Complex probabilistic models for inference, learning and data fusion

Fusion in simple models

Peter Antal

antal@mit.bme.hu

Overview

- Basic concepts of probability theory
  - Joint distribution
  - Conditional probability
  - Bayes’ rule
  - Chain rule
  - Marginalization
  - General inference
  - Independence
    - Conditional independence
    - Contextual independence
    - Direct dependency
    - Independence model
      - Logical properties
- Naive Bayesian networks
  - Definition
  - Inference
  - Full Bayesian treatment
    - Specification
    - Inference
    - Learning

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain
-

E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

$Cavity = false \wedge Toothache = false$

$Cavity = false \wedge Toothache = true$

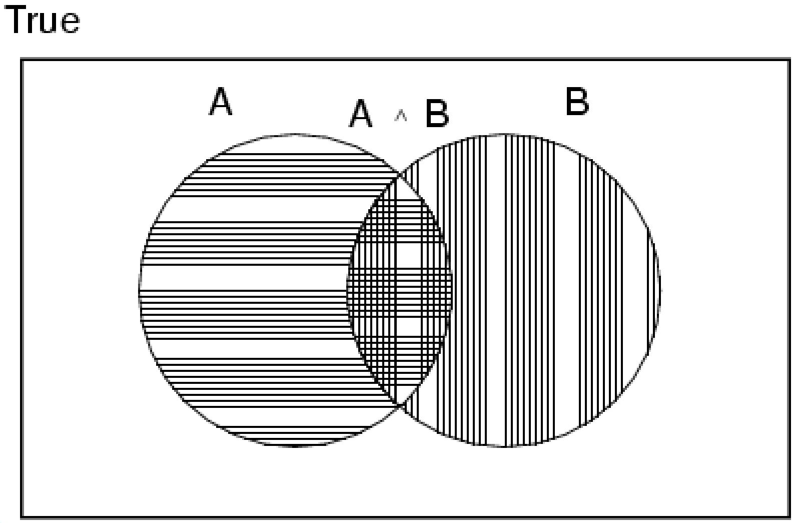
$Cavity = true \wedge Toothache = false$

$Cavity = true \wedge Toothache = true$

- Atomic events are mutually exclusive and exhaustive

Axioms of probability

- For any propositions *A, B*
-
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  -



Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

- Boolean random variables
- e.g., *Cavity* (do I have a cavity?)
-
- Discrete random variables
- e.g., *Weather* is one of $<sunny,rainy,cloudy,snow>$
- Domain values must be exhaustive and mutually exclusive

- Elementary proposition constructed by assignment of a value to a
- random variable: e.g., $Weather = sunny$, $Cavity = false$
- (abbreviated as $\neg cavity$)

- Complex propositions formed from elementary propositions and standard logical connectives e.g., $Weather = sunny \vee Cavity = false$

Joint (probability) distribution

- Prior or unconditional probabilities of propositions
- e.g., $P(\mathit{Cavity} = \text{true}) = 0.1$ and $P(\mathit{Weather} = \text{sunny}) = 0.72$ correspond to belief prior to arrival of any (new) evidence
-
- Probability distribution gives values for all possible assignments:
- $\mathbf{P}(\mathit{Weather}) = <0.72,0.1,0.08,0.1>$ (normalized, i.e., sums to 1)
- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
- $\mathbf{P}(\mathit{Weather},\mathit{Cavity})$ = a 4 × 2 matrix of values:
-

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Conditional probability

- Conditional or posterior probabilities
- e.g., $P(\mathit{cavity} \mid \mathit{toothache}) = 0.8$

  i.e., given that *toothache* is all I know

- (Notation for conditional distributions:
- $\mathbf{P}(\mathit{Cavity} \mid \mathit{Toothache})$ = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
- $P(\mathit{cavity} / \mathit{toothache}, \mathit{cavity}) = 1$

- New evidence may be irrelevant, allowing simplification, e.g.,
- $P(\mathit{cavity} / \mathit{toothache}, \mathit{sunny}) = P(\mathit{cavity} \mid \mathit{toothache}) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) \,/\, P(b)$ if $P(b) > 0$
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(\mathit{Weather}, \mathit{Cavity}) = \mathbf{P}(\mathit{Weather} \,/\, \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})$
- (View as a set of 4 × 2 equations, not matrix mult.)

Bayes' rule

An algebraic triviality

$$p(X\,|\,Y) = \frac{p(Y\,|\,X)\,p(X)}{p(Y)} = \frac{p(Y\,|\,X)\,p(X)}{\sum_X p(Y\,|\,X)\,p(X)}$$

A scientific research paradigm

$$p(Model\,|\,Data) \propto p(Data\,|\,Model)\,p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause\,|\,Effect) \propto p(Effect\,|\,Cause) \times p(Cause)$$

Chain rule

- Chain rule is derived by successive application of product rule:
- $P(X_1, \dots, X_n)$ $= P(X_1,\dots,X_{n-1})\, P(X_n \mid X_1,\dots,X_{n-1})$

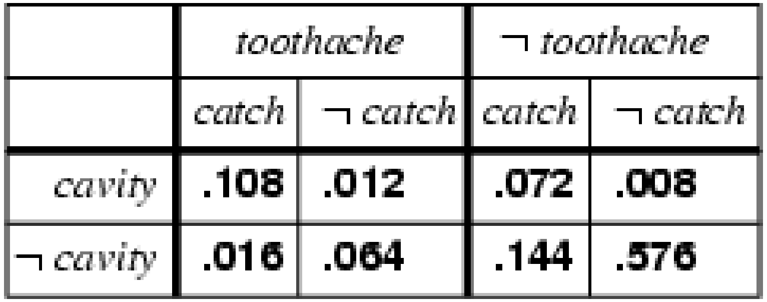
$$= P(X_1,\dots,X_{n-2})\, P(X_{n-1} \mid X_1,\dots,X_{n-2})\, P(X_n \mid X_1,\dots,X_{n-1})$$

$$= \dots$$

$$= \pi\, P(X_i \mid X_1, \dots, X_{i-1})$$

Marginalization

- ~Summing out/averaging out
- Start with the joint probability distribution:

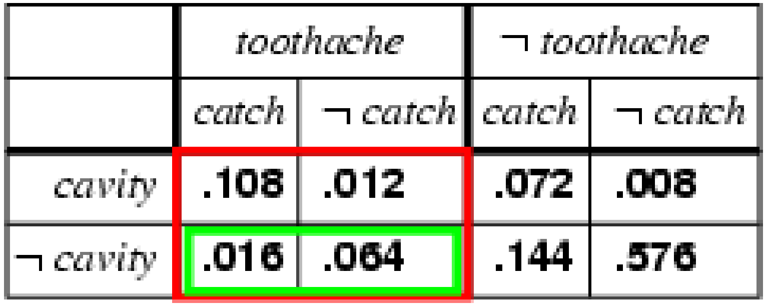


	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

Inference by enumeration

- Start with the joint probability distribution:
-

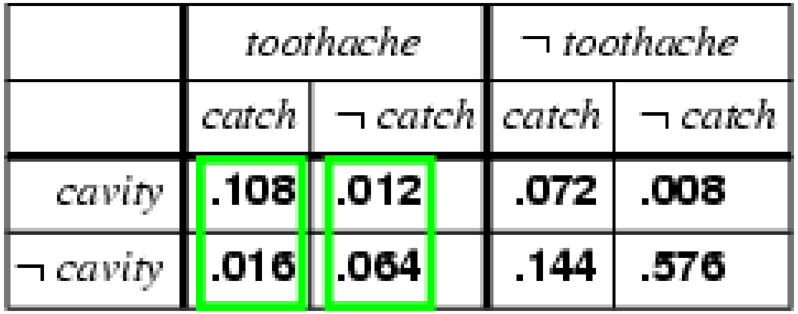


	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Can also compute conditional probabilities:
-

$$
\begin{aligned}
P(\neg \mathit{cavity} \mid \mathit{toothache}) &= \frac{P(\neg \mathit{cavity} \wedge \mathit{toothache})}{P(\mathit{toothache})} \\
&= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064} \\
&= 0.4
\end{aligned}
$$

Normalization



	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Denominator can be viewed as a normalization constant α

$$
\begin{aligned}
\mathbf{P}(\mathit{Cavity} \mid \mathit{toothache}) &= \alpha,\ \mathbf{P}(\mathit{Cavity},\mathit{toothache}) \\
&= \alpha,\ [\mathbf{P}(\mathit{Cavity},\mathit{toothache},\mathit{catch}) + \mathbf{P}(\mathit{Cavity},\mathit{toothache},\neg\ \mathit{catch})] \\
&= \alpha,\ [<0.108,0.016> + <0.012,0.064>] \\
&= \alpha,\ <0.12,0.08> = <0.6,0.4>
\end{aligned}
$$

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Any question about observable events in the domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Independence, Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z defined as follows

for all x,y and z with $P(z)>0$: $P(x;y|z)=P(x|z)\ P(y|z)$

(Almost) alternatively, $I_P(X;Y|Z)$ iff

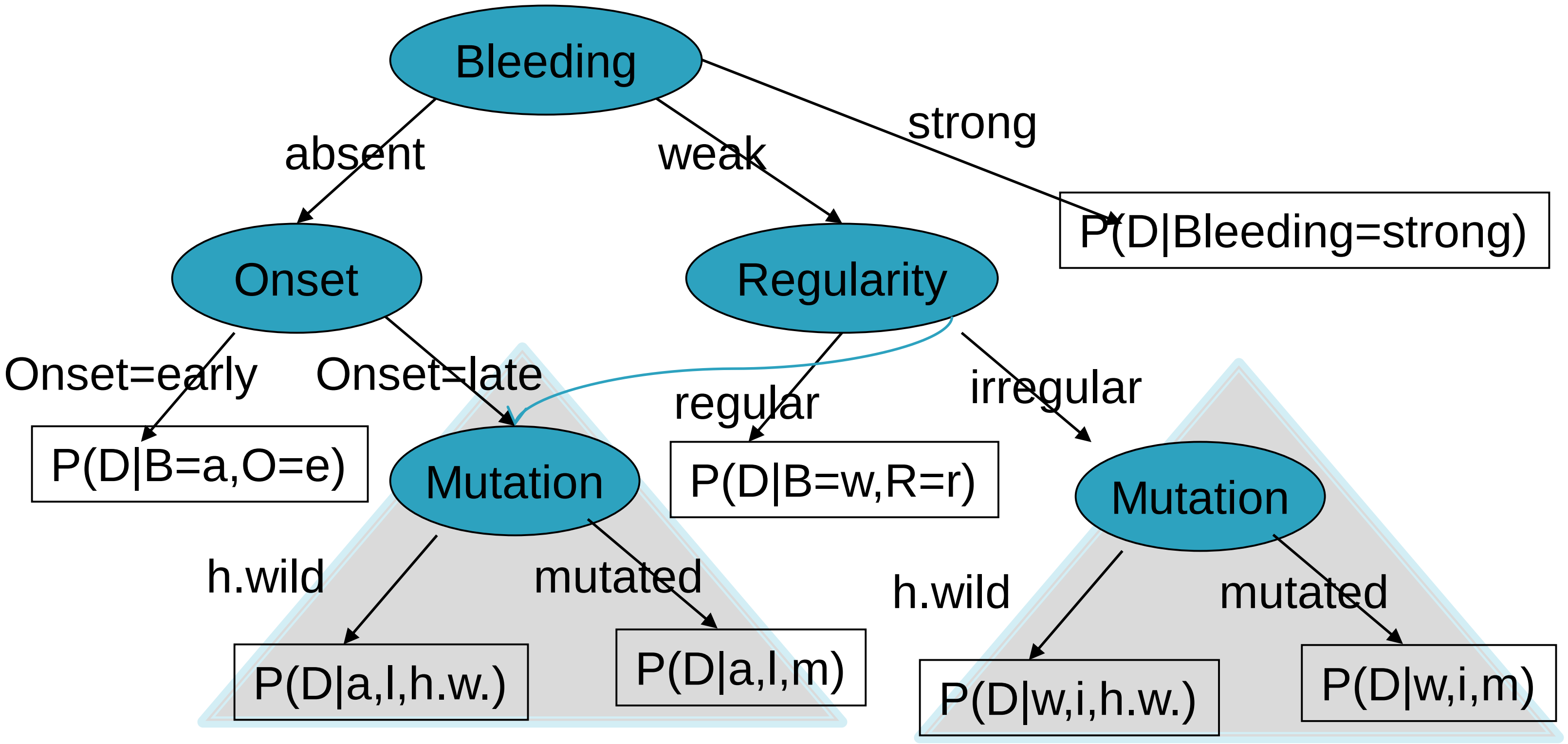
$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z)$ =def= $\neg\ I_P(X;Y|Z)$

Direct dependence: $D_P(X;Y|V/\{X,Y\})$

Context-specific independence

Contextual independence: $I_P(X;Y|Z=z)$ for not all z.



Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.

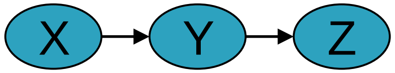
Decision graph: If conditions are equivalent, then subtrees can be merged.

E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$

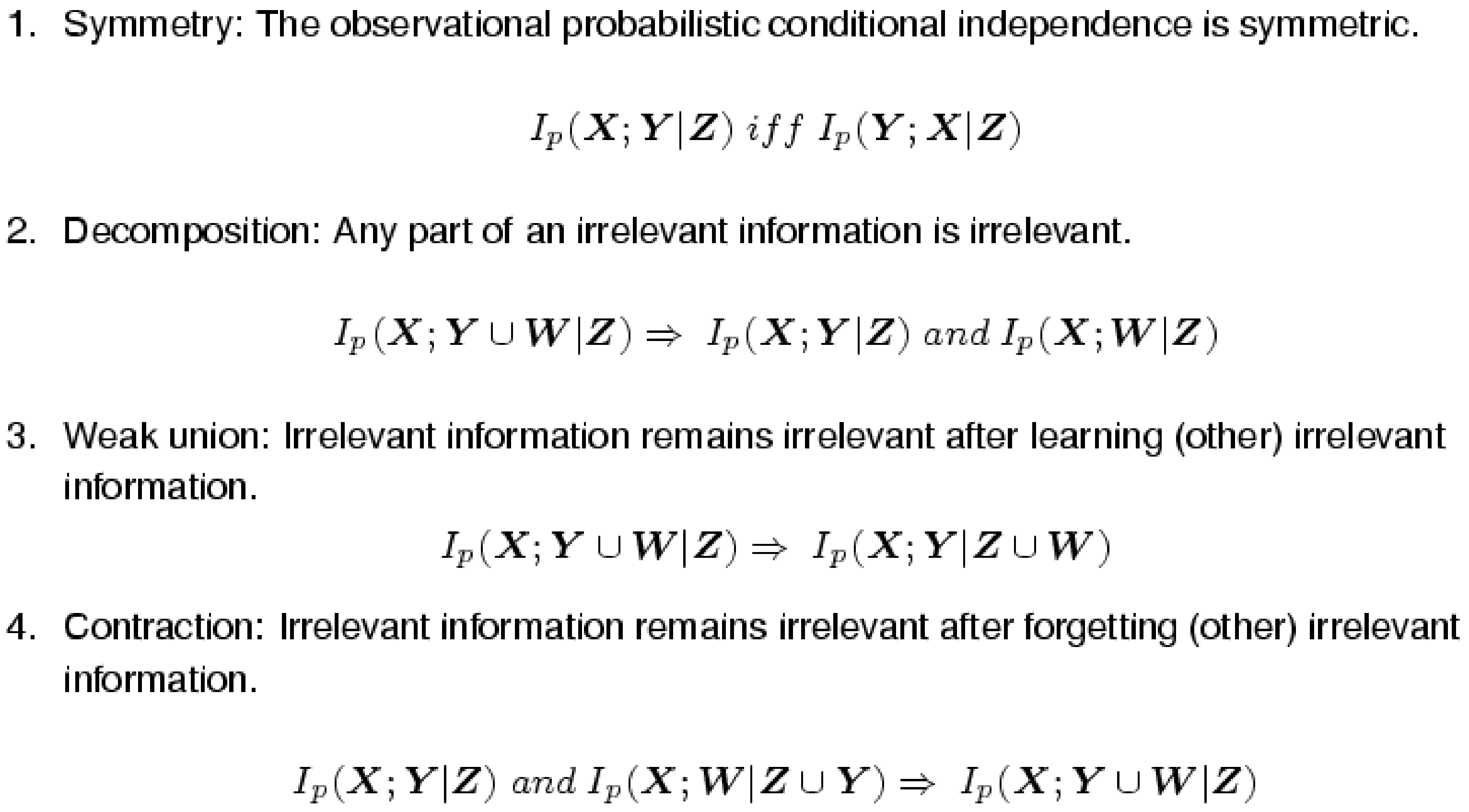


If P(X,Y,Z) is a Markov chain, then
$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

Measures of dependence

- Information theoretic based dependence
  - Entropy: H(X)
  - Conditional entropy: H(X|Y)
  - Kullback-Leibler divergence (KL(p||q))
    - Not distance (asymmetric, triangle inequality)
    - Always positive
  - Mutual information: MI(X;Y), MI(X;Y|Z)
    - MI(X;Y)=H(X)-H(X|Y)
    - MI(X;Y)=KL(p(X,Y)||p(X)p(Y))

The semi-graphoid axioms



1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ iff\ I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

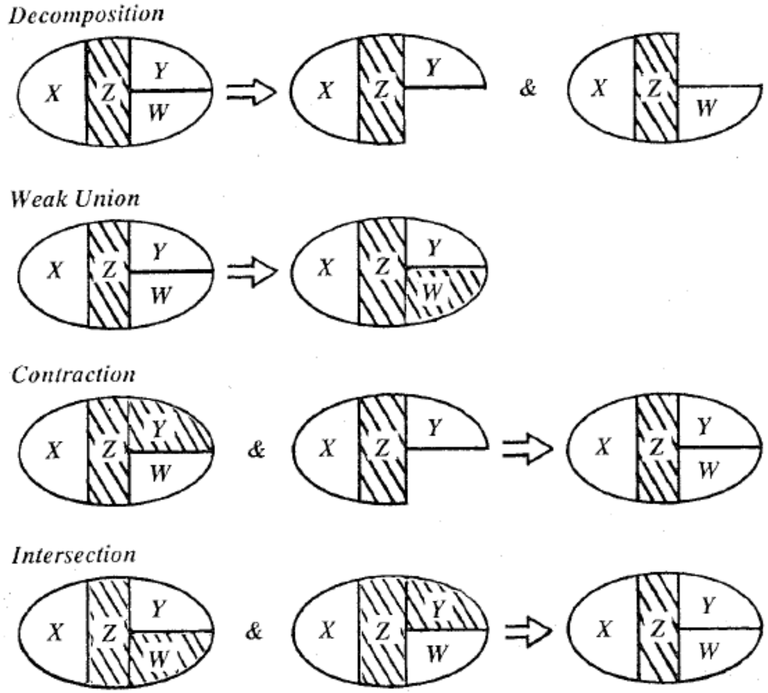
Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution). SG is sound, but incomplete inference.

Graphoids

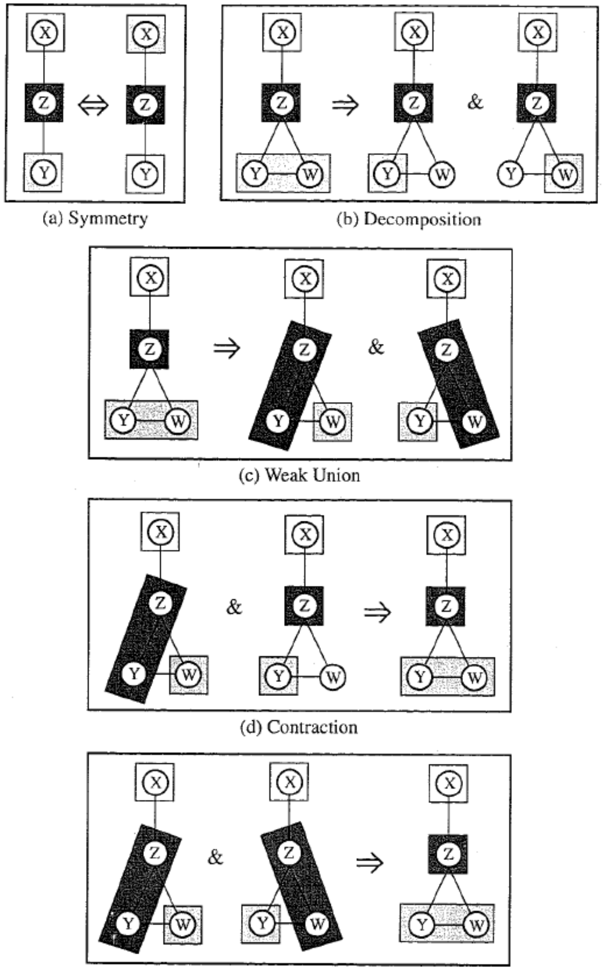
Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}\cup\mathbf{W})\ and\ I_p(\mathbf{X};\mathbf{W}|\mathbf{Z}\cup\mathbf{Y}) \Rightarrow\ I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z})$$

Graphoids: Semi-graphoids+Intersection
(holds only in strictly positive distribution)



J.Pearl: Probabilistic Reasoning in intelligent systems, 1998



(a) Symmetry

(b) Decomposition

(c) Weak Union

(d) Contraction

(e) Intersection

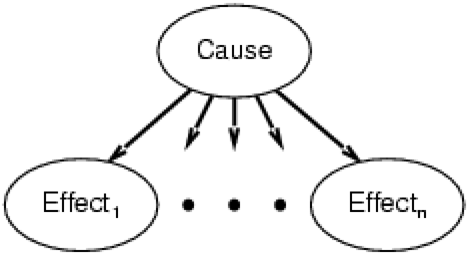
Simple probabilistic models

- Total independence
- **Naive Bayesian networks**
- Hidden Markov Models

Naive Bayesian network

Assumptions:

1, Two types of nodes: a cause and effects.



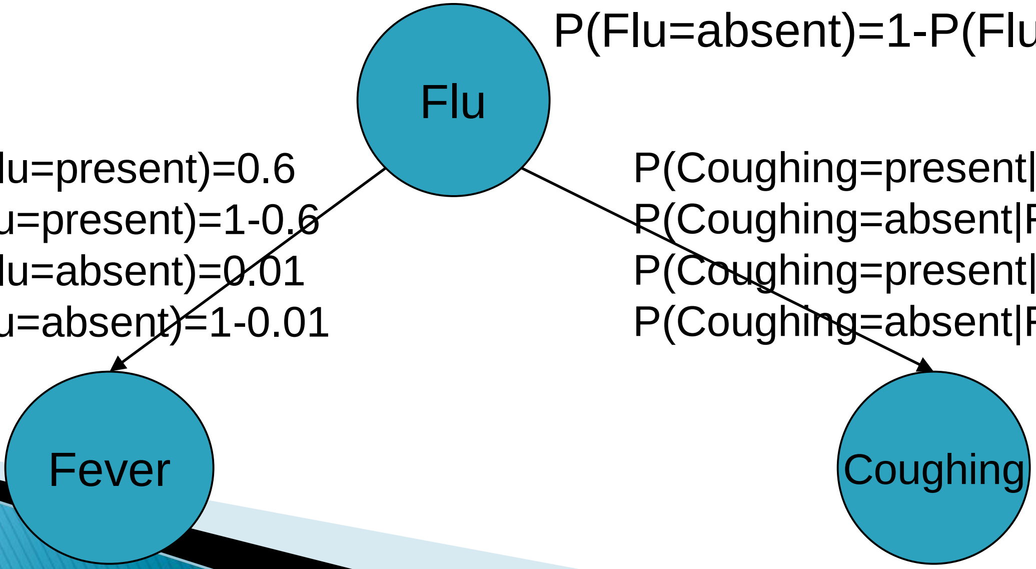
2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model

P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)



P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1-0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1-0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

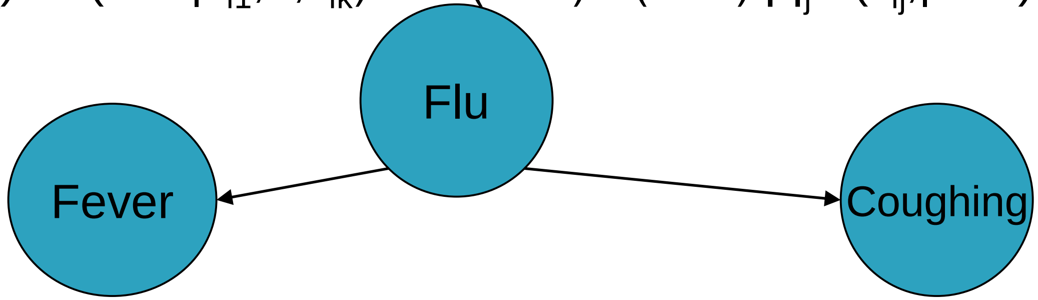
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

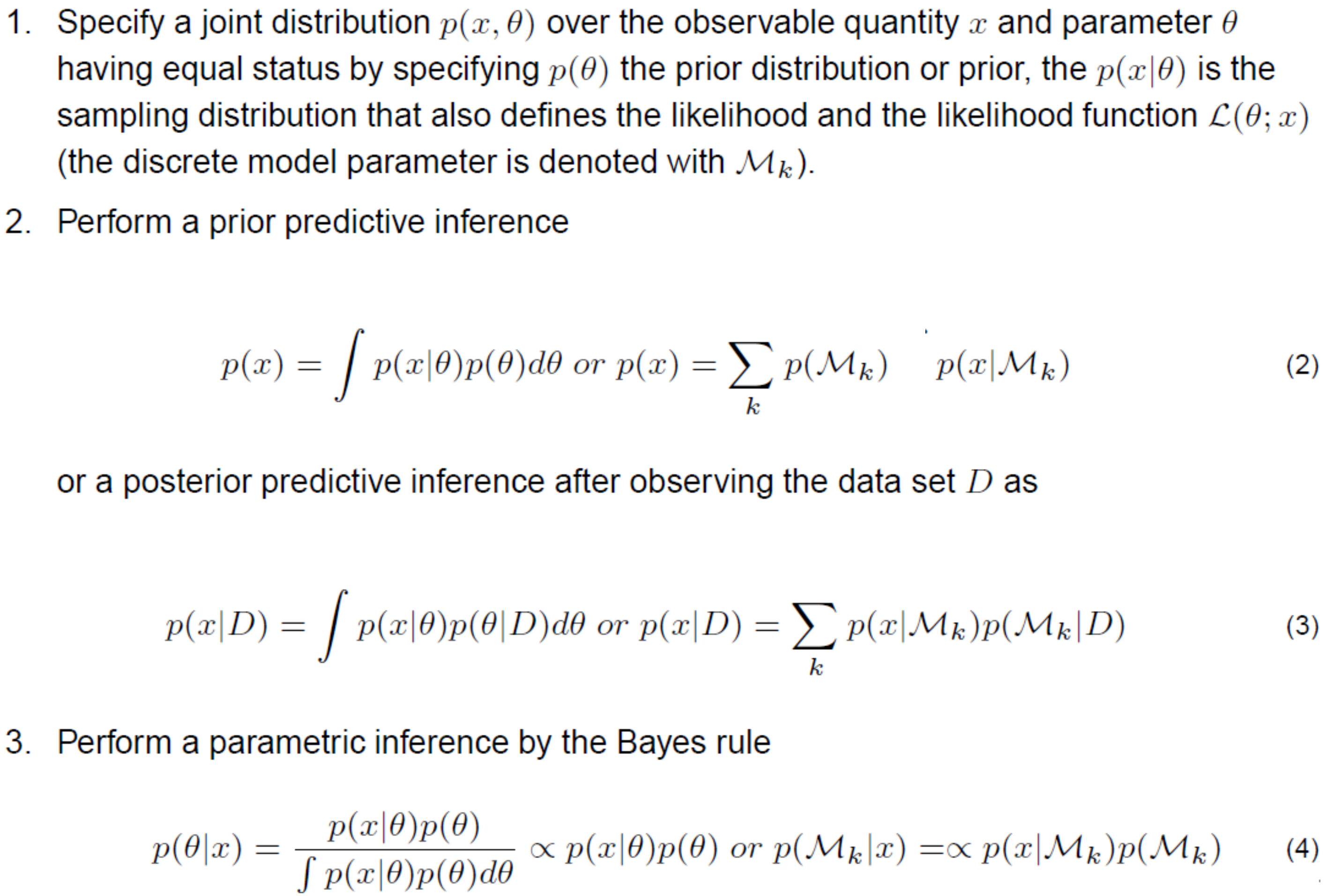
$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \,|\, Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \,|\, Flu = present)\, p(Coughing = present \,|\, Flu = present)$$

The Bayesian framework



1. Specify a joint distribution $p(x, \theta)$ over the observable quantity x and parameter θ having equal status by specifying $p(\theta)$ the prior distribution or prior, the $p(x|\theta)$ is the sampling distribution that also defines the likelihood and the likelihood function $\mathcal{L}(\theta; x)$ (the discrete model parameter is denoted with $\mathcal{M}_k$).
2. Perform a prior predictive inference

$$p(x) = \int p(x|\theta)p(\theta)d\theta \; or \; p(x) = \sum_k p(\mathcal{M}_k) \quad p(x|\mathcal{M}_k) \tag{2}$$

or a posterior predictive inference after observing the data set D as

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta \; or \; p(x|D) = \sum_k p(x|\mathcal{M}_k)p(\mathcal{M}_k|D) \tag{3}$$

3. Perform a parametric inference by the Bayes rule

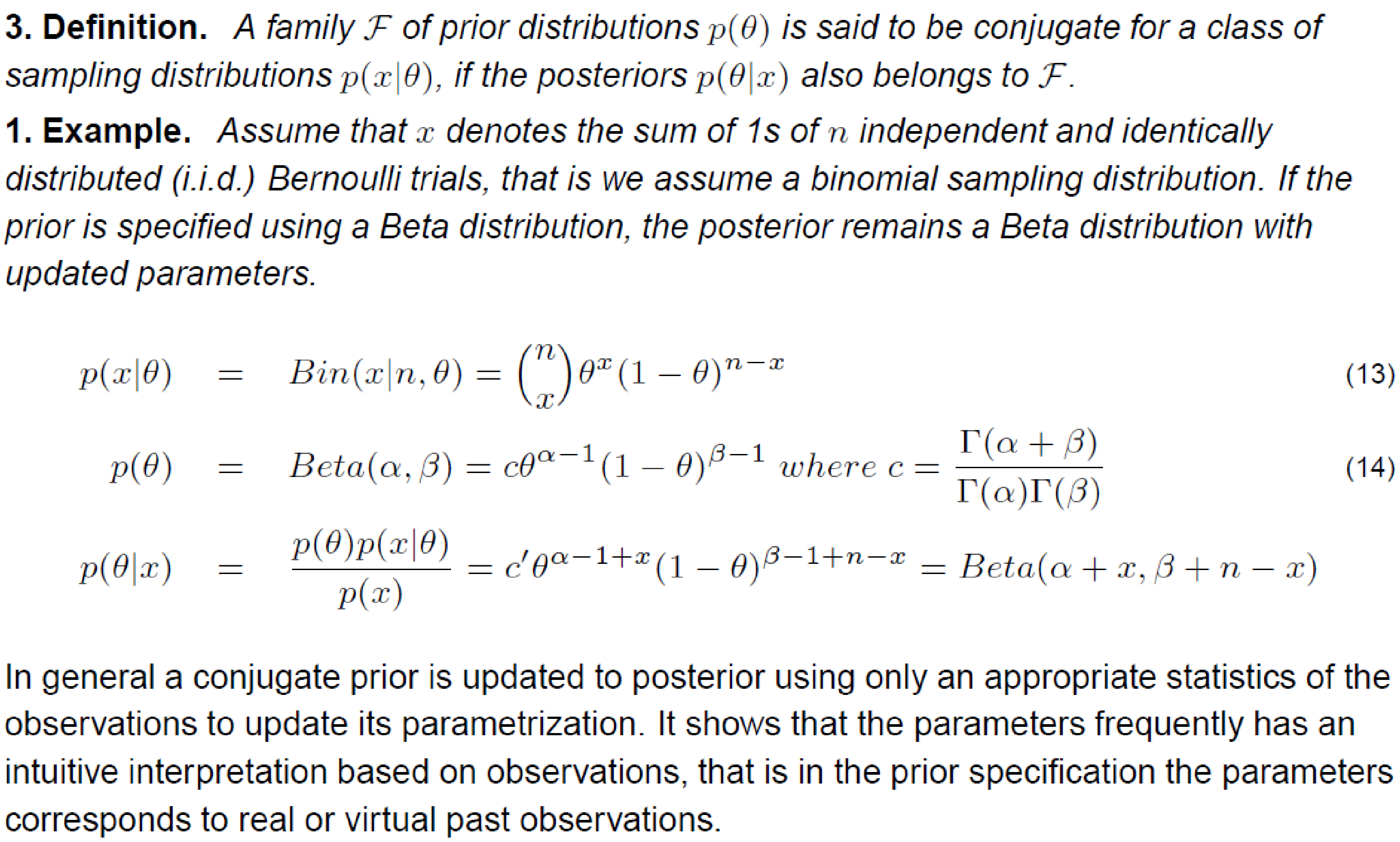
$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta) \; or \; p(\mathcal{M}_k|x) =\propto p(x|\mathcal{M}_k)p(\mathcal{M}_k) \tag{4}$$

IDA: 2.1,2.2,2.3

Full Bayesian naive-BN

- Structure prior: p(G)
  - Specify priors for edges in G
  - Penalize deviation from a prior structure G_0
- Parameter prior: $p(\Theta|G)$
  - θ denotes the complete parametrization for G
  - Specify $p(\Theta|G)$ independently for each variable?
  - Specify $p(\Theta|G)$ using a „convenient” (~conjugate) prior?
- Inference
  - Tractable?

The Beta distribution



3. Definition. *A family $\mathcal{F}$ of prior distributions $p(\theta)$ is said to be conjugate for a class of sampling distributions $p(x|\theta)$, if the posteriors $p(\theta|x)$ also belongs to $\mathcal{F}$.*

1. Example. *Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.*

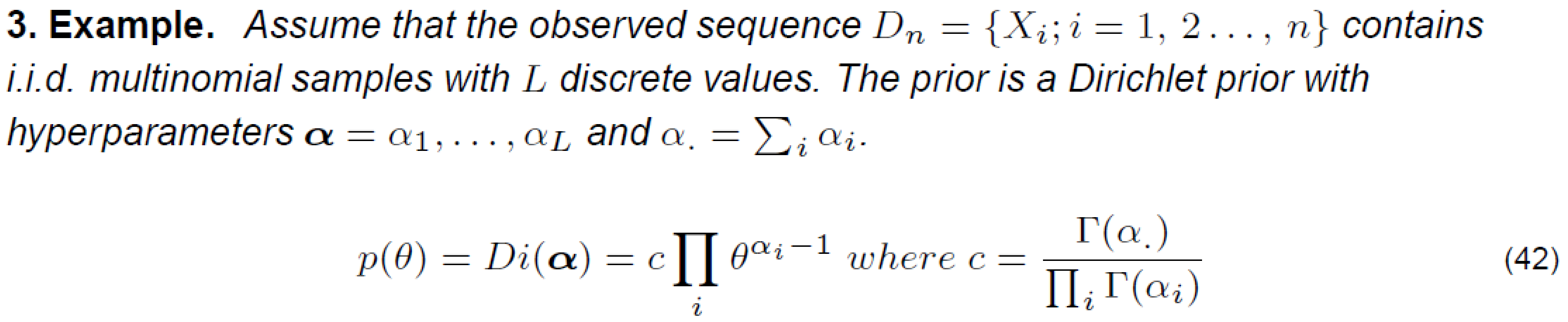
$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x} \tag{13}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \; where \; c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{14}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

In general a conjugate prior is updated to posterior using only an appropriate statistics of the observations to update its parametrization. It shows that the parameters frequently has an intuitive interpretation based on observations, that is in the prior specification the parameters corresponds to real or virtual past observations.

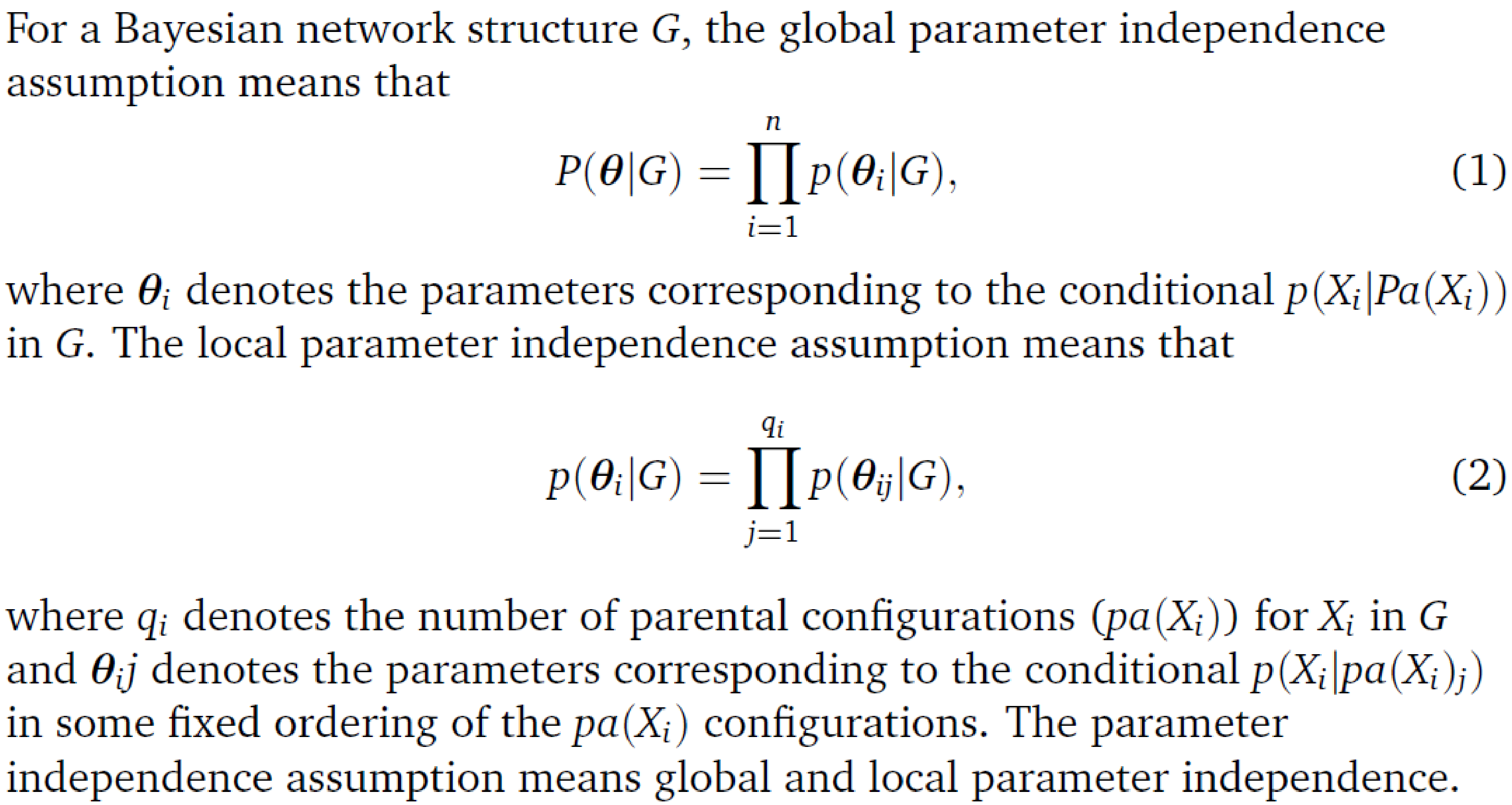
The Dirichlet distribution



3. Example. *Assume that the observed sequence $D_n = \{X_i; i = 1, 2 \ldots, n\}$ contains i.i.d. multinomial samples with L discrete values. The prior is a Dirichlet prior with hyperparameters $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_L$ and $\alpha_. = \sum_i \alpha_i$.*

$$p(\theta) = Di(\boldsymbol{\alpha}) = c \prod_i \theta^{\alpha_i - 1} \; where \; c = \frac{\Gamma(\alpha_.)}{\prod_i \Gamma(\alpha_i)} \tag{42}$$

Parameter independence



For a Bayesian network structure G, the global parameter independence assumption means that

$$P(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i|G), \qquad (1)$$

where $\boldsymbol{\theta}_i$ denotes the parameters corresponding to the conditional $p(X_i|Pa(X_i))$ in G. The local parameter independence assumption means that

$$p(\boldsymbol{\theta}_i|G) = \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|G), \qquad (2)$$

where q_i denotes the number of parental configurations $(pa(X_i))$ for X_i in G and $\boldsymbol{\theta}_i j$ denotes the parameters corresponding to the conditional $p(X_i|pa(X_i)_j)$ in some fixed ordering of the $pa(X_i)$ configurations. The parameter independence assumption means global and local parameter independence.

Full Bayesian inference with N–BNs

- Integration over parameters?
  - Analytical solution!
- Bayesian model averaging over exponential number of structures?
  - Analytical solution!
- Existence of equivalent „super”–parametrization!!
  - DISCUSSION&PROOFS: later
  - PDSS:9.2.5

Summary

- Basic concepts of probability theory
  - On the use of probabilities: PDSS:2.1
  - The Bayesian framework: PDSS:2.2
  - LATER: Indepence models: PDSS:2.3

  https://www.mit.bme.hu/system/files/oktatas/targyak/9383/Antal_Valoszinusegi.pdf
- Naive Bayesian networks
  - Definition, Inference (PDSS:2.5.1)
  - Full Bayesian treatment: LATER
  - ➔IDA:9.2.5 (~9.2)

  https://www.mit.bme.hu/system/files/oktatas/targyak/9383/Antal_IDA.pdf