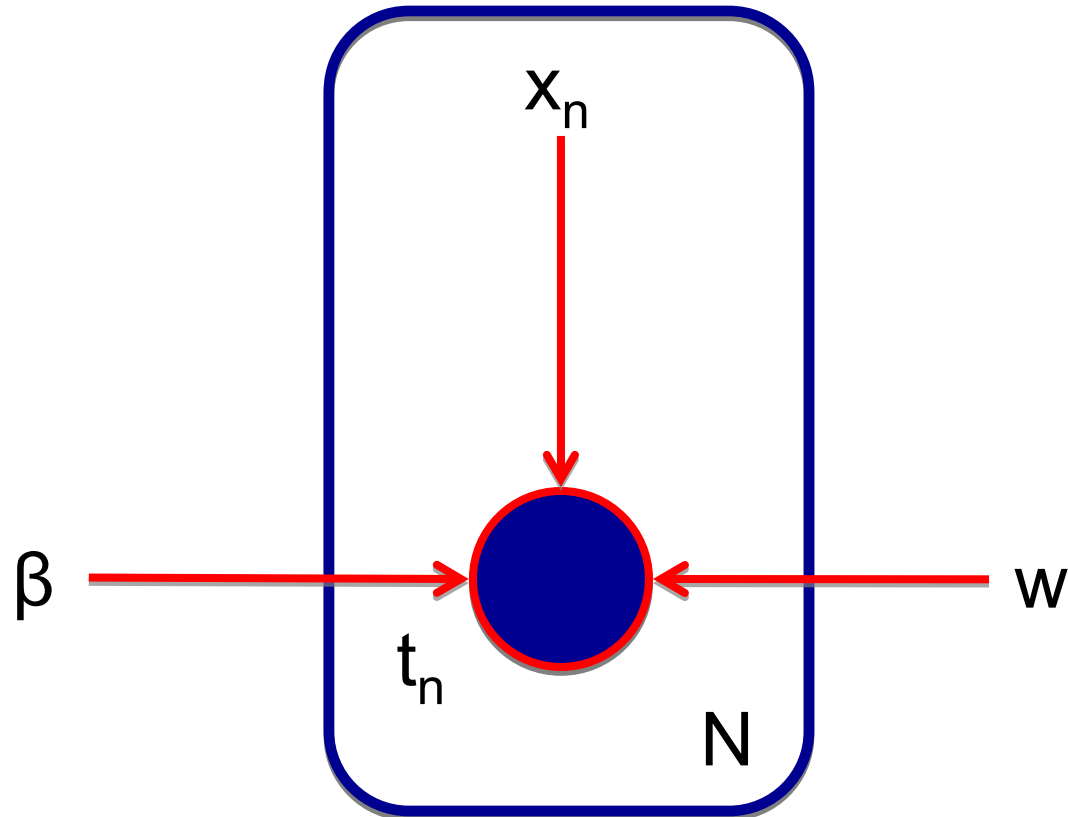
- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- *Structural learning*
- *Known structure*
 - *ex: linear regression*
 - *ex: Gaussian mixture models*
- Bayesian networks versus Bayesian models
- *ex: Bayesian linear regression*

Ex: linear regression

$$p(\mathbf{t} | \mathbf{X}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$

Ex: linear regression

$$p(\mathbf{t} | \mathbf{X}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$



Ex: linear regression

- Maximum likelihood training

$$p(\mathbf{t} | \mathbf{X}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$

$$L(\mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n))^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)$$

- For \mathbf{w}

$$L(\mathbf{w}) \propto -\frac{1}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n))^2 + \text{cst} \quad \text{so that } \mathbf{w}_{\text{ML}} = \mathbf{w}_{\text{LS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t}$$

- For β

$$\beta_{\text{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^N (t_n - y(\mathbf{x}_n))^2$$

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- *Structural learning*
- *Known structure*
 - *ex: linear regression*
 - *ex: Gaussian mixture models*
- Bayesian networks versus Bayesian models
- ex: Bayesian linear regression

Ex: Gaussian mixture model

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$$

$$p(\mathbf{X}) = \prod_{n=1}^N \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right)$$

$$L(\pi, \mu, \Sigma) = \log p(\mathcal{D}) = \log p(\mathbf{X}) = \sum_{n=1}^N \log \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right)$$

intractable...

Ex: Gaussian mixture model

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$$

very hard to learn, so add a latent variable

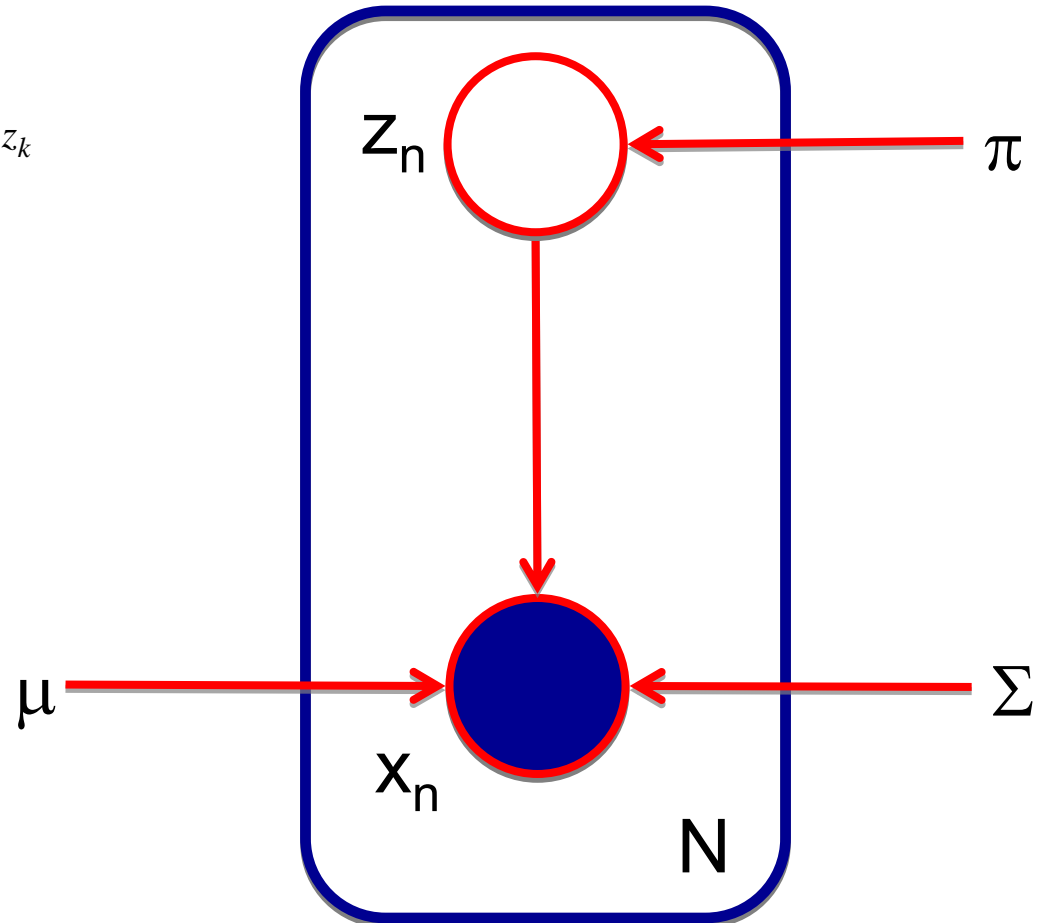
$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k} \quad \text{and} \quad p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)^{z_k}$$

$$\begin{aligned} p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} | \mathbf{z}) \\ &= \sum_{k=1}^K p(z_k) p(\mathbf{x} | z_k) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \end{aligned}$$

Ex: Gaussian mixture model

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

$$p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)^{z_k}$$



Ex: Gaussian mixture model

The EM algorithm

– inference: compute $p(H|O, \theta^{(t)})$

– E-step (expectation): compute

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{H|O, \theta^{(t)}} [\log p(H, O | \theta)]$$

– M-step (maximisation): compute

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)})$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– inference: compute $p(\mathbf{Z}|\mathbf{X}, \theta^{(t)})$

– E-step (expectation): compute

$$Q(\theta|\theta^{(t)}) = E_{\mathbf{Z}|\mathbf{X}, \theta^{(t)}} [\log p(\mathbf{X}, \mathbf{Z} | \theta)]$$

– M-step (maximisation): compute

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)})$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– inference: compute $p(\mathbf{Z}|\mathbf{X}, \theta^{(t)}) = \prod_{n=1}^N p(\mathbf{z}_n | \mathbf{x}_n, \theta^{(t)})$

$$p(\mathbf{z}_n | \mathbf{x}_n, \theta^{(t)}) = \prod_{k=1}^K p(z_{nk} | \mathbf{x}_n, \theta^{(t)})^{z_{nk}}$$

$$p(z_{nk} | \mathbf{x}_n, \theta^{(t)}) = \frac{p(\mathbf{x}_n, z_{nk} | \theta^{(t)})}{p(\mathbf{x}_n | \theta^{(t)})} = \frac{p(z_{nk} | \pi^{(t)}) p(\mathbf{x}_n | z_{nk}, \mu^{(t)}, \Sigma^{(t)})}{\sum_{v=1}^K \pi_v^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_v^{(t)}, \Sigma_v^{(t)})} = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{v=1}^K \pi_v^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_v^{(t)}, \Sigma_v^{(t)})}$$

$$\gamma_{nk} = \mathbf{E}[z_{nk}] = p(z_{nk} | \mathbf{x}_n, \theta^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{v=1}^K \pi_v^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_v^{(t)}, \Sigma_v^{(t)})}$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– E-step (expectation): compute

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} [\log p(\mathbf{X}, \mathbf{Z} | \theta)] = \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} \left[\sum_{n=1}^N \log (p(\mathbf{z}_n | \pi) p(\mathbf{x}_n | \mathbf{z}_n, \mu, \Sigma)) \right] \\ &= \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} \left[\sum_{n=1}^N (\log p(\mathbf{z}_n | \pi) + \log p(\mathbf{x}_n | \mathbf{z}_n, \mu, \Sigma)) \right] \\ &= \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} \left[\sum_{n=1}^N \sum_{k=1}^K (z_{nk} \log \pi_k + z_{nk} \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)) \right] \\ &= \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} \left[z_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)) \right] \\ &= \sum_{n=1}^N \sum_{k=1}^K \mathbb{E}_{\mathbf{z} | \mathbf{X}, \theta^{(t)}} [z_{nk}] (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)) \end{aligned}$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

- M-step (maximisation): compute

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta | \theta^{(t)})$$

$$Q(\theta | \theta^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k))$$

$$\begin{aligned} \frac{\partial Q(\theta | \theta^{(t)})}{\partial \pi_k} &= \frac{\partial}{\partial \pi_k} \left(\sum_{n=1}^N \gamma_{nk} \log \pi_k + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \right) \\ &= \sum_{n=1}^N \gamma_{nk} \frac{\partial}{\partial \pi_k} \log \pi_k + \lambda \frac{\partial}{\partial \pi_k} \pi_k = \sum_{n=1}^N \frac{\gamma_{nk}}{\pi_k} + \lambda \\ &= 0 \Leftrightarrow \sum_{n=1}^N \frac{\gamma_{nk}}{\pi_k^{(t+1)}} + \lambda = 0 \Leftrightarrow \pi_k^{(t+1)} = -\frac{1}{\lambda} \sum_{n=1}^N \gamma_{nk} \end{aligned}$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– M-step (maximisation): compute

$$\sum_{k=1}^K \pi_k^{(t+1)} = 1 \Leftrightarrow \sum_{k=1}^K -\frac{1}{\lambda} \sum_{n=1}^N \gamma_{nk} = 1 \Leftrightarrow -\frac{1}{\lambda} \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} = 1 \Leftrightarrow -\frac{1}{\lambda} \sum_{n=1}^N 1 = 1 \Leftrightarrow \lambda = -N$$

$$\frac{\partial Q(\theta | \theta^{(t)})}{\partial \pi_k} = 0 \Leftrightarrow \pi_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^N \gamma_{nk}$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– M-step (maximisation): compute

$$\begin{aligned}\frac{\partial Q(\theta|\theta^{(t)})}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \left(\sum_{n=1}^N \gamma_{nk} \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right) \\ &= \sum_{n=1}^N \gamma_{nk} \frac{\partial}{\partial \mu_k} \left(-\frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \right) = \frac{1}{2} \sum_{n=1}^N \gamma_{nk} \Sigma_k^{-1} (\mathbf{x}_n - \mu_k)\end{aligned}$$

$$\begin{aligned}\frac{\partial Q(\theta|\theta^{(t)})}{\partial \mu_k} = 0 &\Leftrightarrow \sum_{n=1}^N \gamma_{nk} \Sigma_k^{-1} \mu_k^{(t+1)} = \sum_{n=1}^N \gamma_{nk} \Sigma_k^{-1} \mathbf{x}_n \\ &= 0 \Leftrightarrow \mu_k^{(t+1)} = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \left(\sum_{n=1}^N \gamma_{nk} \mathbf{x}_n \right)\end{aligned}$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

– M-step (maximisation): compute

$$\begin{aligned}\frac{\partial Q(\theta|\theta^{(t)})}{\partial \Sigma_k^{-1}} &= \frac{\partial}{\partial \Sigma_k} \left(\sum_{n=1}^N \gamma_{nk} \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right) \\ &= \sum_{n=1}^N \gamma_{nk} \frac{\partial}{\partial \Sigma_k} \left(\frac{1}{2} \log |\Sigma_k^{-1}| - \frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \right) \\ &= \frac{1}{2} \sum_{n=1}^N \gamma_{nk} \left(\Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \right)\end{aligned}$$

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial \Sigma_k^{-1}} = 0 \Leftrightarrow \Sigma_k^{(t+1)} = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \left(\sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \right)$$

Ex: Gaussian mixture model

The EM algorithm applied to GMMs

- inference: $\gamma_{nk} = \mathbb{E}[z_{nk}] = p(z_{nk} | \mathbf{x}_n, \theta^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{v=1}^K \pi_v^{(t)} \mathcal{N}(\mathbf{x}_n | \mu_v^{(t)}, \Sigma_v^{(t)})}$
- E-step (expectation): $Q(\theta | \theta^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} (\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k))$
- M-step (maximisation):
$$\frac{\partial Q(\theta | \theta^{(t)})}{\partial \theta} = 0 \Leftrightarrow \begin{cases} \pi_k^{(t+1)} = \frac{1}{N} \sum_{n=1}^N \gamma_{nk} & \text{and } \mu_k^{(t+1)} = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \left(\sum_{n=1}^N \gamma_{nk} \mathbf{x}_n \right) \\ \Sigma_k^{(t+1)} = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \left(\sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \mu_k) (\mathbf{x}_n - \mu_k)^T \right) \end{cases}$$

Gaussian mixture models

- Choose a number of clusters K
- Initialise the K priors π_k , the K means μ_k and the K covariances Σ_k
- Repeat until convergence
 - compute the probability $p(k | \mathbf{x}_n)$ of each datapoint \mathbf{x}_n to belong to each cluster k
 - update each cluster prior π_k , mean μ_k and covariance Σ_k by taking the weighted average number / location / variance of all the points, where the weight of point \mathbf{x}_n is $p(k | \mathbf{x}_n)$

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- *Structural learning*
- *Known structure*
 - *ex: linear regression*
 - *ex: Gaussian mixture models*
- *Bayesian networks versus Bayesian models*
- *ex: Bayesian linear regression*

Bayesian networks versus Bayesian models

A Bayesian network

- is called so because it uses Bayes theorem
- does not make the model Bayesian
- what does it mean to be Bayesian then?

Bayesian networks versus Bayesian models

A Bayesian model

- treats its parameters as hidden random variables
- these parameters have distributions of their own, called prior distributions
- “learning the parameters” is replaced by “inferring their posterior distribution”
- predictions are averaged over each possible parameter

Bayesian thinking

- $p(\theta)$ is the prior probability of the parameters, before we can observe the data
- $p(\theta | \mathcal{D})$ is the posterior probability of the parameters, after we have observed the data
- Predictions will take into account every possible θ , weighted by its posterior probability

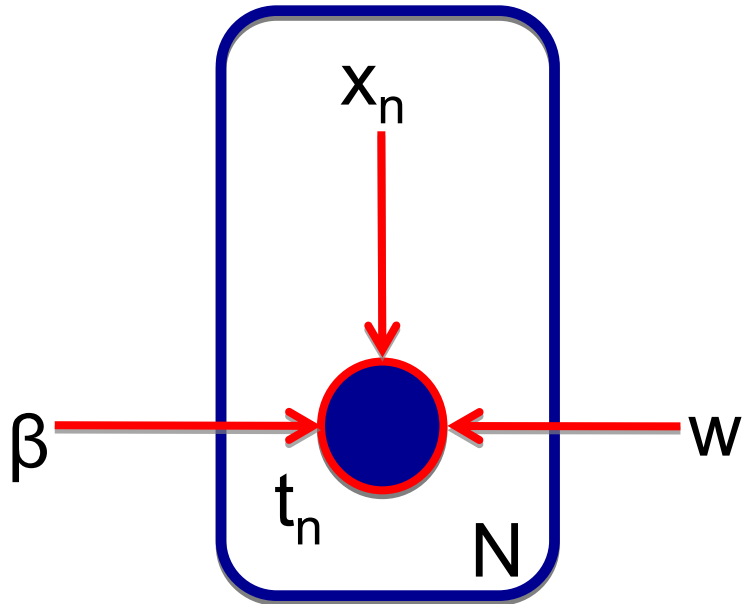
$$p(\hat{\mathbf{x}} | \mathcal{D}) = \int p(\hat{\mathbf{x}} | \theta) p(\theta | \mathcal{D}) d\theta$$

Plan

- *Probability theory*
- *Graphical models*
- *Bayesian networks*
- *Inference*
- *Structural learning*
- *Known structure*
 - *ex: linear regression*
 - *ex: Gaussian mixture models*
- *Bayesian networks versus Bayesian models*
- *ex: Bayesian linear regression*

Ex: Bayesian linear regression

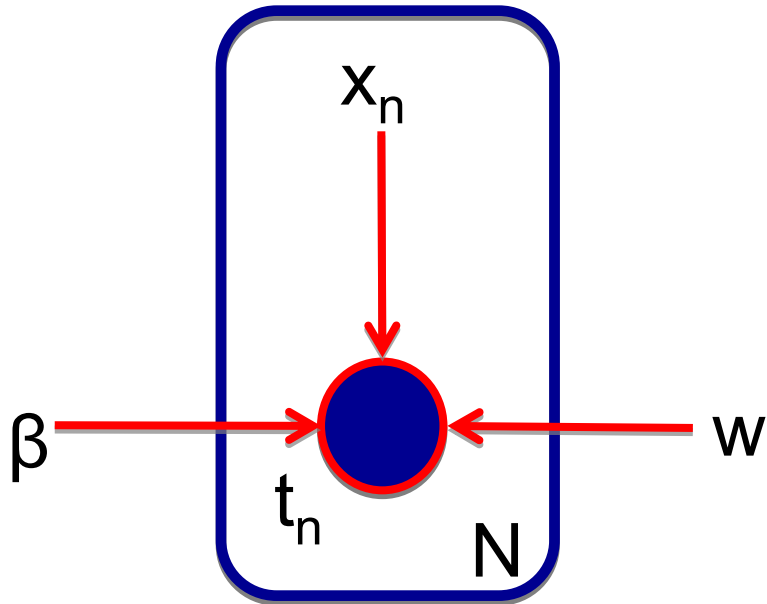
Non-Bayesian



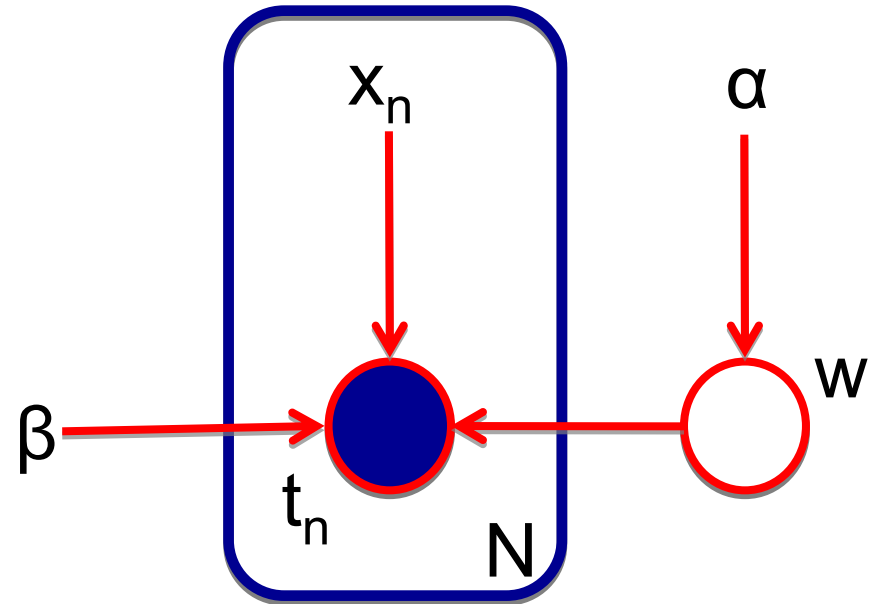
Bayesian

Ex: Bayesian linear regression

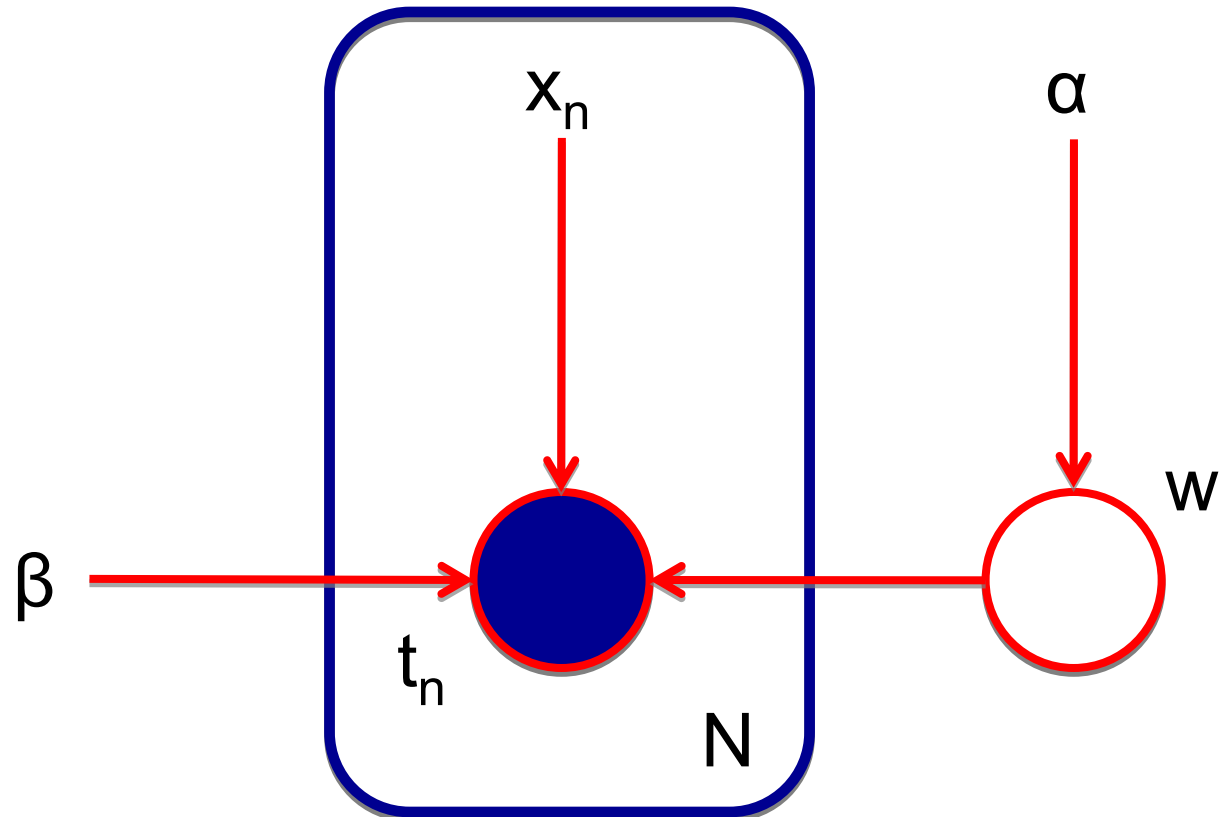
Non-Bayesian



Bayesian

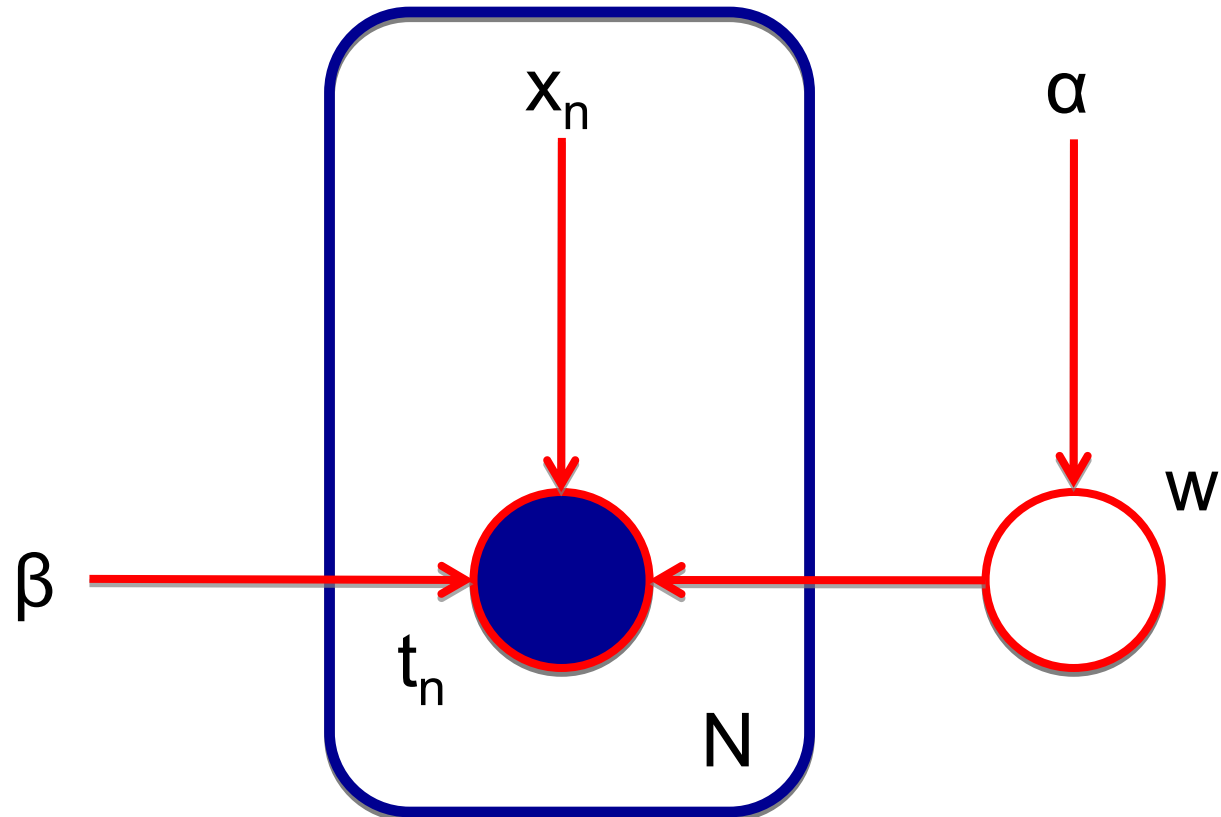


Ex: Bayesian linear regression



Ex: Bayesian linear regression

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$

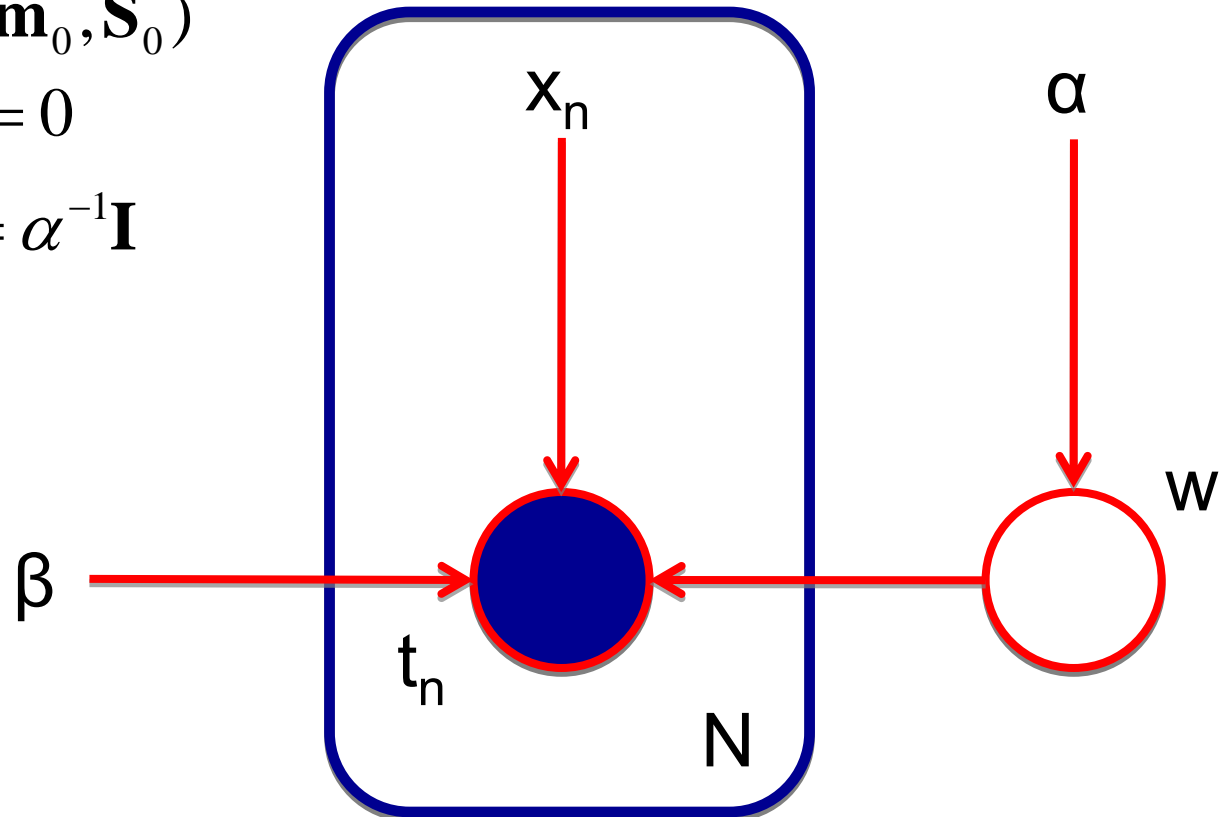


Ex: Bayesian linear regression

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

$$\text{with } \begin{cases} \mathbf{m}_0 = \mathbf{0} \\ \mathbf{S}_0 = \alpha^{-1} \mathbf{I} \end{cases}$$



Ex: Bayesian linear regression

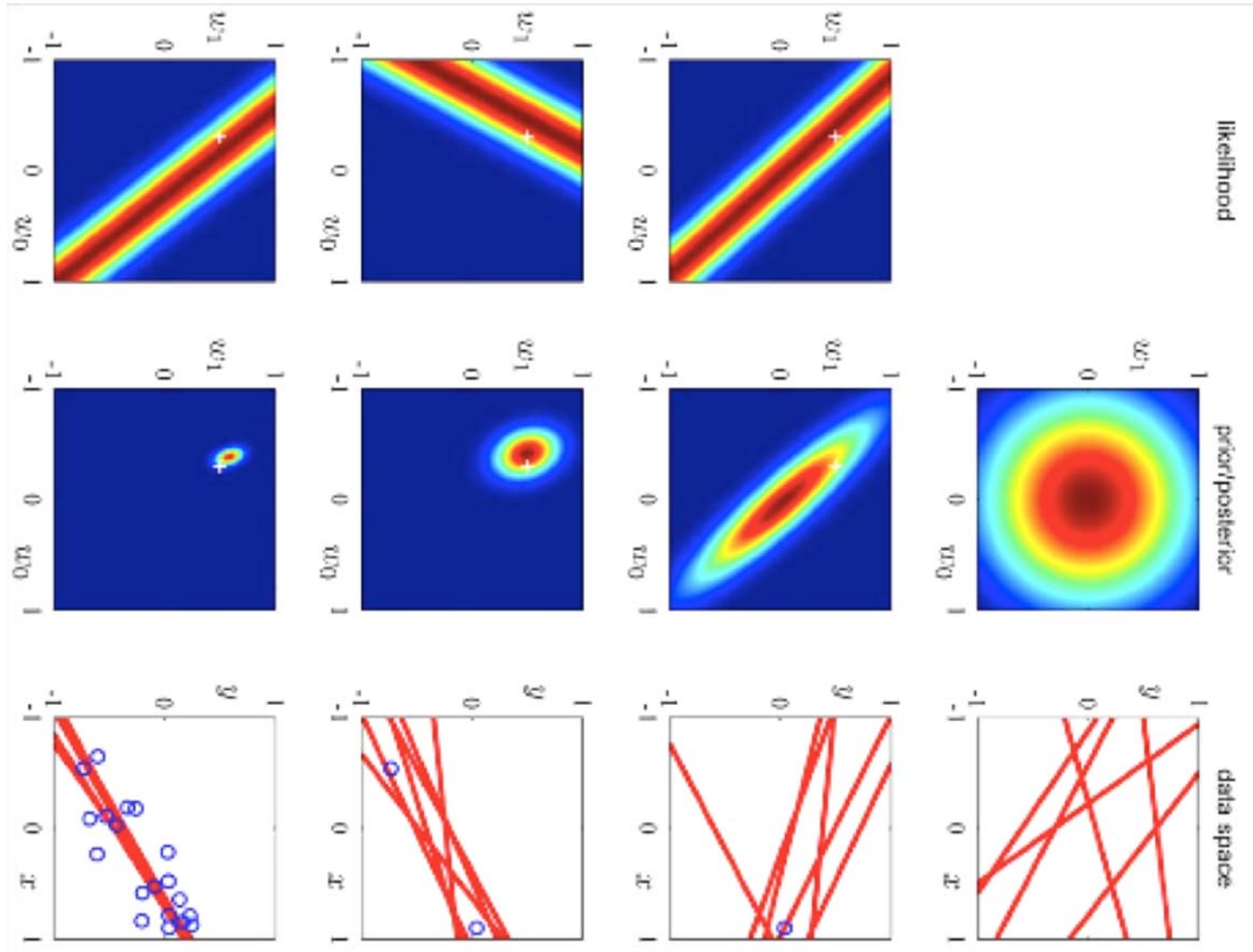
$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | y(\mathbf{x}_n), \beta^{-1})$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | 0, \alpha^{-1} \mathbf{I})$$

$$p(\mathbf{w} | \mathbf{X}, \mathbf{t}) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

$$\log p(\mathbf{w} | \mathbf{X}, \mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \mathbf{x}_n)^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{cst}$$

Ex: Bayesian linear regression



Ex: Bayesian linear regression

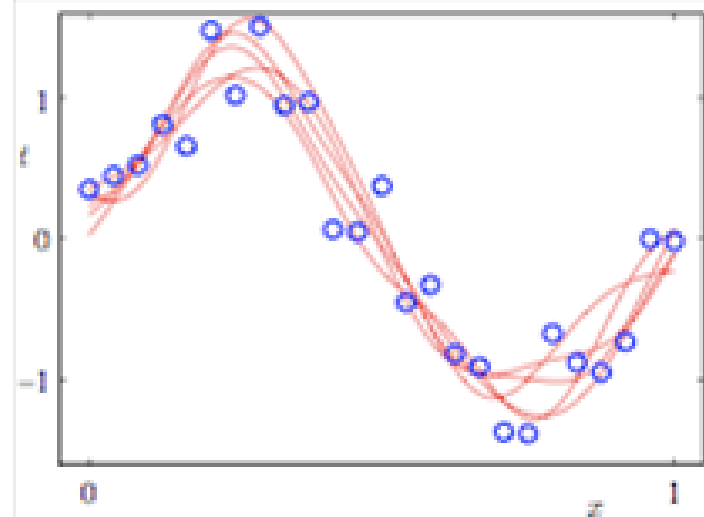
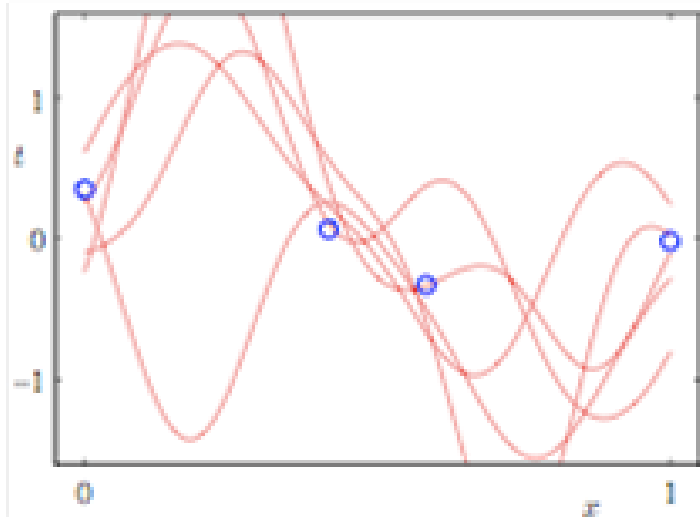
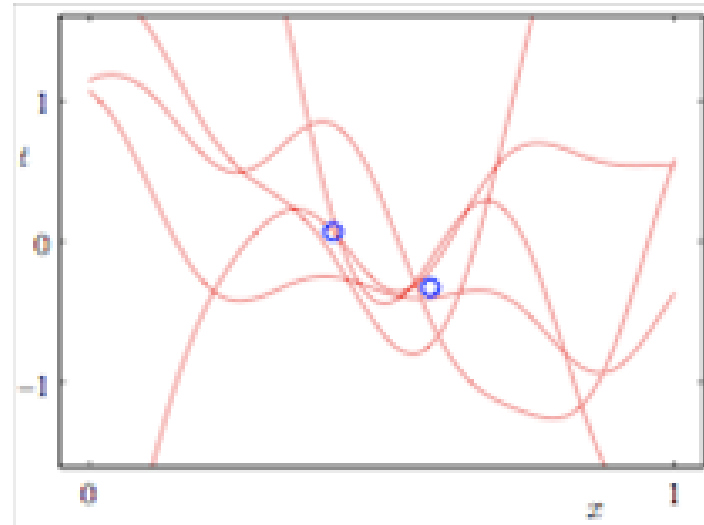
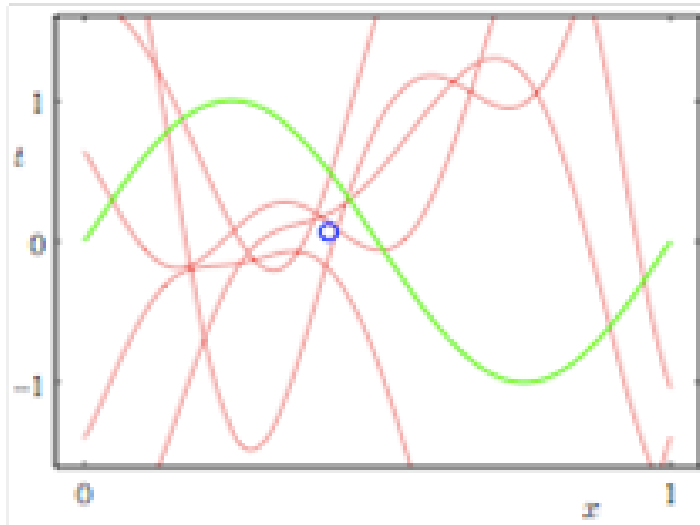
$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)$$

$$\text{with } \begin{cases} \mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{X}^T \mathbf{t}) = \beta \mathbf{S}_N \mathbf{X}^T \mathbf{t} \\ \mathbf{S}_N = (\mathbf{S}_0^{-1} + \beta \mathbf{X}^T \mathbf{X})^{-1} = (\alpha \mathbf{I} + \beta \mathbf{X}^T \mathbf{X})^{-1} \end{cases}$$

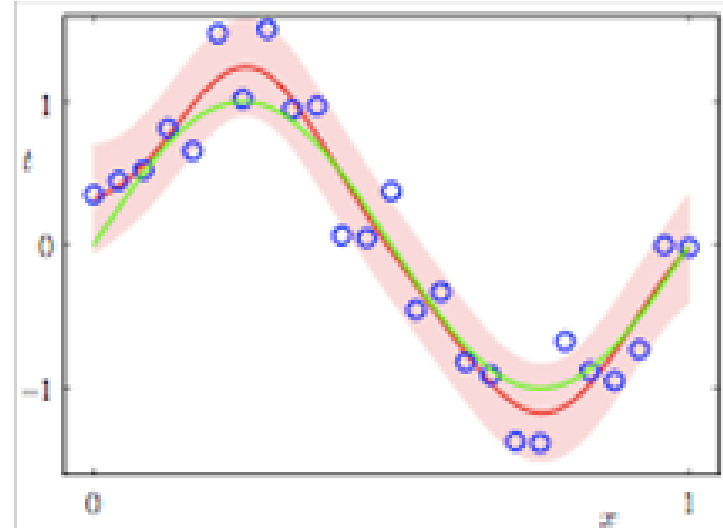
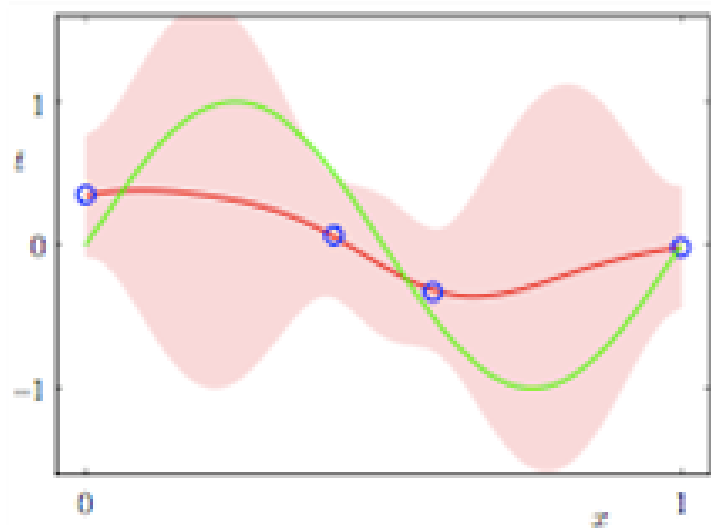
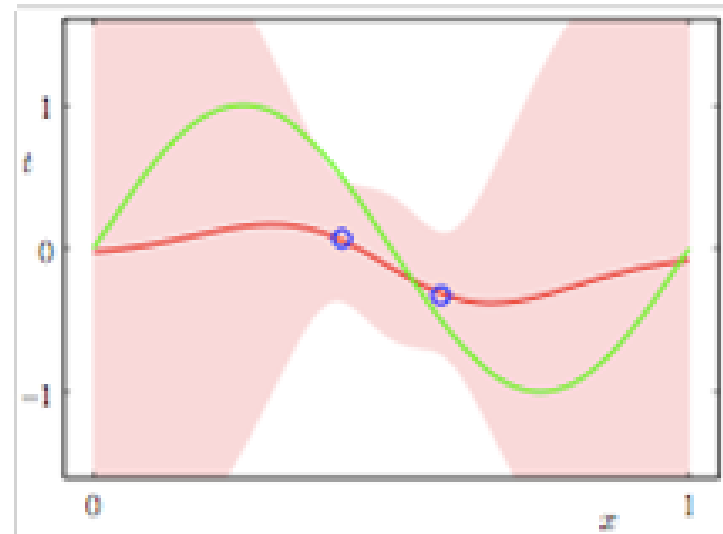
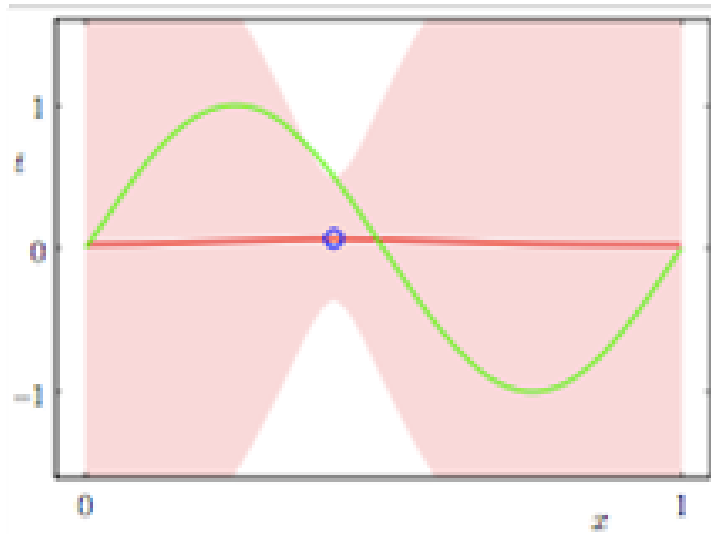
predictive distribution

$$\begin{aligned} p(\hat{t} \mid \hat{\mathbf{x}}, \mathbf{X}, \mathbf{t}) &= \int p(\hat{t} \mid \hat{\mathbf{x}}, \mathbf{w}) p(\mathbf{w} \mid \mathbf{X}, \mathbf{t}) d\mathbf{w} \\ &= \mathcal{N}\left(\hat{t} \mid \mathbf{m}_N^T \hat{\mathbf{x}}, \underbrace{\beta^{-1}}_{\text{noise}} + \underbrace{\hat{\mathbf{x}}^T \mathbf{S}_N \hat{\mathbf{x}}}_{\text{uncertainty in } \mathbf{w}}\right) \end{aligned}$$

Ex: Bayesian linear regression



Ex: Bayesian linear regression



Bayesian models

Plan

- Bayesian learning = inference
- Predictions
- Bayesian thinking
- Priors
- Bayesian model selection

Bayesian models

What is a probability?

- classical statistics: the limiting frequency when an experiment is repeated infinitely many times
- everyday language: how strongly one believes in something
 - what is the probability that it rains tomorrow? What are the chances that Alice goes to Bob's party?
 - ask 2 different people... do they agree?
 - Bayesian statistics ~ everyday language

Plan

- *Bayesian learning = inference*
- Predictions
- Bayesian thinking
- Priors
- Bayesian model selection

Bayesian learning = inference

Find the posterior distribution over the set of parameters θ

$$p(\theta \mid \mathcal{D}, \mathcal{M})$$

training data

parameters model

Bayesian learning = inference

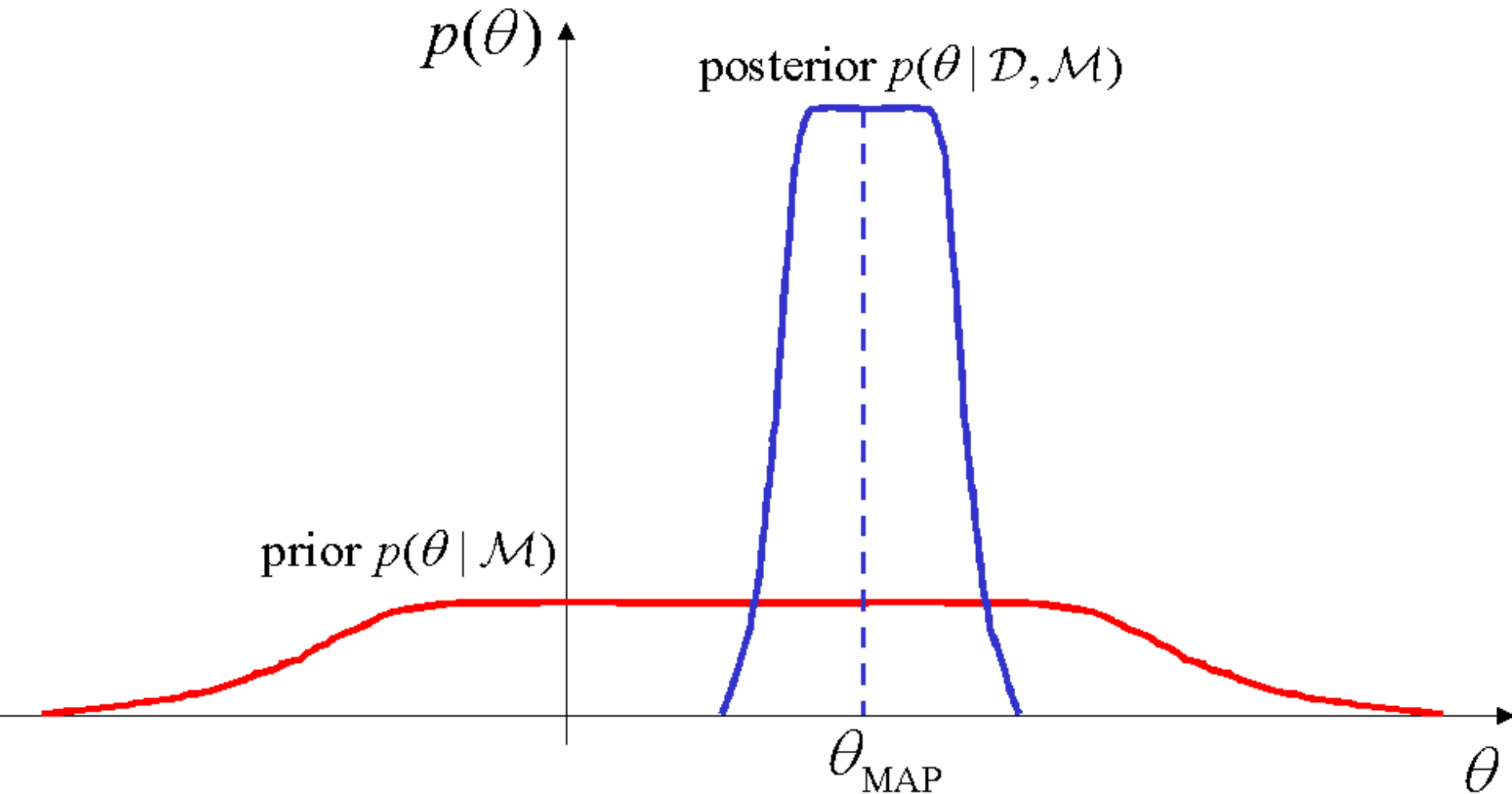
Bayes theorem

$$p(\theta | \mathcal{D}, \mathcal{M}) = \frac{\overset{\text{joint}}{p(\mathcal{D}, \theta | \mathcal{M})}}{\underset{\text{model evidence}}{p(\mathcal{D} | \mathcal{M})}} = \frac{\overset{\text{likelihood}}{p(\mathcal{D} | \theta, \mathcal{M})} \overset{\text{prior}}{p(\theta | \mathcal{M})}}{\underset{\text{marginal likelihood}}{p(\mathcal{D} | \mathcal{M})}}$$

model evidence

$$p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\theta$$

Bayesian learning = inference



Bayesian learning = inference

- inference

$$p(\theta | \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M})}{p(\mathcal{D} | \mathcal{M})}$$

- maximum a posteriori

$$\begin{aligned}\theta_{\text{MAP}} &= \arg \max_{\theta} p(\theta | \mathcal{D}, \mathcal{M}) \\ &= \arg \max_{\theta} p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M})\end{aligned}$$

- maximum likelihood

$$\theta_{\text{ML}} = \arg \max_{\theta} p(\mathcal{D} | \theta, \mathcal{M})$$

Plan

- *Bayesian learning = inference*
- *Predictions*
- Bayesian thinking
- Priors
- Bayesian model selection

Predictions

- Bayesian decision

$$p(\hat{t}|\hat{\mathbf{x}}, \mathcal{D}, \mathcal{M}) = \int p(\hat{t}|\hat{\mathbf{x}}, \theta, \mathcal{M}) p(\theta | \mathcal{D}, \mathcal{M}) d\theta$$

- Point estimate θ^*

- data plentiful: $p(\theta | \mathcal{D}, \mathcal{M}) \approx \delta_{\theta, \theta^*}$ with $\theta^* = \theta_{\text{MAP}}$
- data really plentiful: $p(\theta | \mathcal{M})$ has no influence $\Rightarrow \theta^* \approx \theta_{\text{ML}}$
- decision:
$$p(\hat{t}|\hat{\mathbf{x}}, \mathcal{D}, \mathcal{M}) \approx \int p(\hat{t}|\hat{\mathbf{x}}, \theta, \mathcal{M}) \delta_{\theta, \theta^*} d\theta$$
$$= p(\hat{t}|\hat{\mathbf{x}}, \theta^*, \mathcal{M})$$

Plan

- *Bayesian learning = inference*
- *Predictions*
- *Bayesian thinking*
- Priors
- Bayesian model selection

Bayesian thinking

Bayesian view

- prior probabilities are subjective
- there is no absolute probability since background assumptions may vary

Frequentist view

- prior what?
- an absolute probability is assigned by repeating experiments many many many times

Bayesian thinking

Bayesian view

- $p(\mathcal{D} | \theta, \mathcal{M})$
- θ : random variable
 - only one dataset D
 - the uncertainty in the parameters is expressed through a distribution over θ

Frequentist view

- $p(\mathcal{D} | \theta, \mathcal{M})$
- θ : estimator
 - determined according to D
 - with error bars coming from its distribution

Bayesian thinking

Bayesians

- use Bayes theorem

$$p(\theta | \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} | \theta, \mathcal{M})p(\theta | \mathcal{M})}{p(\mathcal{D} | \mathcal{M})}$$

Frequentists

- use estimators, ex:
 - value: maximum likelihood
 - error bars: cross-validation

Bayesian thinking

- $p(\theta | \mathcal{M})$ is the prior probability of the parameters, before we can observe the data
- $p(\theta | \mathcal{D}, \mathcal{M})$ is the posterior probability of the parameters, after we have observed the data

Bayesian thinking

Bayesians

- use Bayes theorem

$$p(\theta | \mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D} | \theta, \mathcal{M})p(\theta | \mathcal{M})}{p(\mathcal{D} | \mathcal{M})}$$

- predict using

$$p(\hat{t} | \hat{\mathbf{x}}, \mathcal{D}, \mathcal{M}) = \int p(\hat{t} | \hat{\mathbf{x}}, \theta, \mathcal{M})p(\theta | \mathcal{D}, \mathcal{M}) d\theta$$

Frequentists

- use estimators, ex:

- value: maximum likelihood
- error bars: cross-validation

- predict using

$$p(\hat{t} | \hat{\mathbf{x}}, \mathcal{D}, \mathcal{M}) = p(\hat{t} | \hat{\mathbf{x}}, \theta^*, \mathcal{M})$$

Plan

- *Bayesian learning = inference*
- *Predictions*
- *Bayesian thinking*
- *Priors*
- Bayesian model selection

Priors

- Objective (non-informative) priors
 - have good frequentist properties but are hard to implement (ex: continuous variables)
- Subjective priors
 - should capture out beliefs as well as possible, and that's it!
- Conjugate priors

Priors

- Hierarchical priors

$p(\theta | \alpha)$? put a probability distribution $p(\alpha | \beta)$ on the hyperparameters... this can go on for several levels...

- Empirical priors

- $\alpha^* = \arg \max_{\alpha} p(\mathcal{D} | \alpha)$

- predictions: $p(\hat{\mathbf{x}} | \mathcal{D}, \alpha^*) = \int p(\hat{\mathbf{x}} | \theta) p(\theta | \mathcal{D}, \alpha^*) d\theta$

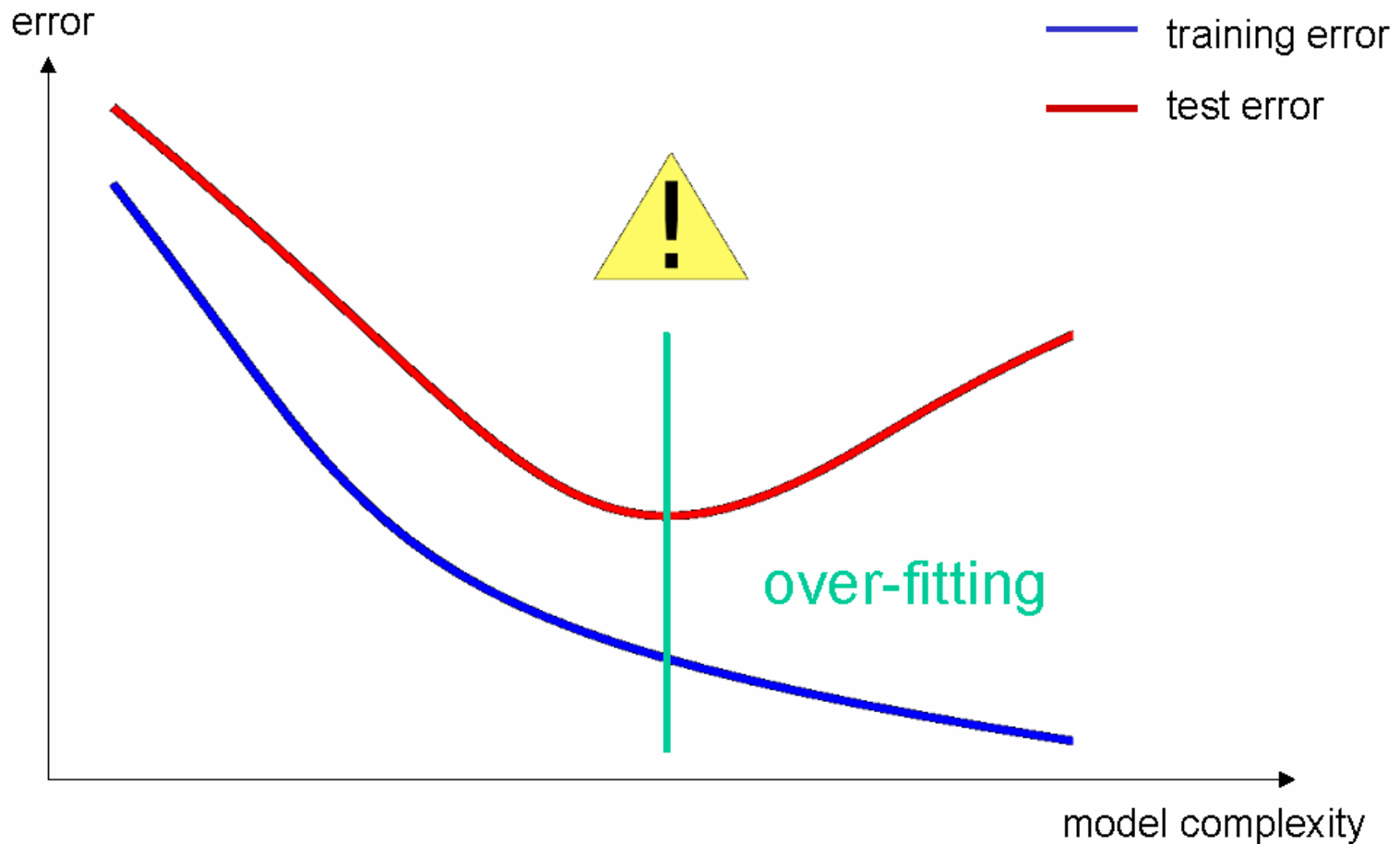
- more robust but double counting of evidence and prone to overfitting

Plan

- *Bayesian learning = inference*
- *Predictions*
- *Bayesian thinking*
- *Priors*
- *Bayesian model selection*

Bayesian model selection

Remember this plot



Bayesian model selection

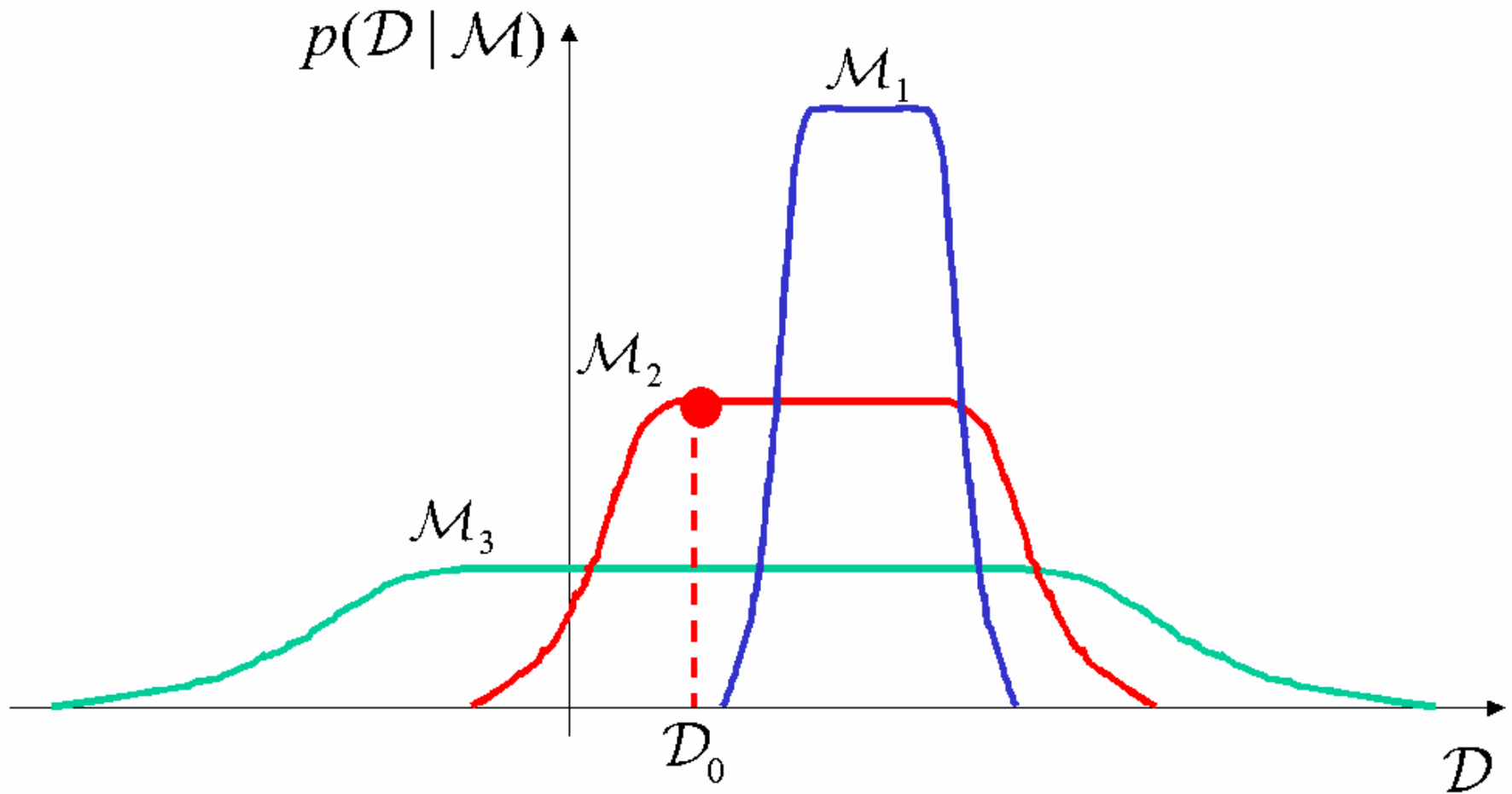
- Q: how do I tell which model generalizes best?
- Bayesian model selection

$$p(\mathcal{M} | \mathcal{D}) = \frac{p(\mathcal{D} | \mathcal{M}) p(\mathcal{M})}{p(\mathcal{D})} \propto p(\mathcal{M}) p(\mathcal{D} | \mathcal{M})$$

$$p(\mathcal{M} | \mathcal{D}) \propto p(\mathcal{M}) \int p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\theta$$

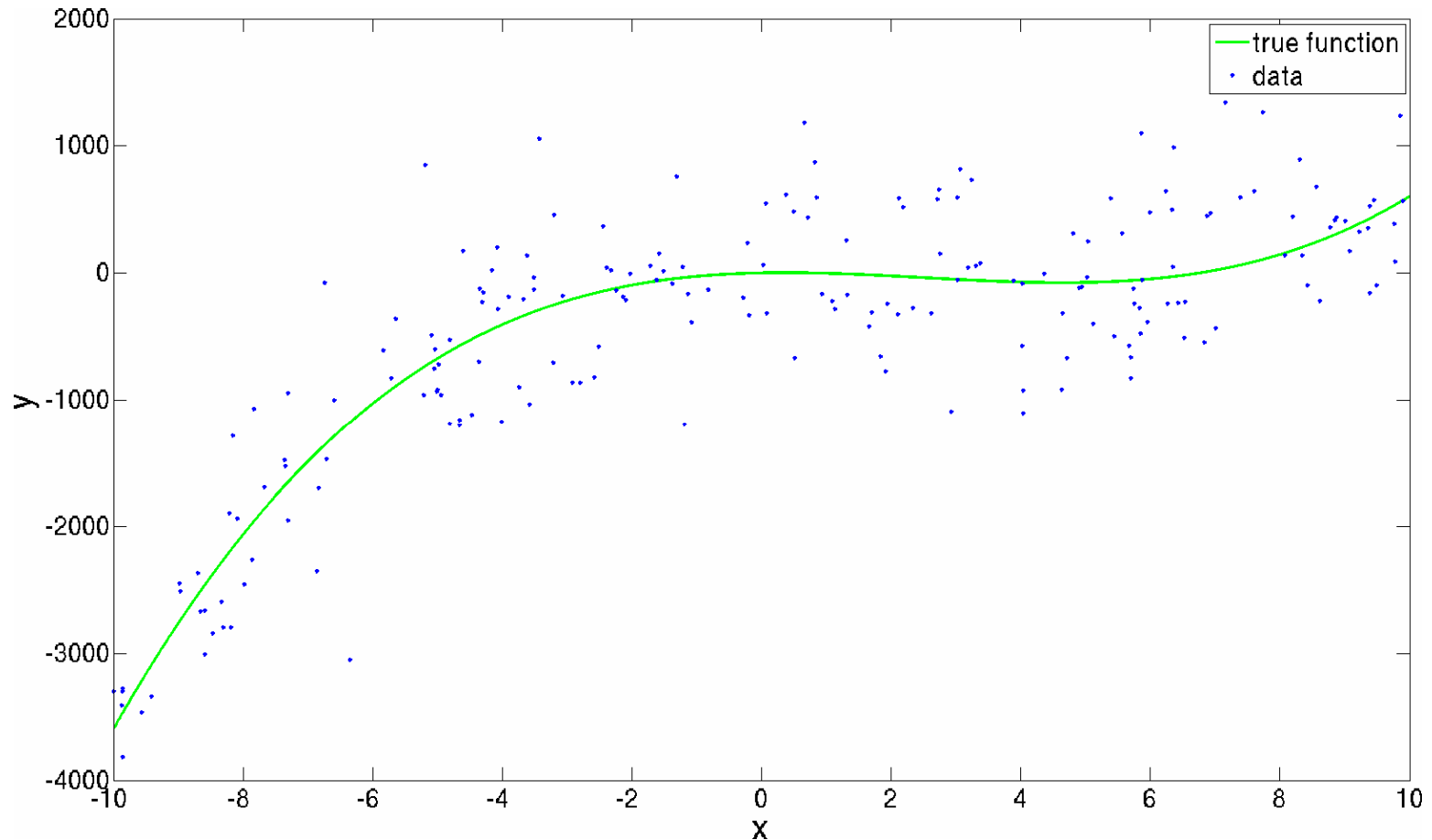
=> all on training data

Bayesian model selection



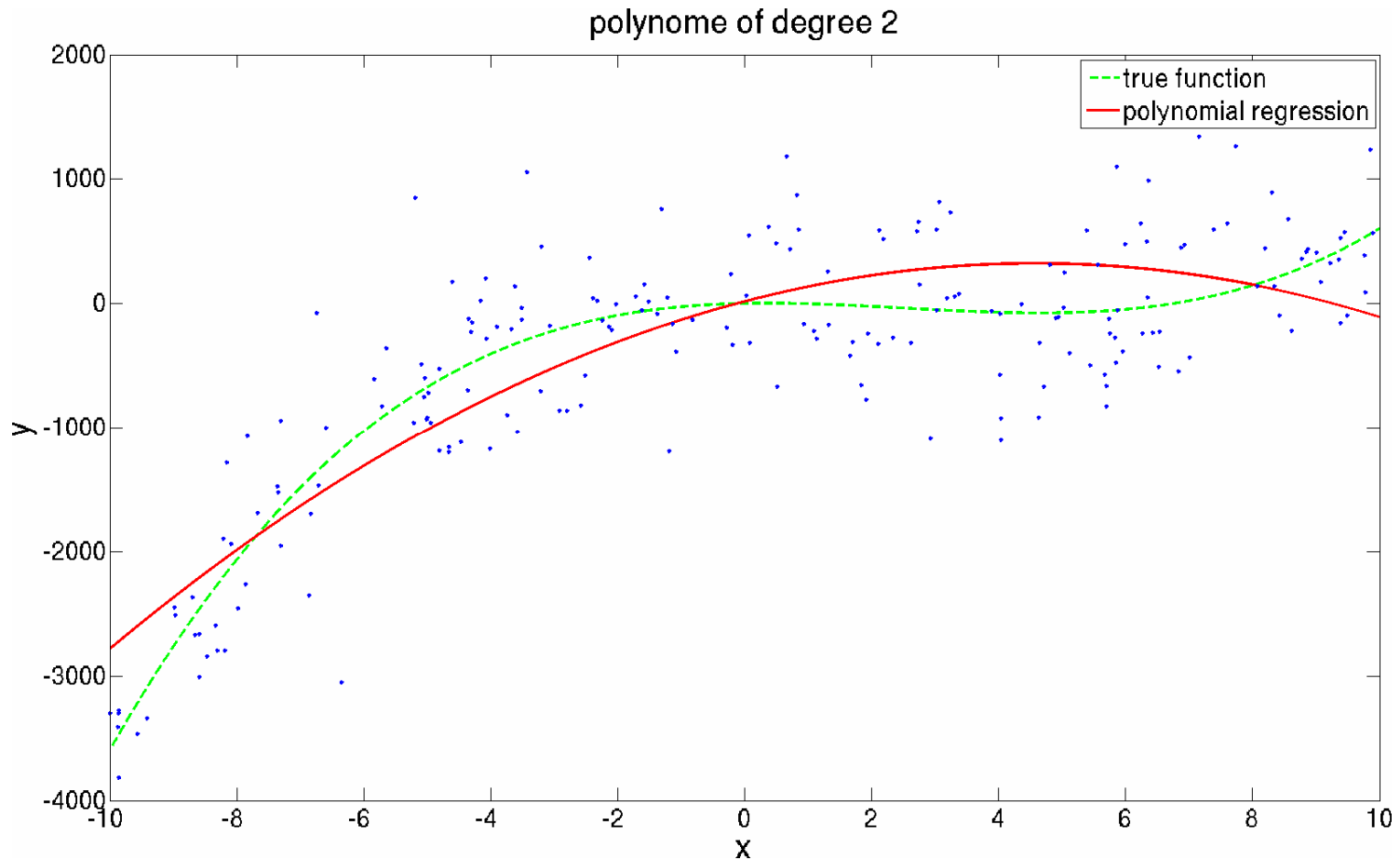
Bayesian model selection

Polynomial regression – true function (d. 3)



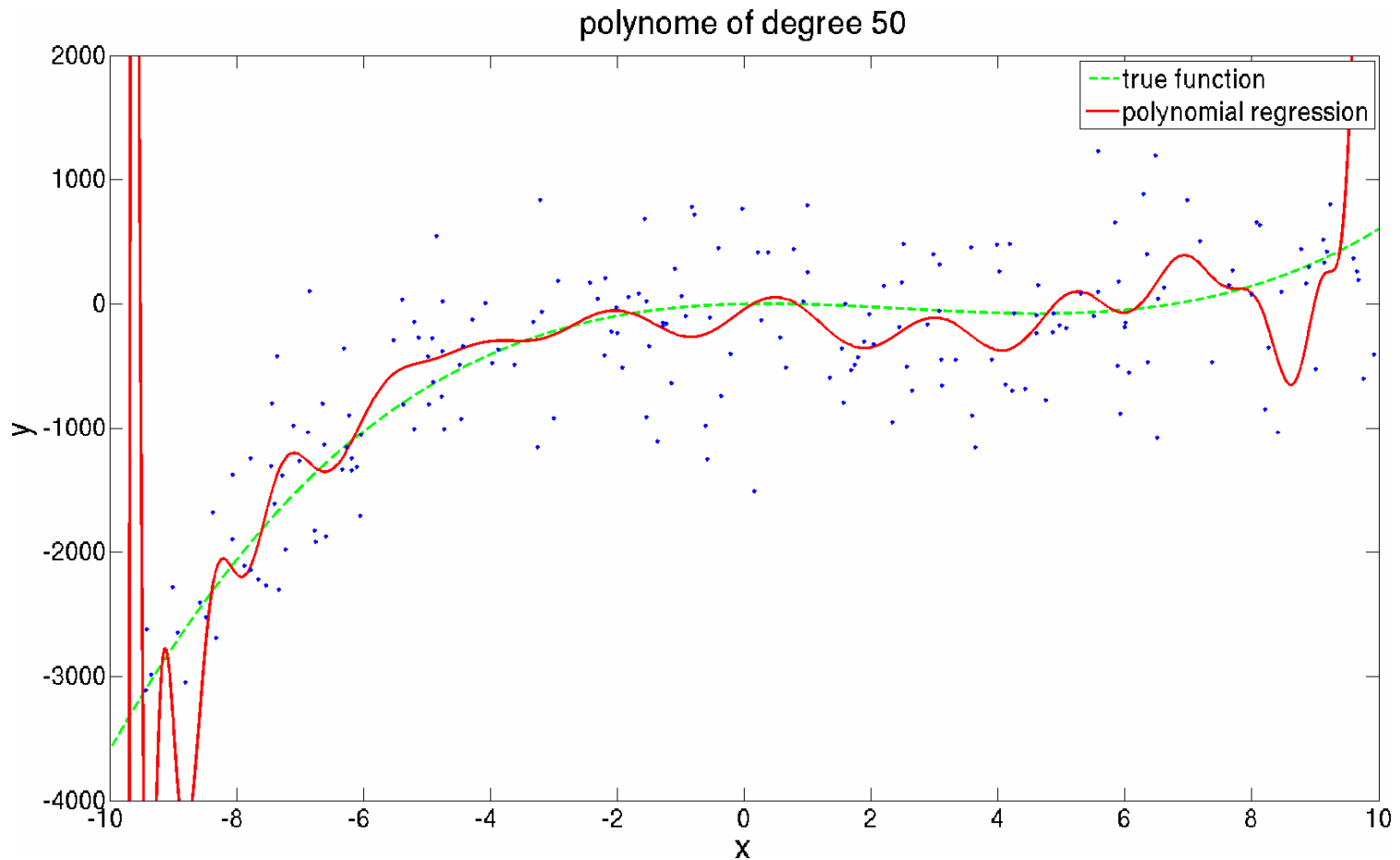
Bayesian model selection

Polynomial regression – degree 2



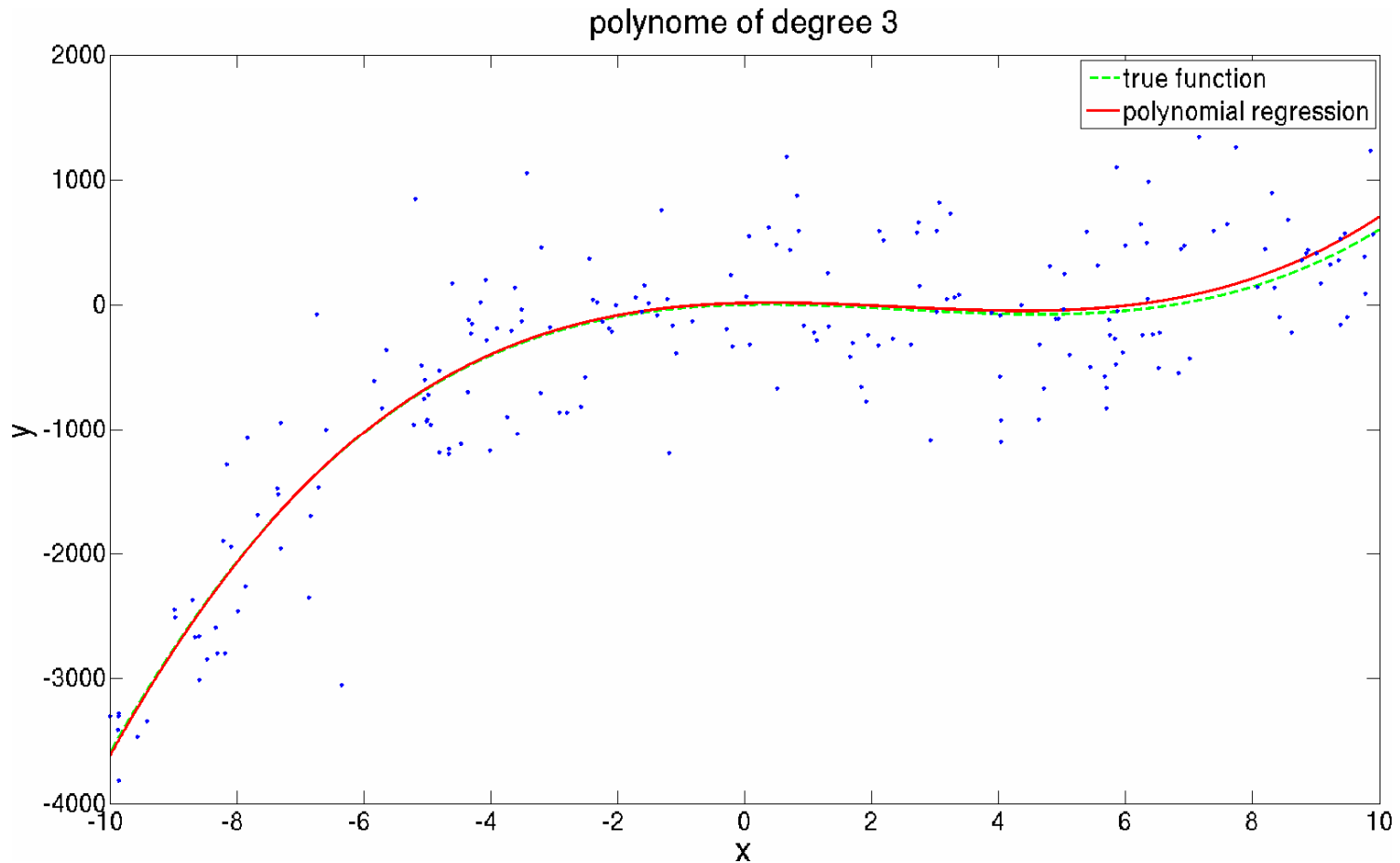
Bayesian model selection

Polynomial regression – degree 50



Bayesian model selection

Polynomial regression – (real) degree 3



Bayesian model selection

- We have

$$p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\mathbf{w}$$

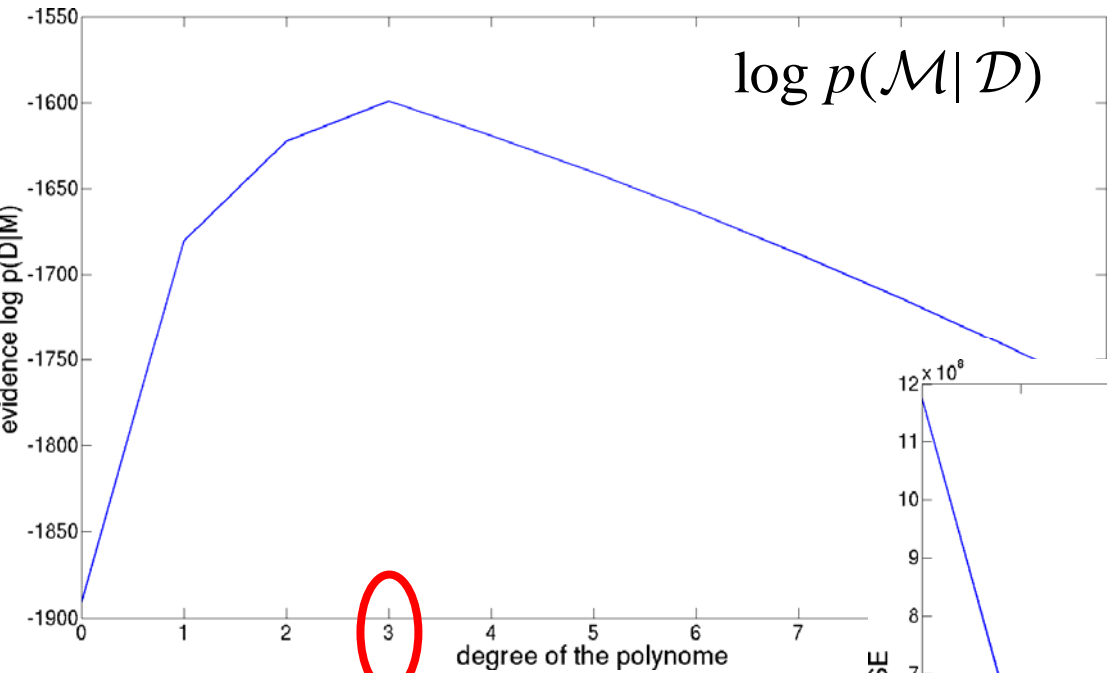
- Here $\mathcal{D} = \{\mathbf{X}, \mathbf{t}\}$, where \mathbf{X} is not modelled and acts as a parameter
 $\theta = \mathbf{w}$

- So we have

$$\begin{aligned} p(\mathcal{D} | \mathcal{M}) &= \int p(\mathbf{t} | \mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\ &= -\frac{N}{2} \log \left(\frac{2\pi}{\beta} \right) - \frac{\beta}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n))^2 \end{aligned}$$

Bayesian model selection

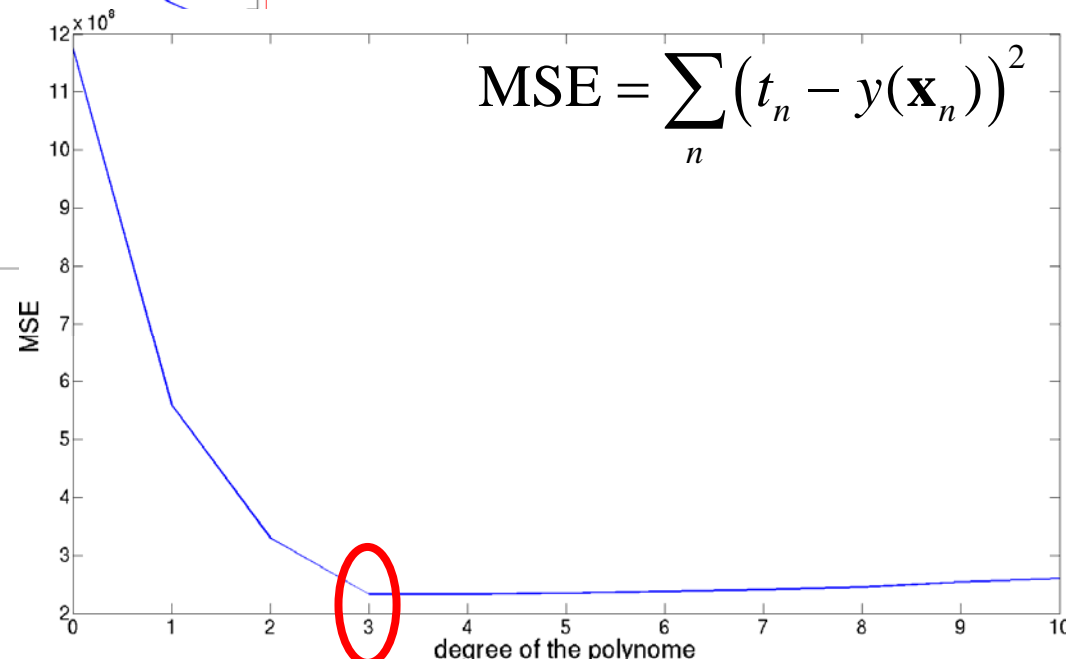
Model's posterior probability



$\log p(\mathcal{M}|\mathcal{D})$

$$p(\mathcal{M}|\mathcal{D}) \propto$$

$$p(\mathcal{M}) \int p(\mathcal{D}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta$$

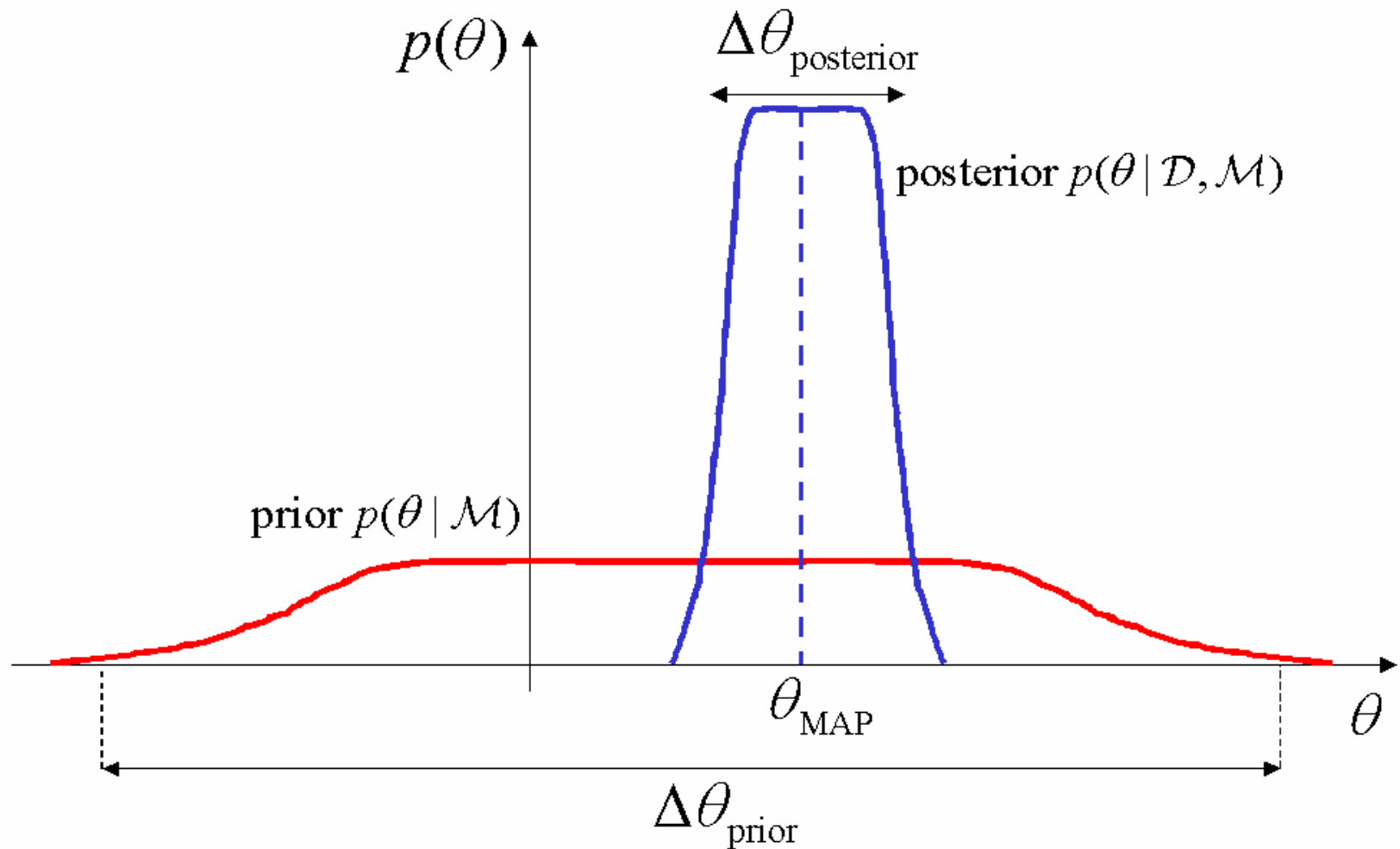


$$\text{MSE} = \sum_n (t_n - y(\mathbf{x}_n))^2$$

Bayesian model selection

- $p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\theta$
- summation is hard in practice
- approximations

Bayesian model selection



Bayesian model selection

- $p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \theta, \mathcal{M}) p(\theta | \mathcal{M}) d\theta$
- summation is hard in practice
- approximations

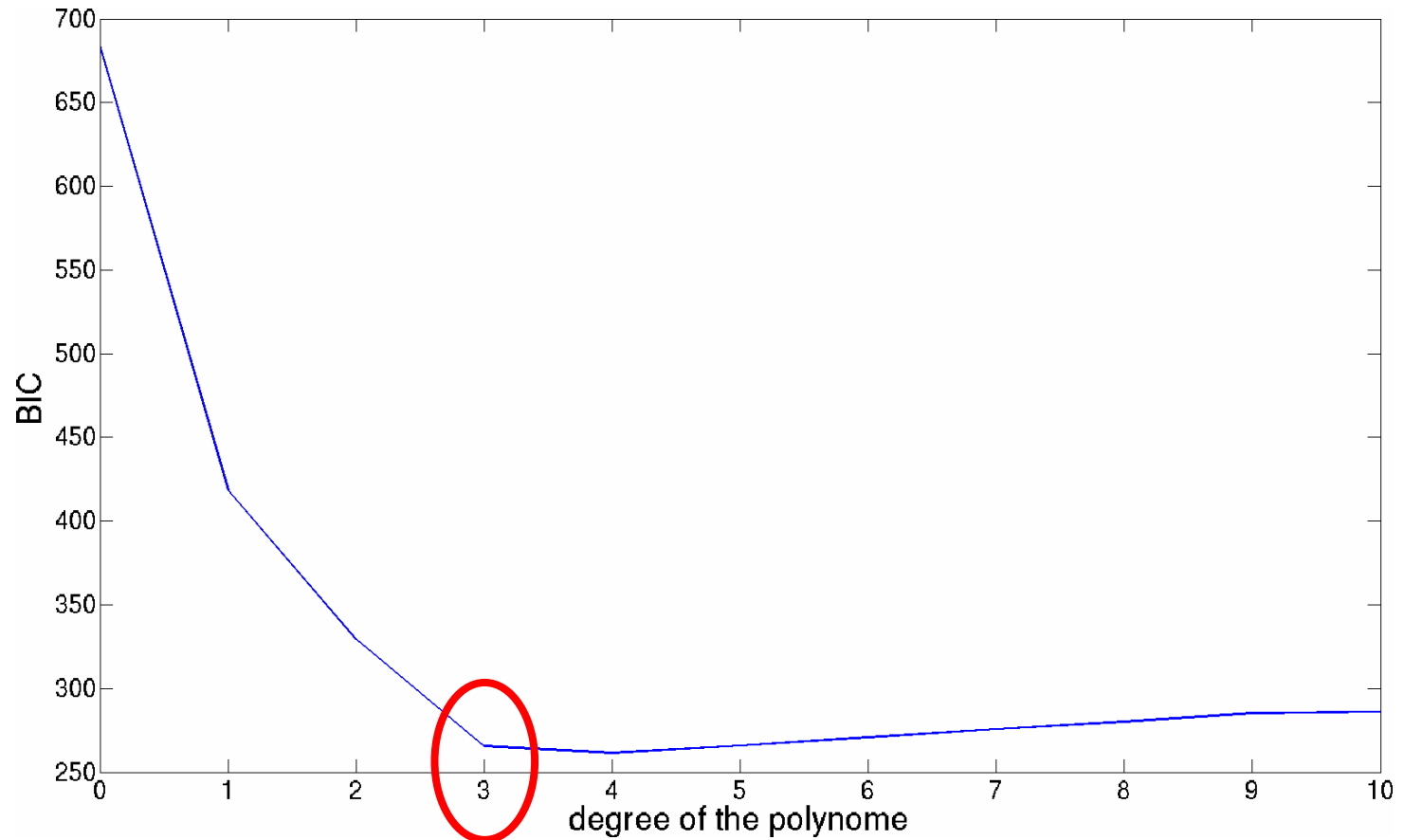
$$p(\mathcal{D} | \mathcal{M}) \approx p(\mathcal{D} | \theta_{\text{MAP}}, \mathcal{M}) \left(\frac{\Delta \theta_{\text{posterior}}}{\Delta \theta_{\text{prior}}} \right)^{\text{dim}(\theta)}$$

$$\log p(\mathcal{D} | \mathcal{M}) \approx \log p(\mathcal{D} | \theta_{\text{MAP}}, \mathcal{M}) + \text{dim}(\theta) f(\mathcal{D}, \mathcal{M})$$

$$\text{BIC} \Leftrightarrow f(\mathcal{D}, \mathcal{M}) = -\log N$$

Bayesian model selection

BIC criterion



Conclusion

Modern machine learning is mostly probabilistic and very often Bayesian

- graphical models are powerful
- although SVMs are a fierce competitor
- Bayesian networks do NOT encode causality

Conclusion

Bayesian statistics

- model the uncertainty in the parameters through a prior distribution that can contain knowledge 😊
 - ➔ priors (and therefore probabilities in general) can be subjective but they are not arbitrary!
- average predictions over the possible parameters 😊
- require to choose a good prior 😞
- are very hard in practice 😞

Conclusion

- Predictions of all the models sample from our posterior distribution are averaged!
- No need for a validation set, model selection happens on the training set

- Predictive distribution

$$p\left(\hat{t} \mid \hat{\mathbf{x}}, \mathcal{D}, \mathcal{M}\right) = \int p\left(\hat{t} \mid \hat{\mathbf{x}}, \theta, \mathcal{M}\right) p(\theta \mid \mathcal{D}, \mathcal{M}) d\theta$$

- Ex: Bayesian linear regression

$$p\left(\hat{t} \mid \hat{\mathbf{x}}, \mathbf{X}, \mathbf{t}\right) = \int p\left(\hat{t} \mid \hat{\mathbf{x}}, \mathbf{w}\right) p(\mathbf{w} \mid \mathbf{X}, \mathbf{t}) d\mathbf{w}$$