Advanced AI Techniques

I. Bayesian Networks / 2. Parameter Learning

Wolfram Burgard, Luc de Raedt,
Bernhard Nebel, Lars Schmidt-Thieme

Institute of Computer Science
University of Freiburg
http://www.informatik.uni-freiburg.de/

Advanced AI Techniques

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

1. Maximum Likelihood Parameter Estimates

2. Bayesian Parameter Estimates / One Variable

3. Bayesian Parameter Estimates / Several Variables

Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute of Computer Science, University of Freiburg, Germany,
Course on Advanced AI Techniques, winter term 2005 1/41

Given

- a bayesian network structure $G := (V, E)$ on a set of variables V and
- a data set $D \in \mathrm{dom}(V)^*$ of cases.

Learning the parameters of the bayesian network means to find vertex potentials

$$(p_v)_{v \in V}$$

s.t. some **optimality criterion** w.r.t. G and D holds.

Advanced AI Techniques / 1. Maximum Likelihood Parameter Estimates
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

The simplest criterion is the **maximum likelihood criterion**, i.e., the probability of the data given the bayesian network is maximal:

find $(p_v)_{v \in V} s.t.\ p(D)$ is maximal,

where p denotes the JPD build from $(p_v)_{v \in V}$.

$$p(D) = \prod_{d \in D} p(d) = \prod_{d \in D} \prod_{v \in V} p_v(d|_{\mathrm{fam}(v)})$$

$(p_v)_{v \in V}$ with maximal $p(D)$ are called **maximum likelihood estimates**. p is also called **likelihood**.

data:

X	Y	p_1(case)	p_2(case)
H	T	0.25	0.25
H	H	0.25	0.2
T	H	0.25	0.3
T	T	0.25	0.25
T	H	0.25	0.3
T	H	0.25	0.3
H	T	0.25	0.25
T	T	0.25	0.25
p(D)=		$1.5259 \cdot 10^{-5}$	$2.1093 \cdot 10^{-5}$

JPD$_1$:

Y =	H	T
X = H	.25	.25
T	.25	.25

JPD$_2$:

Y =	H	T
X = H	.2	.25
T	.3	.25

Lemma 1. $p(D)$ *is maximal iff*

$$p_v(x|y) := \frac{|\{d \in D \mid d|_v = x, d|_{\mathrm{pa}(v)} = y\}|}{|\{d \in D \mid d|_{\mathrm{pa}(v)} = y\}|}$$

(if there are $d \in D$ with $d|_{\mathrm{pa}(v)} = y$, otherwise $p_v(x|y)$ can be choosen arbitrarily – $p(D)$ does not depend on it).

Instead of the likelihood p often $\log p$ is used, called **log-likelihood**.

Proof. Due to independence of the cases and the factorization of the JPD in bayesian networks, $p(D)$ factors as

$$p(D) = \prod_{d\in D} p(d) = \prod_{d\in D}\prod_{v\in V} p_v(d|_{\text{fam}(v)}) = \prod_{v\in V}\prod_{d\in D|_{\text{fam}(v)}} p_v(d)$$

which is maximal if for all $v \in V$

$$p_v(D) := \prod_{d\in D|_{\text{fam}(v)}} p_v(d) = \prod_{x\in\text{dom}(v)}\prod_{y\in\text{dom}(\text{pa}(v))} p_v(x|y)^{n_D(x,y)}$$

is maximal, with count data

$$n_D(x,y) := |\{d\in D \mid d|_v = x, d|_{\text{pa}(v)} = y\}|$$

p_v in turn is maximal if for all $x \in \text{dom}(v)$

$$\prod_{y\in\text{dom}(\text{pa}(v))} p_v(x|y)^{n_D(x,y)}$$

is maximal. As beneath $p_v(x|y) \in [0,1]$ the only constraint to $p_v(x|y)$ is

$$\sum_{y\in\text{dom}(\text{pa}(v))} p_v(x|y) = 1$$

we have with an arbitrary $y_0 \in \text{dom}(\text{pa}(v))$

$$\prod_{\substack{y\in\text{dom}(\text{pa}(v))\\ y\neq y_0}} p_v(x|y)^{n_D(x,y)} \cdot (1 - \sum_{\substack{y\in\text{dom}(\text{pa}(v))\\ y\neq y_0}} p_v(x|y))^{n_D(x,y_0)}$$

Taking logarithms we get

$$\sum_{\substack{y\in\text{dom}(\text{pa}(v))\\ y\neq y_0}} n_D(x,y)\log p_v(x|y) + n_D(x,y_0)\log(1 - \sum_{\substack{y\in\text{dom}(\text{pa}(v))\\ y\neq y_0}} p_v(x|y))$$

Let $\text{dom}(\text{pa}(v)) := \{y_0, \ldots, y_n\}$ an enumeration and write $x_i := (x, y_i)$, $p_i := p_v(x|y_i)$ and $n_i := n_D(x, y_i)$, then we can simplify notation to

$$L(p) := \sum_{i=1}^{n} n_i \log p_i + n_0 \log(1 - \sum_{i=1}^{n} p_i)$$

To be minimal, derivatives have to vanish:

$$\frac{\partial L}{\partial p_j} = n_j \frac{1}{p_j} - n_0 \frac{1}{1-\sum_{i=1}^n p_i} \stackrel{!}{=} 0$$

which yields

$$p_j = \frac{n_j}{n_0}(1 - \sum_{i=1}^{n} p_i)$$

Summing over $j = 1, \ldots, n$ we get

$$\sum_{i=1}^{n} p_i = (\sum_{i=1}^{n}\frac{n_i}{n_0})(1 - \sum_{i=1}^{n} p_i)$$

Solving for $\sum_{i=1}^n p_i$ yields

$$\sum_{i=1}^{n} p_i = \frac{\sum_{i=1}^n n_i}{n_0 + \sum_{i=1}^n n_i}$$

and substituting this in the equations for p_j finally yields

$$p_j = \frac{n_j}{n_0}(1 - \sum_{i=1}^{n} p_i) = \frac{n_j}{n_0 + \sum_{i=1}^n n_i}$$

□

data:

X	Y	p_1(case)	p_{opt}(case)
H	T	0.25	0.25
H	H	0.2	0.125
T	H	0.3	0.375
T	T	0.25	0.25
T	H	0.3	0.375
T	H	0.3	0.375
H	T	0.25	0.25
T	T	0.25	0.25
p(D)=		$2.1093 \cdot 10^{-5}$	$2.5749 \cdot 10^{-5}$

JPD$_2$:

Y =	H	T
X = H	.2	.25
T	.3	.25

JPD$_{opt}$:

Y =	H	T
X = H	0.125	0.25
T	0.375	0.25

1. Maximum Likelihood Parameter Estimates

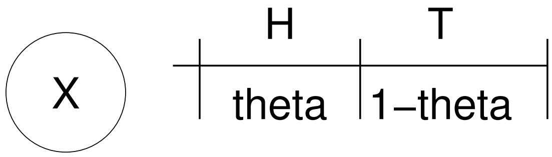
2. Bayesian Parameter Estimates / One Variable

3. Bayesian Parameter Estimates / Several Variables

Simplest case:

- one variable X,
- varable is binary (= 2 states, e.g., H and T)

$\rightsquigarrow$ 1 parameter θ.



Advanced AI Techniques / 2. Bayesian Parameter Estimates / One Variable
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

Example I: flip a coin

We flip a coin with possible outcomes head (H) or tail (T).

actual sample:

actual	H	T	H	H	H

parameter estimation:

$$\hat{p}(X = \mathsf{H}) = \frac{4}{5} = 0.8$$

Requirements for Parameter Estimation (1/3)

1. Be able to combine
   - **prior** / **background knowledge** with
   - **actual observations** / new **data**

Example II: "flip a cat"

We observe cats falling from window seats with possible outcomes

- lands on its paws (P) or
- does not land on its paws (O = ouch)

actual sample:

actual	O	P	O	O	O

parameter estimation:

$$\hat{p}(X = \mathsf{O}) = \frac{4}{5} = 0.8$$

Requirements for Parameter Estimation (2/3)

2. Be able to express different prior probabilities:
   - all events have same prior probability (e.g., coin, dice)
   - each event has a specific prior probability (e.g., cats landing on paws vs. not).

Advanced AI Techniques / 2. Bayesian Parameter Estimates / One Variable
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

Requirements for Parameter Estimation (3/3)

3. Be able to express different **strengths of prior believes**:

strong prior believes:

Many contradicting actual observations are necessary to overwrite prior believes.

"I am quite sure in advance."

weak prior believes:

Already a few contradicting actual observations are sufficient to overwrite prior believes.

"I guess, but really do not know in advance."

A Simple Model for Prior Believes

We model

- prior probabilities by a probability distribution p_{prior}.
- the strength of the prior believes by a **prior sample size** n_{prior}.

$$\hat{p} := \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}}\, p_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}}\, \hat{p}_{\text{actual}}$$

Prior sample size quantifies, how many actual observations we need s.t. prior and actual estimates have the same influence on our final estimates.

A Simple Model for Prior Believes / Example

actual sample:

actual	H	T	H	H	H

$n_{\text{actual}} = 5$

	H	T
$\hat{p}_{\text{actual}}$	0.8	0.2

	H	T
p_{prior}	0.5	0.5

$n_{\text{prior}} = 10$

combined estimate:

$$\hat{p} := \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}}\, p_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}}\, \hat{p}_{\text{actual}}$$

$$= \frac{10}{15} \cdot 0.5 + \frac{5}{15} \cdot 0.8 = 0.6$$

	H	T
$\hat{p}$	0.6	0.4

Prior Sample Size

Prior sample size can be understood literally as

the size of a prior sample

that we combine with the actual sample for our estimations.

actual sample:

actual	H	T	H	H	H

$n_{\text{actual}} = 5$

	H	T
$\hat{p}_{\text{actual}}$	0.8	0.2

	H	T
p_{prior}	0.5	0.5

$n_{\text{prior}} = 10$

$\rightsquigarrow$ prior sample:

prior	H	H	H	H	H	T	T	T	T	T

combined sample:

combined	H	H	H	H	H	T	T	T	T	T	H	T	H	H	H

(first ten: prior sample; last five: actual sample)

	H	T
$\hat{p}$	0.6	0.4

Prior Sample Size

But,

- not all prior probabilities and prior sample sizes can be expressed equivalently as prior samples, e.g.,

	H	T
p_{prior}	0.5	0.5

$n_{\text{prior}} = 5$

- prior sample size also can be choosen as fractional value, e.g.,

$$n_{\text{prior}} = 0.1$$

So far, if we specify

	H	T
p_{prior}	0.5	0.5

$n_{\text{prior}} = 10$

then ...

...for the discrete attribute X in our model

we specify its prior distribution

$$p_{\text{prior}}(X)$$

consisting of

$$p_{\text{prior}}(X = H) = \theta = 0.5$$
$$p_{\text{prior}}(X = T) = 1 - \theta = 0.5$$

...for the parameter $\Theta := p_{\text{prior}}(X = H)$

we specify its expected value $\hat{\theta} = 0.5$.

Prior Parameter Distribution

Is the prior distribution $p_{\text{prior}}(X)$ / expected parameter value $\hat{\theta}$ sufficient to answer more complex queries?

E.g., what is the prior probability that a coin is fair to some extent, i.e., $\theta \geq 0.4$ and $\theta \leq 0.6$?

$$\int_{0.4}^{0.6} p(\Theta) d\Theta$$

$\rightsquigarrow$ we need to specify a prior distribution $p(\Theta)$ of the parameter θ itself !

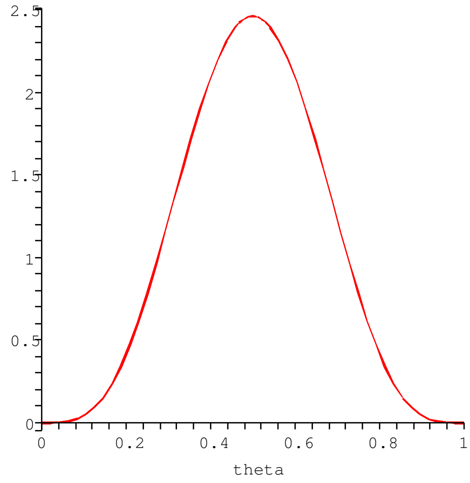


Figure 1: $p(\Theta) = \beta_{5,5}(\Theta)$: we expect the true parameter to be at 0.5.

$$\hat{\theta} = 0.5$$

$$\int_{0.4}^{0.6} p(\Theta) d\Theta = .467$$

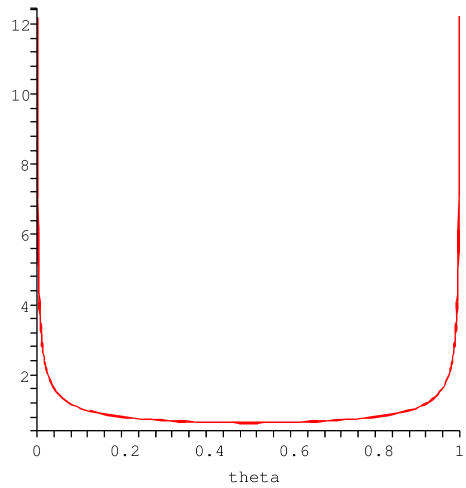
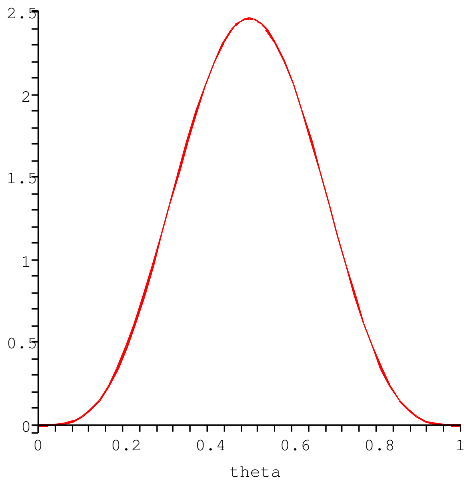


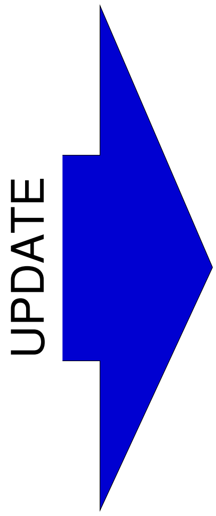
Figure 2: $p(\Theta) = \beta_{.5,.5}(\Theta)$: we expect the true parameter to be at 0 or at 1.

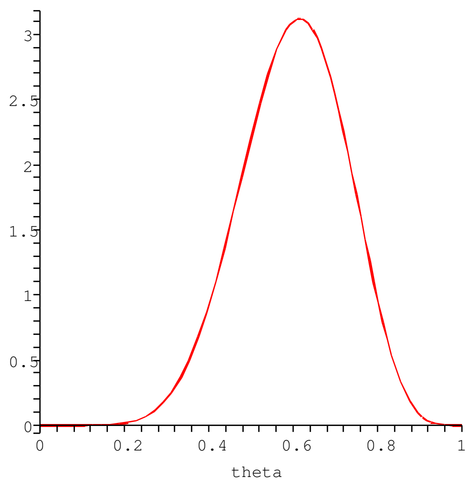
$$\hat{\theta} = 0.5$$

$$\int_{0.4}^{0.6} p(\Theta) d\Theta = .128$$

DATA
d







prior distribution
p(Theta)

a posterior distribution
p(Theta | d)

Compute the expected value θ of its a posterior distribution

$$\hat{\theta}_{\text{MAP}} := E(p(\theta \mid d))$$

called **maximal a posterior estimator (MAP)** of Θ.

Use Bayes' formula:

$$p(\theta \mid d) = \frac{p(d \mid \theta)\, p(\theta)}{p(d)}$$

$$p(d \mid \theta) = \prod_{x \in d} \theta^{\delta_{x=H}} (1-\theta)^{\delta_{x=T}}$$
$$= \theta^{|\{x \in d \mid x = H\}|} (1-\theta)^{|\{x \in d \mid x = T\}|}$$

actual sample:

actual	H	T	H	H	H

$$p(d \mid \theta) = \theta^{|\{x \in d \mid x = H\}|} (1-\theta)^{|\{x \in d \mid x = T\}|}$$
$$= \theta^4 (1-\theta)^1$$

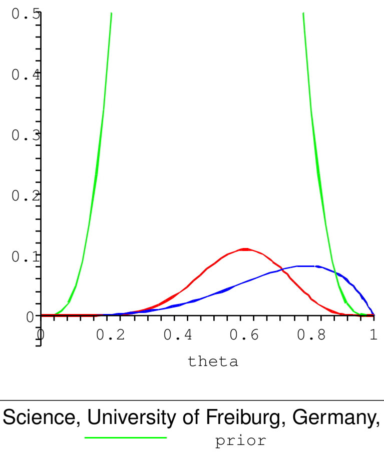
$$p(\theta) = \beta_{5,5}(\theta)$$

Computing expectation of

$$p(\theta \mid d) = \frac{p(d \mid \theta)\, p(\theta)}{p(d)} = \frac{\theta^4 (1-\theta)^1 \beta_{5,5}(\theta)}{p(d)}$$

leads to

$$\hat{p}_{\text{MAP}} = 0.6$$



For general priors, we have to solve a 1-dimensional integration problem:

$$p(\theta \,|\, d) = \frac{p(d \,|\, \theta)\, p(\theta)}{p(d)} = \frac{\theta^s\, (1-\theta)^t\, p(\theta)}{p(d)}$$

where $s := |\{x \in d \,|\, x = H\}|$ and $t := |\{x \in d \,|\, x = T\}|$.

nice:

- Problem depends on the data only via summary statistics s, t.

not so nice:

- Complicated priors $p(\theta)$ may not have analytical solutions.

Gamma function

Definition 1. Gamma function

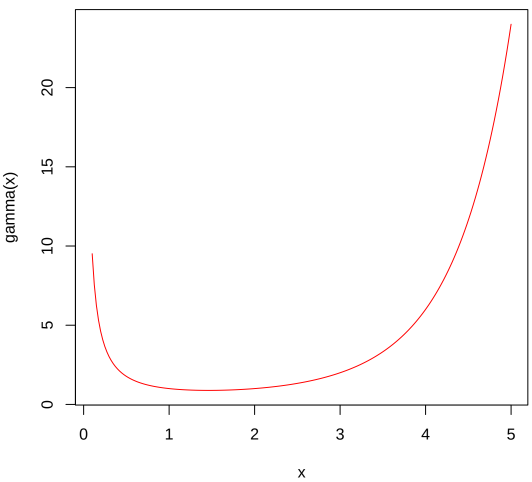
$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

converging for $x > 0$.

Lemma 2 (Γ is generalization of factorial).

(i) $\Gamma(n) = (n-1)!$ *for* $n \in \mathbb{N}$.

(ii) $\frac{\Gamma(x+1)}{\Gamma(x)} = x$.



Beta function

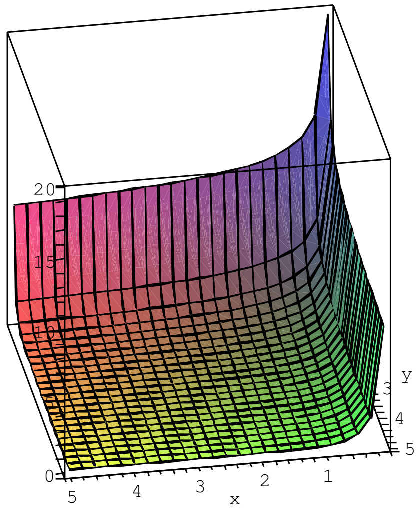
Definition 2. Beta function

$$\beta(x, y) := \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

defined for $x, y > 0$.

Lemma 3 (β is generalization of binomial).

$$1/\beta(n-m, m) = m\binom{n-1}{m}$$ *for* $n, m \in \mathbb{N}, n >$

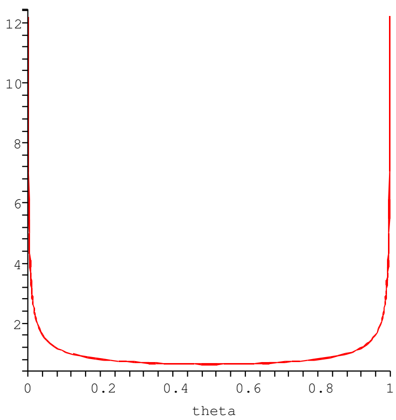


Beta distribution (1/2)

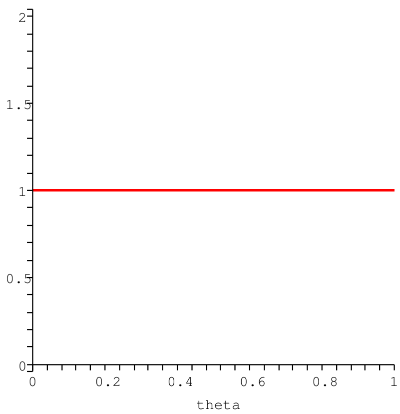
Definition 3. Beta distribution has density

$$\beta_{a,b}(x) := \frac{1}{\beta(a, b)} x^{a-1}(1-x)^{b-1}$$

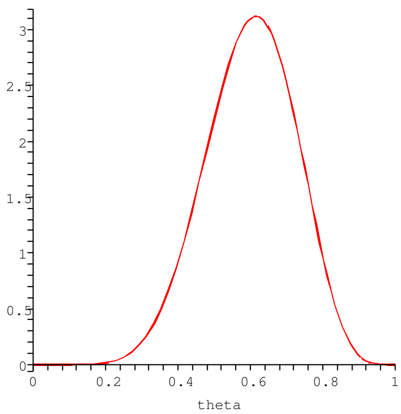
defined on $[0, 1]$.



$\beta_{.5,.5}$



$\beta_{1,1}$



$\beta_{9,6}$

Beta distribution (2/2)

Lemma 4.

$$E(\beta_{a,b}(\theta)) = \frac{a}{a+b}$$

Proof.

$$\begin{aligned} E(\beta_{a,b}(\theta)) &= \int_0^1 \theta \beta_{a,b}(\theta) d\theta \\ &= \int_0^1 \frac{1}{\beta(a,b)} \theta\, \theta^{a-1}(1-\theta)^{b-1} d\theta \\ &= \frac{\beta(a+1,b)}{\beta(a,b)} \int_0^1 \beta_{a+1,b}(\theta) d\theta \\ &= \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(a)\Gamma(b)} \\ &= \frac{a}{a+b} \end{aligned}$$

□

Lemma 5 (beta is conjungated prior for binomial samples). *For a beta prior, the a posterior again is beta:*

$$p(\theta \mid d) = \beta_{s+a,t+b}(\theta)$$

for $p_{prior}(\theta) = \beta_{a,b}(\theta)$ *and* $s := |\{x \in d \mid x = H\}|$ *and* $t := |\{x \in d \mid x = T\}|$.

Proof.

$$p(\theta \mid d) = \frac{p(d \mid \theta) p(\theta)}{p(d)}$$

$$\begin{aligned} p(d \mid \theta) p(\theta) &= \theta^s (1-\theta)^t p(\theta) \\ &= \theta^s (1-\theta)^t \beta_{a,b}(\theta) \\ &= \theta^s (1-\theta)^t \frac{1}{\beta(a,b)} \theta^{a-1} (1-\theta)^{b-1} \\ &= \frac{\beta(s+a,t+b)}{\beta(a,b)} \beta_{s+a,t+b}(\theta) \end{aligned}$$

$$\begin{aligned} p(d) &= E(\theta^s(1-\theta)^t) \\ &= \int \theta^s (1-\theta)^t d\theta \\ &= \int_0^1 \theta^s (1-\theta)^t \frac{1}{\beta(a,b)} \theta^a (1-\theta)^b dx \\ &= \frac{\beta(s+a,t+b)}{\beta(a,b)} \int_0^1 \beta_{s+a,t+b}(x) dx \\ &= \frac{\beta(s+a,t+b)}{\beta(a,b)} \end{aligned}$$

$$p(\theta \mid d) = \frac{p(d \mid \theta)p(\theta)}{p(d)} = \beta_{s+a,t+b}(\theta)$$

Summary

If we choose a beta distribution as prior, i.e.,

$$p(\theta) = \beta_{a,b}(\theta)$$

then we can compute the a posterior analytically:

$$p(\theta \mid d) = \beta_{s+a,t+b}(\theta)$$

and by taking expectations, compute parameter values also analytically:

$$\hat{\theta}_{\text{MAP}} = E(p(\theta \mid d)) = E(\beta_{s+a,t+b}(\theta)) = \frac{s+a}{s+a+t+b}$$

Advanced AI Techniques / 2. Bayesian Parameter Estimates / One Variable
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

A closer look at

$$\hat{\theta}_{\text{MAP}} = \frac{s+a}{s+a+t+b}$$

With

$$\theta_{\text{prior}} := \frac{a}{a+b}, \quad n_{\text{prior}} := a+b$$

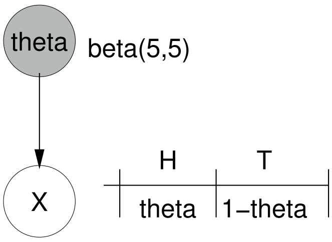
and

$$\hat{\theta}_{\text{actual}} = \frac{s}{s+t}, \quad n_{\text{actual}} = s+t$$

we have exactly

$$\hat{\theta}_{\text{MAP}} = \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}}\,\theta_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}}\,\hat{\theta}_{\text{actual}}$$

Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute of Computer Science, University of Freiburg, Germany,
Course on Advanced AI Techniques, winter term 2005



Advanced AI Techniques / 2. Bayesian Parameter Estimates / One Variable

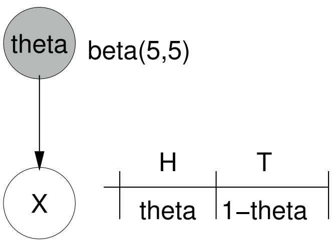
ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

DATA

H	T	H	H	H

SUFFICIENT STATISTICS

H	T
4	1



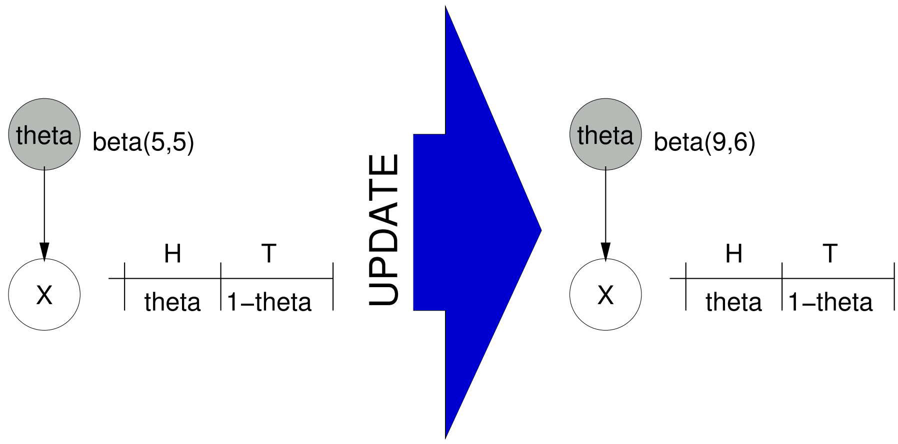
Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute of Computer Science, University of Freiburg, Germany,
Course on Advanced AI Techniques, winter term 2005 31/41

DATA

H	T	H	H	H

SUFFICIENT STATISTICS

H	T
4	1

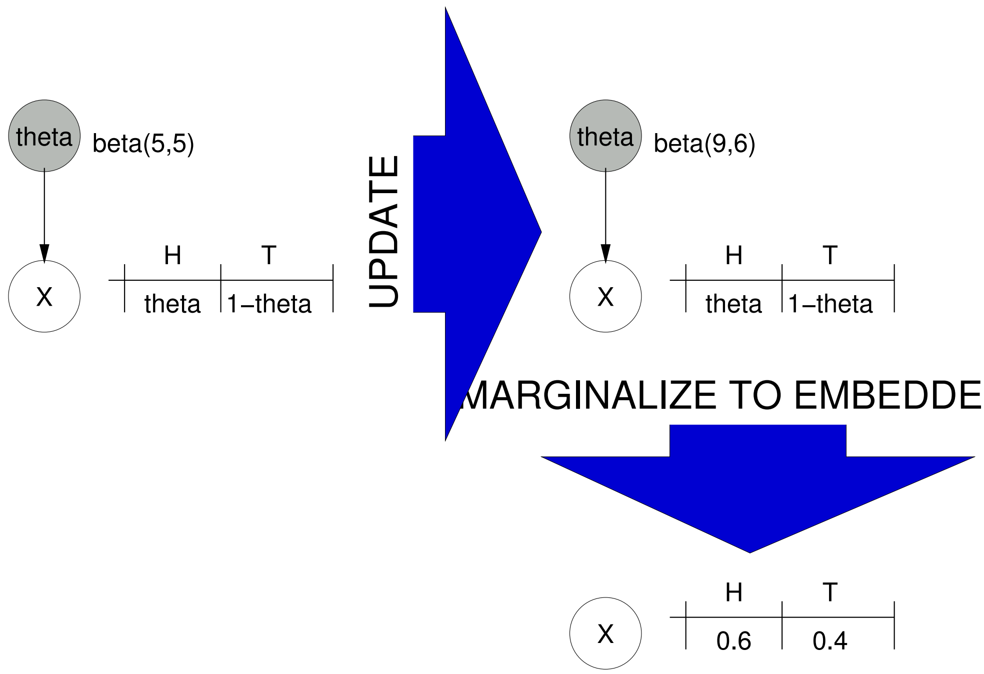


DATA

H	T	H	H	H

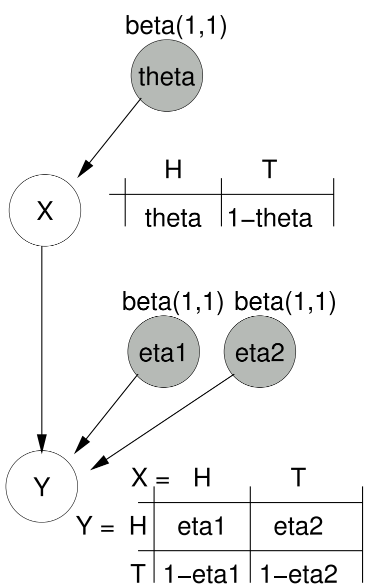
SUFFICIENT STATISTICS

H	T
4	1

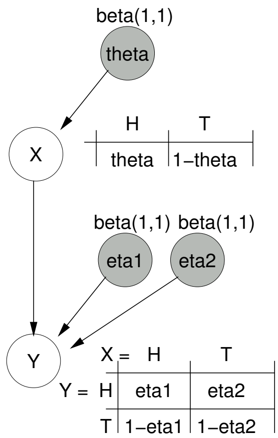


1. **Maximum Likelihood Parameter Estimates**
2. **Bayesian Parameter Estimates / One Variable**
3. **Bayesian Parameter Estimates / Several Variables**

More than one variable



More than one variable



Parameter priors are independent (as roots in a BN):

- priors for parameters of different variables (e.g., θ and $\{\eta_1, \eta_2\}$; **global parameter independence**)
- as well as priors for different parameters of the same variable (e.g., η_1 and η_2; **local parameter independence**).

More than one variable

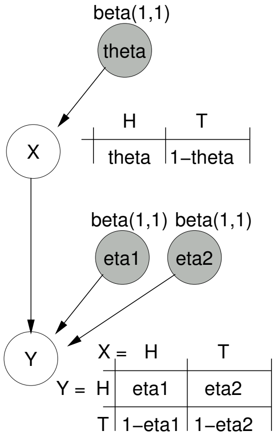
Lemma 6 (global and local parameter posterior independence).

$$p(\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{n,q_n} \,|\, d) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{i,j} \,|\, d)$$

Proof see Theorem 6.12, p. 337 of Neapolitan.

This means:

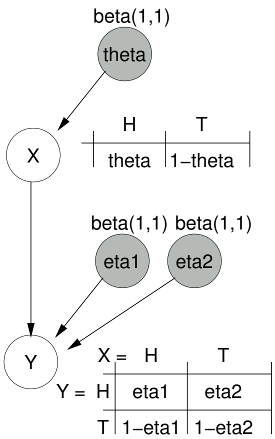
- we can compute each parameter on its own.
- same technique as for one parameter seen before.



Advanced AI Techniques / 3. Bayesian Parameter Estimates / Several Variab
ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

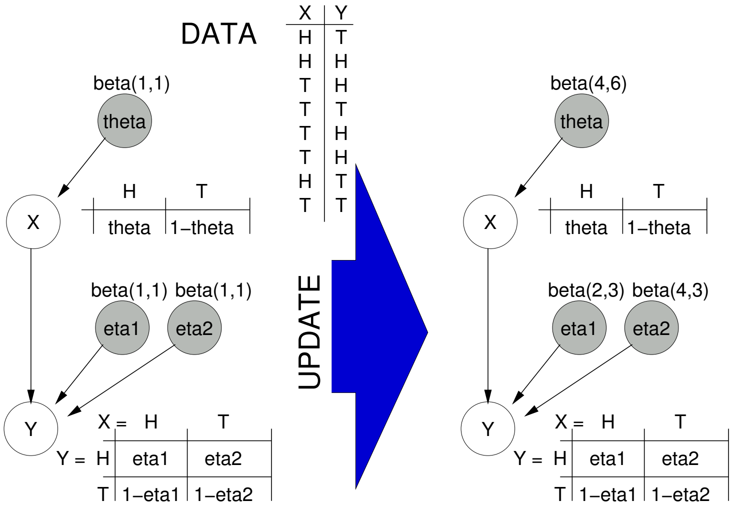
DATA

X	Y
H	T
H	H
T	H
T	T
T	H
T	H
H	T
T	T



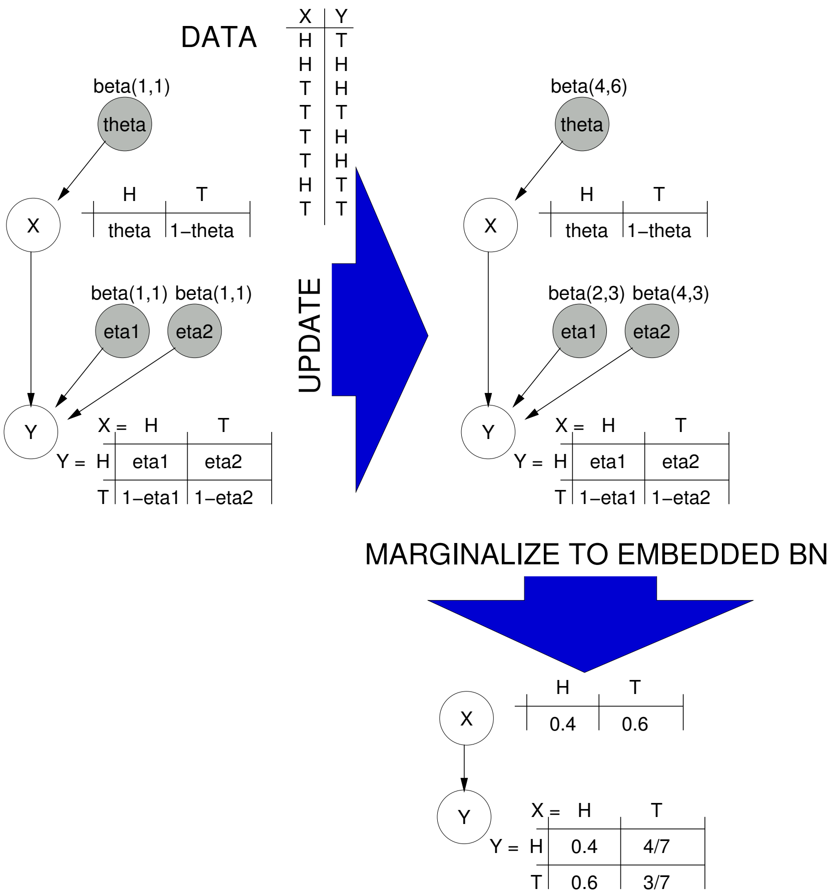
Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute of Computer Science, University of Freiburg, Germany,
Course on Advanced AI Techniques, winter term 2005 34/41

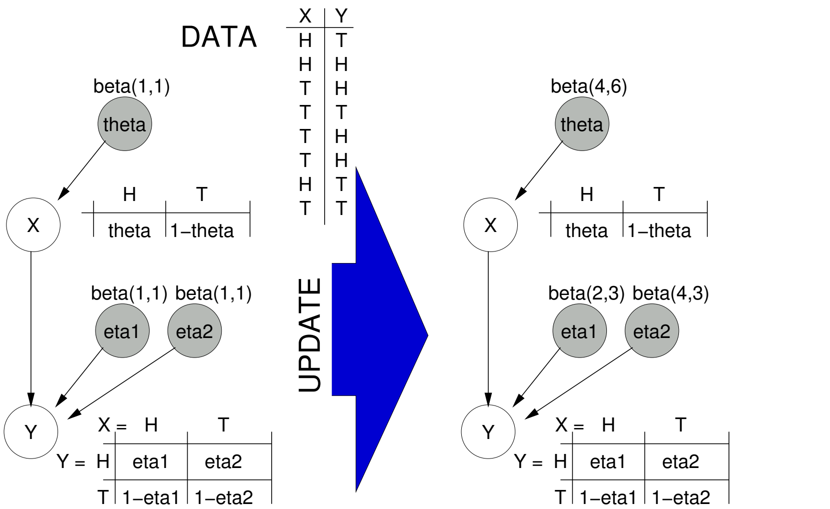
ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

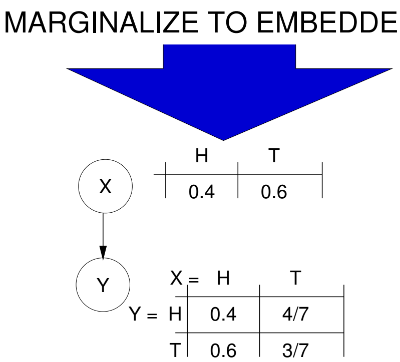


Advanced AI Techniques / 3. Bayesian Parameter Estimates / Several Variab

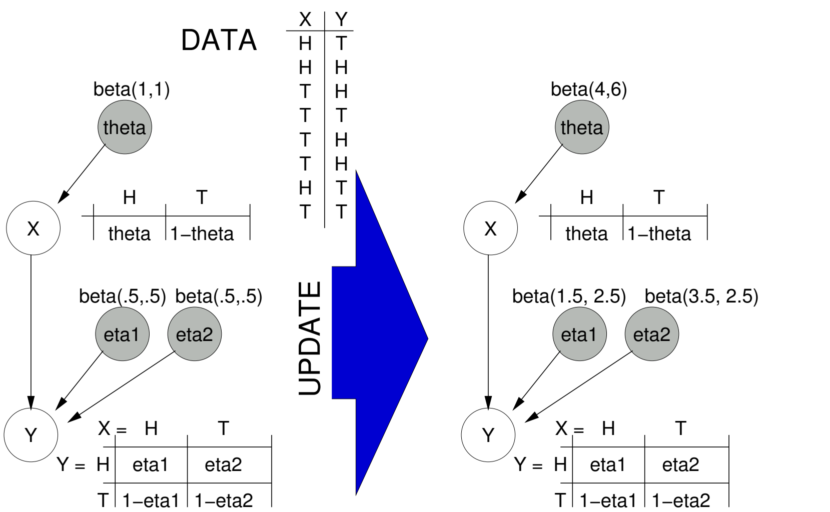
ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

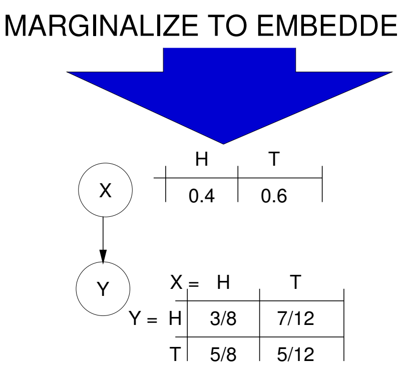






$$\hat{p}(Y = H) = 0.4 \cdot 0.4 + 0.6 \cdot \frac{4}{7} = 0.503$$





$$\hat{p}(Y = H) = 0.4 \cdot \frac{3}{8} + 0.6 \cdot \frac{7}{12} = 0.5$$

Equivalent Sample Size

Definition 4. Let $\beta_{a_{i,j},b_{i,j}}$ the priors in an augmented BN ($i = 1, \ldots, n; j = 1, \ldots, q_i$).
If there is a number N with

$$a_{i,j} + b_{i,j} = p(\mathsf{pa}_{i,j}) \cdot N$$

for all i and j, the BN is said to have **equivalent sample size** N.

If all variables are binary,
use equivalent sample size 2 to express an uninformative prior.

Multinomial Variables

So far we have looked at

- one binary variable (i.e., having two different values) and
- several binary variables

Now we look at

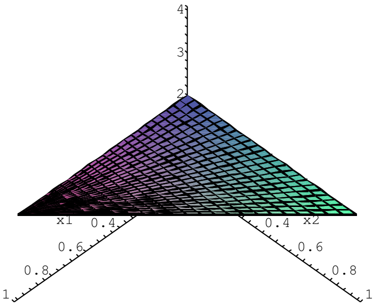
- one or several multinomial variables (i.e., having n different values)

Dirichlet distribution (1/2)

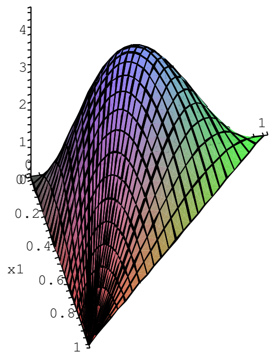
Definition 5. Dirichlet distribution has density

$$\mathrm{Dir}_{a_1,a_2,\ldots,a_n}(x_1,x_2,\ldots,x_{n-1}) := \frac{\Gamma(a_1+a_2+\ldots+a_n)}{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_n)} x_1^{a_1-1} x_2^{a_2-1} \cdots x_{n-1}^{a_{n-1}-1} (1-x_1-x_2-\ldots-x_{n-1})^{a_n-1}$$

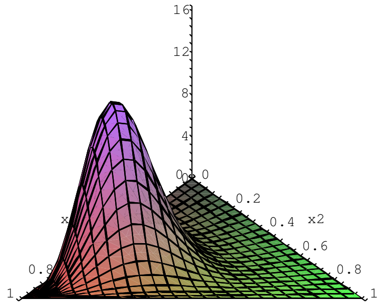
defined on $\{x \in [0,1]^{n-1} \,|\, x_1+x_2+\cdots+x_{n-1} \leq 1\}$.



$\mathrm{Dir}_{1,1,1}$



$\mathrm{Dir}_{2,2,2}$



$\mathrm{Dir}_{9,3,4}$

Dirichlet distribution (2/2)

For the special case of a binary variable ($n = 2$):

$$\mathrm{Dir}_{a_1,a_2}(x) = \beta_{a_1,a_2}(x)$$

Lemma 7. *For* $i = 1,\ldots,n$*:*

$$E_{\mathrm{Dir}_{a_1,a_2,\ldots,a_n}}(X_i) = \frac{a_i}{a_1+a_2+\cdots+a_n}$$

(where $X_n := 1 - X_1 - X_2 - \ldots - X_{n-1}$*).*

Lemma 8 (Dirichlet is conjungated prior for multinomial samples). *For a Dirichlet prior, the a posterior again is Dirichlet:*

$$p(\theta \mid d) = \mathrm{Dir}_{a_1+s_1, a_2+s_2, \ldots, a_n+s_n}(\theta)$$

for $p_{\textit{prior}}(\theta) = \mathrm{Dir}_{a_1, a_2, \ldots, a_n}(\theta)$ *and* $s_i := |\{x \in d \mid x = i\}|$ *(*$i = 1, \ldots, n$*).*

This means:

- We can compute each parameter estimate by counting, as in the binary case.

- Due to global and local posterior parameter independence, we can estimate each parameter on its own.

  $\rightsquigarrow$ same procedure as for binary variables seen before.

Summary

There a two different techniques to estimate parameters (for a Bayesian Network):

	maximum likelihood (ML)	maximum a posterior (MAP)
optimality	maximize (log-)likelihood $p(d \mid \theta)$	maximize posterior $E(p(\theta \mid d))$
result	point-estimate $\hat{\theta}$	posterior distribution $\hat{p}(\theta \mid d)$
assumption	all parameter values have same prior probability ($\hat{=}$ uninformative prior)	known prior distribution of parameter values (parameter is a random variable)
computation	can be solved analytically and boils down to relative frequencies (for complete data)	can also be solved analytically and boils down to update by relative frequencies (beta/Dirichlet distributions are conjugated priors for binomial/multinomial data)

- Priors should be choosen to have an equivalent sample size to avoid unintuitive behavior.

- In both cases, parameters can be estimated independently (global and local parameter independence).