# Probabilistic Reasoning

Russell Chap 12, 13, 14

# Uncertainty

- Why uncertainty?
  - Agent can not access the whole truth of its environment
    - Partial observability (Wumpus world)
    - Incomplete understanding (Diagnosis)

      ⇨ Use probability

- Example
  - Causal rule (cause → effect)
    - Cavity $\Rightarrow$ Toothache : True
  - Diagnostic rule (effect → cause)
    - Toothache $\Rightarrow$ Cavity : Not always True
    - Toothache $\Rightarrow$ Cavity $\vee$ Abscess $\vee$ …

      ⇨ Toothache $\Rightarrow$ Cavity : 0.85

# Probability

- Outcomes
  - An experiment produce outcomes $\omega$ (rolling a die)
- Sample space
  - S - set of all outcomes ({⚀, ⚁, …, ⚅})
- Event
  - E - subset of sample sapce (E: even = {⚁, ⚃, ⚅})
- Probability
  - $0 \leq P(\omega) \leq 1$
  - $\sum P(\omega) = 1$
  - $P(E) = |E| / |S|$ if all outcomes are equally likely
  - Ex> P(even) = 3/6

# Probability

- Random variables
  - Function: outcome $\omega$ → value $a$
    - Ex> X: outcome of rolling a die → {1, 2, … , 6}

      X(⚀)=1, X(⚁)=2, X(⚂)=3, …

      P(X=1) = 1/6
  - $P(X=a) = \sum_{(X(\omega)=a)} P(\omega)$
    - Ex> X: outcome of rolling a die → {0(even), 1(odd)}

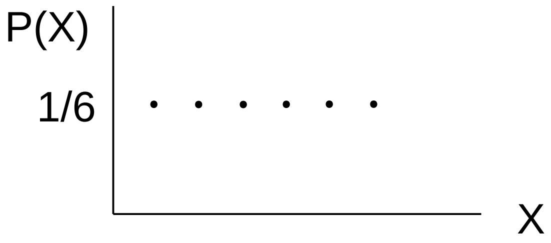
      X(⚀)=1, X(⚁)=0, E(⚂)=1, …

      P(X=0) = 1/6 + 1/6 + 1/6 = 1/2
  - Boolean r.v. : {T, F}
    - Ex> P(Cavity) = 0.1 is same as P(Cavity = True) = 0.1
  - Discrete r.v. : {a, b, c, … }
    - Ex> P(weather = sunny) = 0.1
  - Continuous r.v.
    - Ex> P(temp < 36.5) = 0.1

# Probability Distribution

- Discrete variable
  - P(X = a) : Probability for value of X = a

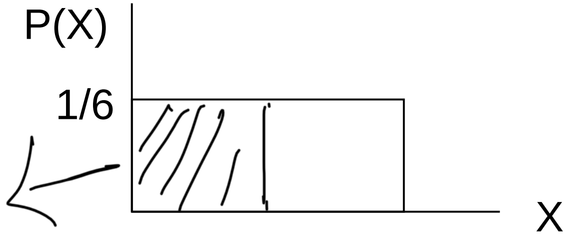
| S | X | P(X) |
|---|---|---|
| 1 | 1 | 1/6 |
| 2 | 2 | 1/6 |
| … | … | … |
| 6 | 6 | 1/6 |



- Continuous variable – probability density
  - P(X ≤ a) = probability for value of X ≤ a = area between 0 and a

$P(X \le 3)$

$\frac{1}{2}$



# Probability Distribution

- Joint probability distribution
  - P(X, Y) : represents joint probabilities for values of X and Y
  - Example
    - Gas{true,false}, Meter{empty, full}, Start{yes, no}
    - P(G,M,S):

| Gas | Meter | Start | P(G,M,S) |
|---|---|---|---|
| false | empty | no | 0.1386 |
| false | empty | yes | 0.0014 |
| false | full | no | 0.0594 |
| false | full | yes | 0.0006 |
| true | empty | no | 0.0240 |
| true | empty | yes | 0.0560 |
| true | full | no | 0.2160 |
| true | full | yes | 0.5040 |

# Probability Distribution

| Gas | Meter | Start | P(G,M,S) |
|---|---|---|---|
| false | empty | no | 0.1386 |
| false | empty | yes | 0.0014 |
| false | full | no | 0.0594 |
| false | full | yes | 0.0006 |
| true | empty | no | 0.0240 |
| true | empty | yes | 0.0560 |
| true | full | no | 0.2160 |
| true | full | yes | 0.5040 |

- P(Gas=false, Meter=empty, Start=no) = 0.1386
- P(Start=yes) = 0.5620

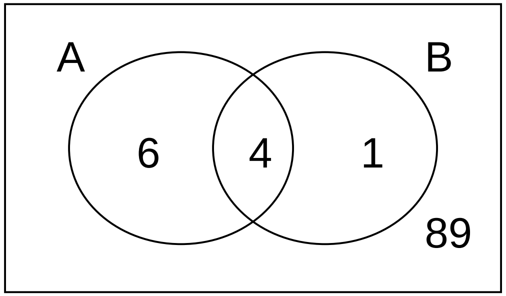
  ⇨ *Prior probability (unconditional probability)*

- P(Start=yes), given that Meter=empty ? = 0.2609

  ⇨ *Posterior probability (conditional probability)*

# Conditional Probability

- Conditional probability
  - $P(A \mid B) = P(A \wedge B) / P(B)$
  - Example
    - $P(A \wedge B) = 4 / 100 = 0.04$
    - $P(B) = 5 / 100 = 0.05$
    - $P(A \mid B) = 4 / 5 = 0.80$



- Independence
  - $P(A \mid B) = P(A)$
  - $P(A \wedge B) = P(A \mid B)\, P(B)$

    $= P(A)\, P(B)$ if A, B are independent

# Bayes' Rule

- Probability of $A \Rightarrow B$ is $P(B \mid A)$

$$\begin{aligned} P(A \wedge B) &= P(A \mid B) \cdot P(B) \\ &= P(B \mid A) \cdot P(A) \end{aligned}$$

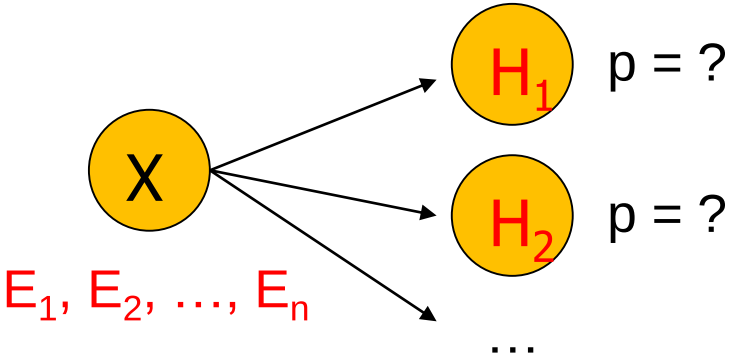
$$\therefore \quad \boxed{P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}} \longrightarrow \textit{Bayes' Rule}$$

- Assessing $P(\text{Cause} \mid \text{Effect})$ from $P(\text{Effect} \mid \text{Cause})$

$$P(\text{Cause} \mid \text{Effect}) = \frac{P(\text{Effect} \mid \text{Cause})\, P(\text{Cause})}{P(\text{Effect})}$$

# Bayesian Classifier

- Statistical classifier
  - It performs probabilistic prediction (classification)
  - From evidence E to hypothesis H



- Performance
  - Bayesian Classification provides a standard of optimal decision making
  - A simple Naïve Bayesian Classifier has comparable performance with other statistical / machine learning methods

# Bayesian Classifier

- If we want to conclude among $H_1$, $H_2$, … (classes) given evidence E,

$$P(H_1 \mid E) = \frac{P(E \mid H_1) \cdot P(H_1)}{P(E)}$$

$$P(H_2 \mid E) = \frac{P(E \mid H_2) \cdot P(H_2)}{P(E)}$$

*same*

$$\therefore \quad \boxed{P(H_i \mid E) = \alpha\, P(E \mid H_i) \cdot P(H_i)}$$

# Bayesian Classifier

- Example
  - Evidence: headache
  - Hypothesis: 1. flu ?, 2. covid-19 ?

P(headache | flu) = 0.4
P(headache | covid-19) = 0.5
P(flu) = 1/100
P(covid-19) = 1/500

$P(f \mid \text{headache}) = \alpha\, P(h \mid f)\, P(f) = \alpha \times 0.4 \times 1/100 = 0.004\, \alpha$
$P(c \mid \text{headache}) = \alpha\, P(h \mid c)\, P(c) = \alpha \times 0.5 \times 1/500 = 0.001\, \alpha$

$0.004\, \alpha + 0.001\, \alpha = 1$
⇨ Probability of flu = P(f | headache) = 0.8

# Bayesian Classifier

$E_1\ E_2 \cdots \leftarrow \begin{matrix} H_1 \\ H_2 \end{matrix}$

- Naïve Bayesian Classifier
  - For multiple, *conditionally independent* evidences

$$P(E_1, E_2, \ldots E_n \mid H_i) = P(E_1 \mid H_i) \cdot P(E_2 \mid H_i) \cdot \ldots \cdot P(E_n \mid H_i)$$

$$\therefore \boxed{\begin{aligned} & P(H_i \mid E_1, E_2, \ldots E_n) \\ & \quad = \alpha\, P(E_1, E_2, \ldots E_n \mid H_i) \cdot P(H_i) \\ & \quad = \alpha\, P(E_1 \mid H_i) \cdot P(E_2 \mid H_i) \cdot \ldots \cdot P(E_n \mid H_i) \cdot P(H_i) \\ & \quad = \alpha\, P(H_i) \cdot \prod_{k=1..n} P(E_k \mid H_i) \end{aligned}}$$

*These probabilities can be obtained from sample data*

# Example

| No. | age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|---|
| 1 | <=30 | high | no | fair | no |
| 2 | <=30 | high | no | excellent | no |
| 3 | 31…40 | high | no | fair | yes |
| 4 | >40 | medium | no | fair | yes |
| 5 | >40 | low | yes | fair | yes |
| 6 | >40 | low | yes | excellent | no |
| 7 | 31…40 | low | yes | excellent | yes |
| 8 | <=30 | medium | no | fair | no |
| 9 | <=30 | low | yes | fair | yes |
| 10 | >40 | medium | yes | fair | yes |
| 11 | <=30 | medium | yes | excellent | yes |
| 12 | 31…40 | medium | no | excellent | yes |
| 13 | 31…40 | high | yes | fair | yes |
| 14 | >40 | medium | no | excellent | no |

# Example

- $P(H_i)$
  - P(Yes) = 9/14 = 0.643
  - P(No) = 5/14 = 0.357
- $P(E_k \mid H_i)$
  - P(age = “<30” | Yes) = 2/9 = 0.222
  - P(income = “medium” | Yes) = 4/9 = 0.444
  - P(student = “yes” | Yes) = 6/9 = 0.667
  - P(credit = “fair” | Yes) = 6/9 = 0.667

  ...
  - P(age = “<30” | No) = 3/5 = 0.600
  - P(income = “medium” | No) = 2/5 = 0.400
  - P(student = “yes” | No) = 1/5 = 0.200
  - P(credit = “fair” | Yes) = 2/5 = 0.400

  ...

# Example

E X: (age="<30", income="medium", student="yes", credit="fair") → yes / no ? H

- $P(\text{yes} \mid X) = \alpha\, P(X \mid \text{yes}) \cdot P(\text{yes})$
  $= \alpha\, P(<30 \mid \text{yes}) \cdot P(m \mid \text{yes}) \cdot P(y \mid \text{yes}) \cdot P(f \mid \text{yes}) \cdot P(\text{yes})$
  $= \alpha\ 2/9 \cdot 4/9 \cdot 6/9 \cdot 6/9 \cdot 9/14 = 0.028\ \alpha$
- $P(\text{no} \mid X) = \alpha\, P(X \mid \text{no}) \cdot P(\text{no})$
  $= \alpha\, P(<30 \mid \text{no}) \cdot P(m \mid \text{no}) \cdot P(y \mid \text{no}) \cdot P(f \mid \text{no}) \cdot P(\text{no})$
  $= \alpha \cdot 3/5 \cdot 2/5 \cdot 1/5 \cdot 2/5 \cdot 5/14 = 0.007\ \alpha$

⇨ X is classified to **yes**

$$P(\text{yes} \mid X) = \frac{0.028}{0.028 + 0.007} = 0.8$$

# Laplace Correction

- Eliminate too strong probability estimation (0, 1)
  - For K distinct values for feature(evidence) Z,

$$P(Z=i)=\frac{n_i+1}{n_0+\ldots+n_k+K}$$

- Example
  - If for 100 yes class data → income : medium:90, high:10, low:0
  - P(y | age="<30", income="low", student="yes", credit="fair")

    = P(y) P(<30 | y) P(low | y) P(yes | y) P(fair | y) = 0

⇨ P(medium | y) = 91/103

P(high | y) = 11/103

P(low | y) = 1/103

# Advantage and Disadvantage

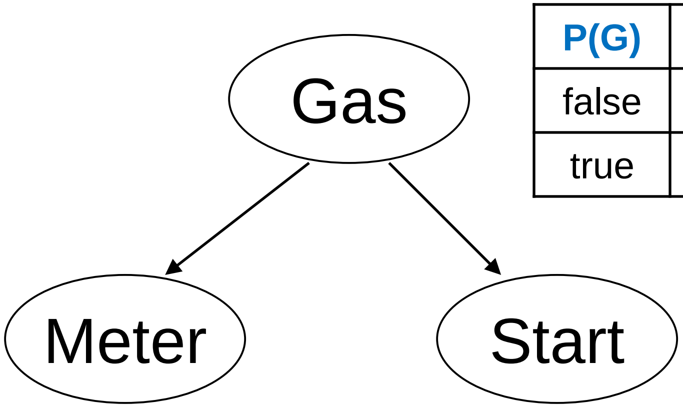
- Advantage
  - Easy to implement
  - Show good performance in most cases
- Disadvantage
  - Assume conditional independence
  - Actually there are dependencies among variables (evidences)

    Ex> ‘age’ and ‘student’

➡ *Bayesian Networks*

# Bayesian Networks

- A directed graph that represents dependencies among r.v.
  - Nodes: Random variables
  - Edges: Direct influence
  - Each node X stores P(X | parents(X))



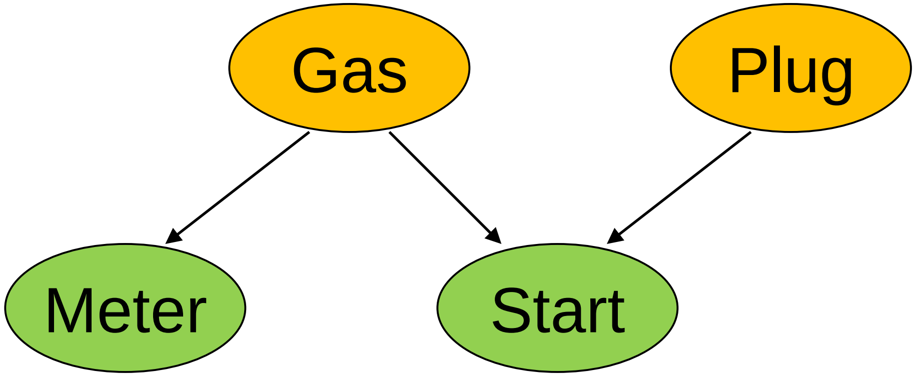
| P(G) | |
|---|---|
| false | 0.2 |
| true | 0.8 |

| P(M\|G) | G = f | G = t |
|---|---|---|
| empty | 0.7 | 0.1 |
| full | 0.3 | 0.9 |

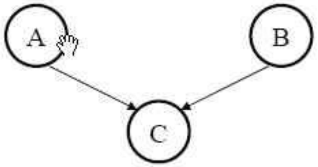
| P(S\|G) | G = f | G = t |
|---|---|---|
| no | 0.99 | 0.30 |
| yes | 0.01 | 0.70 |

# Dependency

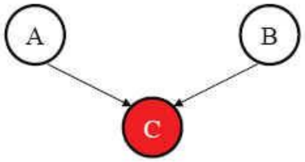
- Example
  - Gas and Plug are independent
  - Gas and Plug are conditionally dependent given Start
  - Meter and Start are dependent
  - Meter and Start are conditionally independent given Gas
    - P(S | M, G) = P(S | G)



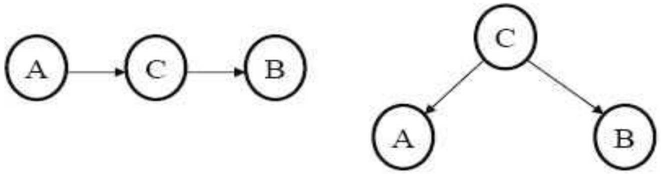
# Dependency



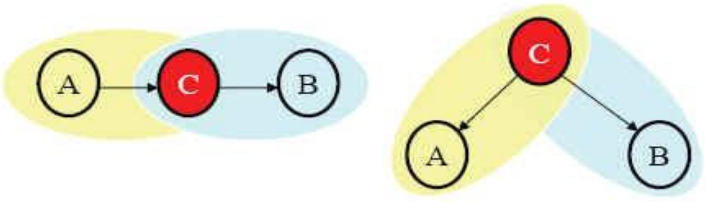
- A, B are independent
  - P(B|A) = P(B)



- A, B are *conditionally* dependent
  - P(B|A,C) ≠ P(B|C)



- A, B are dependent
  - P(B|A) ≠ P(B)



- A, B are *conditionally* independent
  - P(B|A,C) = P(B|C)

# Chain Rule

- $P(A,B) = \frac{P(A,B)}{P(A)} P(A) = P(B|A)\,P(A)$
- $P(A,B,C) = \frac{P(A,B,C)}{P(A,B)} \frac{P(A,B)}{P(A)} P(A) = P(C|B,A)\,P(B|A)\,P(A)$
- $P(A,B,C,D) = P(D|C,B,A)\,P(C|B,A)\,P(B|A)\,P(A)$

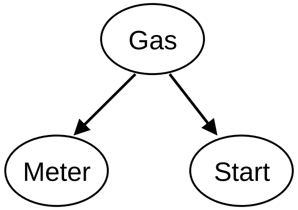
- $P(G,M,S) = P(S \mid G, M)\,P(M \mid G)\,P(G)$ (Chain rule)

  $= P(S \mid G)\,P(M \mid G)\,P(G)$

  (If S, M are conditionally independent given G)

  $= P(S)\,P(M)\,P(G)$

  (If S, M, G are all independent)

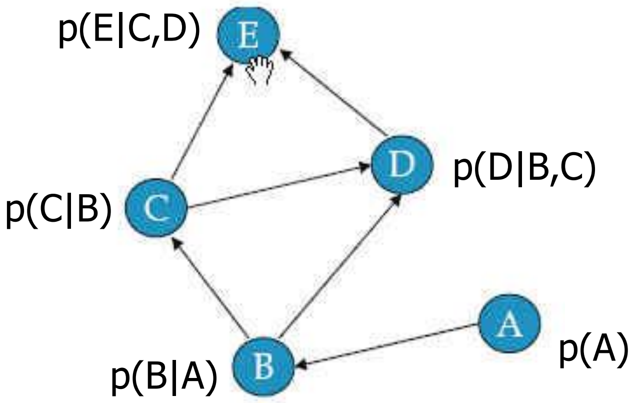


# Computing Joint Distribution

- In Bayesian networks, joint probability = *product of conditional probabilities*

$$P(X_1, X_2, \ldots , X_n) = P(X_n \mid X_1, \ldots , X_{n-1})\, P(X_{n-1} \mid X_1, \ldots , X_{n-2}) \ldots P(X_2 \mid X_1)\, P(X_1)$$
$$= \prod P(Xi \mid parents(Xi))$$

- Ex>



$$\Rightarrow\; P(A,B,C,D,E) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)P(E|A,B,C,D)$$
$$= P(A)P(B|A)P(C|B) \quad P(D|B,C) \quad P(E|C,D)$$

# Computing Joint Distribution

- Example

| Gas | Meter | Start | P(G,M,S) |
|---|---|---|---|
| false | empty | no | 0.1386 |
| false | empty | yes | 0.0014 |
| false | full | no | 0.0594 |
| false | full | yes | 0.0006 |
| true | empty | no | 0.0240 |
| true | empty | yes | 0.0560 |
| true | full | no | 0.2160 |
| true | full | yes | 0.5040 |

=

| P(G) | |
|---|---|
| false | 0.2 |
| true | 0.8 |

| P(M\|G) | G = f | G = t |
|---|---|---|
| empty | 0.7 | 0.1 |
| full | 0.3 | 0.9 |

| P(S\|G) | G = f | G = t |
|---|---|---|
| no | 0.99 | 0.30 |
| yes | 0.01 | 0.70 |

# Inference with Full Distribution

- Inference for P(X | E) with P(X, E, Y)

| Gas | Meter | Start | P(G,M,S) |
|---|---|---|---|
| false | empty | no | 0.1386 |
| false | empty | yes | 0.0014 |
| false | full | no | 0.0594 |
| false | full | yes | 0.0006 |
| true | empty | no | 0.0240 |
| true | empty | yes | 0.0560 |
| true | full | no | 0.2160 |
| true | full | yes | 0.5040 |

- P(S=yes | M=full) = P(S=yes, M=full) / P(M=full)

  = (0.5040+0.0006) / (0.5040+0.0006+0.2160+0.0594)

  = 0.6469

# Inference with Bayesian Network

- Inference for P(X | E) with P(Xi | Parents(Xi))

$$\boxed{P(X \mid E) = \alpha\, P(X, E) = \alpha \textstyle\sum_Y P(X, E, Y)}$$

From the product of probability table, P(X, E, Y)

1. Remove all rows except E
2. Compute product $\alpha$ P(X, Y | E)
3. Sum over irrelevant variables $\alpha$ P(X | E)
4. Normalize P(X | E)

# Inference with Bayesian Network

- Example - P(S=yes | M=full)

1.

| P(G) |  |
|---|---|
| false | 0.2 |
| true | 0.8 |

Remove M=empty

| P(M\|G) | G = f | G = t |
|---|---|---|
| ~~empty~~ | ~~0.7~~ | ~~0.1~~ |
| full | 0.3 | 0.9 |

| P(S\|G) | G = f | G = t |
|---|---|---|
| no | 0.99 | 0.30 |
| yes | 0.01 | 0.70 |

P(S, M, G)

2.

| S | G = f | G = t |
|---|---|---|
| no | 0.0594 | 0.2160 |
| yes | 0.0006 | 0.5040 |

Product

$\alpha$ P(S, G | M)

3.

| S |  |
|---|---|
| no | 0.2754 |
| yes | 0.5046 |

Sum over G

$\alpha$ P(S | M)

4.

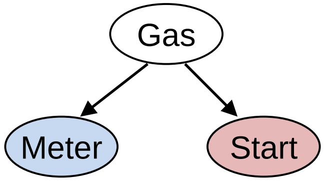
| S |  |
|---|---|
| no | 0.3531 |
| yes | **0.6469** |

Normalize

P(S | M)

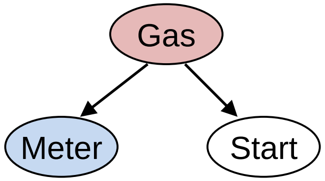
# Inference with Bayesian Network

- Let P(A=true) → P(a)
- Computing P(s | m)



$$P(s \mid m) = \alpha\, P(s, m) = \alpha\, \Sigma_G\, P(s, m, G)$$
$$= \alpha\, \Sigma_G\, P(G)\, P(s \mid G)\, P(m \mid G)$$

- Computing P(g | m)



$$P(g \mid m) = \alpha\, P(g, m) = \alpha\, \Sigma_S\, P(g, m, S)$$
$$= \alpha\, \Sigma_S\, P(g)\, P(S \mid g)\, P(m \mid g)$$
$$= \alpha\, P(g)\, P(m \mid g)\, \Sigma_S\, P(S \mid g)$$
$$= \alpha\, P(g)\, P(m \mid g)$$

⇨ *Every variable that is not an ancestor of Evidence or Query is irrelevant*

# Advantage of Bayesian Network

- Assume 20 boolean variables: 19 E → 1 H
- Compute $P(H \mid E_1, E_2, \ldots, E_{19})$

$$= \frac{P(E_1, E_2, \ldots, E_{19} \mid H)\, P(H)}{P(E_1, E_2, \ldots, E_{19})}$$

1) From full joint distribution

$\mathbf{P}(H, E_1, E_2, \ldots, E_{19})$

⇨ We need to know $2^{20}$ = 1,048,576 prob.

⇨ Almost impossible

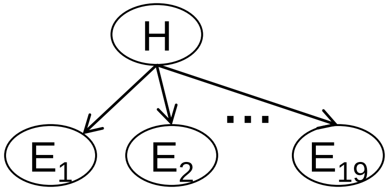
| H | E1 | E2 | … | $P(H,E_1,E_2,\ldots)$ |
|---|---|---|---|---|
| T | T | T | … | 0.xxxx |
|  |  |  | … | … |
| F | F | F | … | 0.xxxx |

# Advantage of Bayesian Network

2) Assume complete independences (Naïve Bayesian)

$\mathbf{P}(E_1 \mid H)\ \mathbf{P}(E_2 \mid H) \ldots \mathbf{P}(E_{19} \mid H)\ \mathbf{P}(H)$

⇨ 4*19 + 2 = 78 prob.
(1 parent cond. prob. table has 4 values)



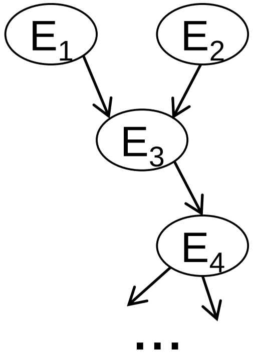
⇨ But the evidences are not actually independent

3) Consider actual dependences (Bayesian Network)

$\mathbf{P}(E_1)\ \mathbf{P}(E_2)\ \mathbf{P}(E_3 \mid E_1, E_2)\ \mathbf{P}(E_4 \mid E_3) \ldots$

(assume less than 2 parents)

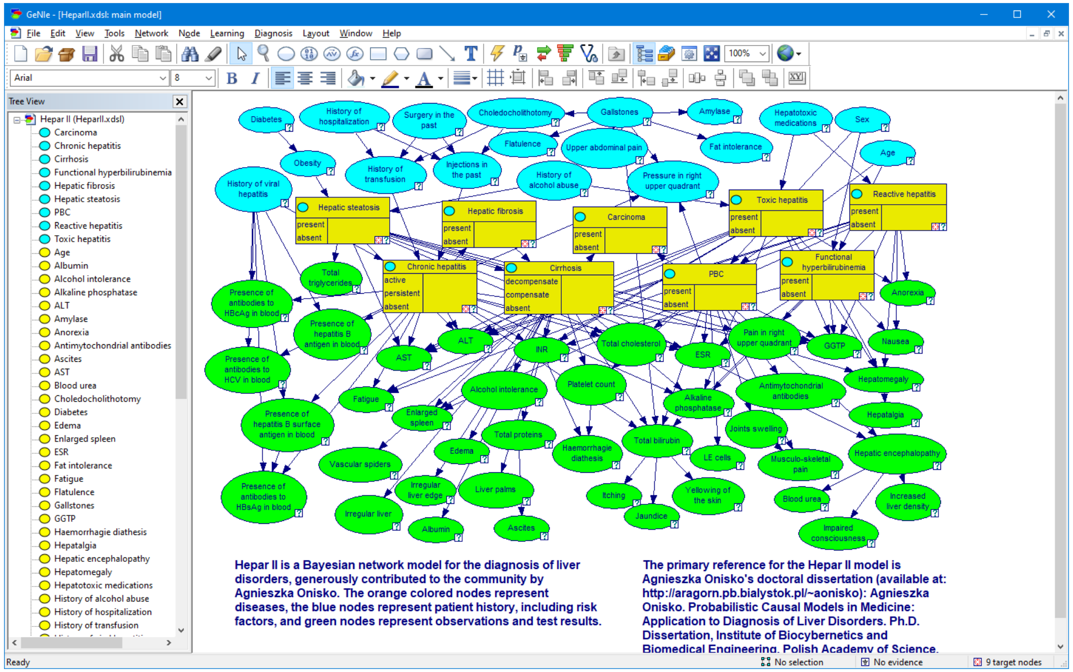
⇨ less than 8*20 = 160 prob.
(2 parent cond. prob. table has 8 values)



⇨ Actual dependences are considered,
yet need small # of probabilities.

# Example: Patient Monitoring

- GeNIe ([www.bayesfusion.com](http://www.bayesfusion.com))
  - An interactive model building, learning, and exploration tool for Bayesian networks
  - Based on the SMILE library for probabilistic graphical models



Hepar II is a Bayesian network model for the diagnosis of liver disorders, generously contributed to the community by Agnieszka Onisko. The orange colored nodes represent diseases, the blue nodes represent patient history, including risk factors, and green nodes represent observations and test results.

The primary reference for the Hepar II model is Agnieszka Onisko's doctoral dissertation (available at: http://aragorn.pb.bialystok.pl/~aonisko): Agnieszka Onisko. Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders. Ph.D. Dissertation, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science.

# Example: Patient Monitoring

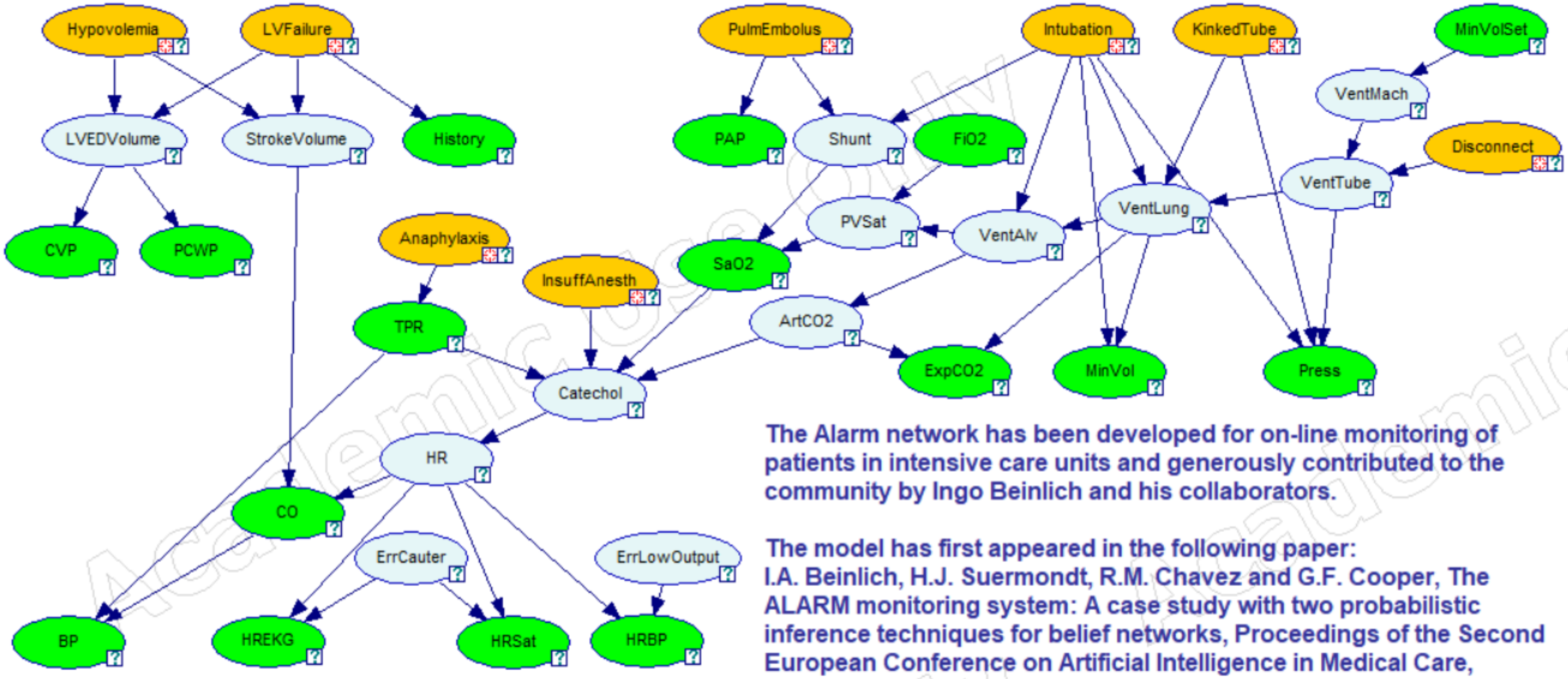
- The ALARM Bayesian network designed to provide an alarm message system for patient monitoring
- 37 variables
  - CVP (central venous pressure): a three-level factor with levels LOW, NORMAL and HIGH.
  - PCWP (pulmonary capillary wedge pressure): a three-level factor with levels LOW, NORMAL and HIGH.
  - HIST (history): a two-level factor with levels TRUE and FALSE.
  - TPR (total peripheral resistance): a three-level factor with levels LOW, NORMAL and HIGH.
  - BP (blood pressure): a three-level factor with levels LOW, NORMAL and HIGH.
  - CO (cardiac output): a three-level factor with levels LOW, NORMAL and HIGH.
  - HRBP (heart rate / blood pressure): a three-level factor with levels LOW, NORMAL and HIGH.
  - HREK (heart rate measured by an EKG monitor): a three-level factor with levels LOW, NORMAL and HIGH.
  - HRSA (heart rate / oxygen saturation): a three-level factor with levels LOW, NORMAL and HIGH.
  - PAP (pulmonary artery pressure): a three-level factor with levels LOW, NORMAL and HIGH.
  - SAO2 (arterial oxygen saturation): a three-level factor with levels LOW, NORMAL and HIGH.
  - FIO2 (fraction of inspired oxygen): a two-level factor with levels LOW and NORMAL.
  - PRSS (breathing pressure): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
  - ECO2 (expelled CO2): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
  - MINV (minimum volume): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
  - MVS (minimum volume set): a three-level factor with levels LOW, NORMAL and HIGH.

# Example: Patient Monitoring

- HYP (hypovolemia): a two-level factor with levels TRUE and FALSE. (혈량저하)
- LVF (left ventricular failure): a two-level factor with levels TRUE and FALSE. (좌심실부전)
- APL (anaphylaxis): a two-level factor with levels TRUE and FALSE. (아나필락시스)
- ANES (insufficient anesthesia/analgesia): a two-level factor with levels TRUE and FALSE. (불충분마취)
- PMB (pulmonary embolus): a two-level factor with levels TRUE and FALSE. (폐색전)
- INT (intubation): a three-level factor with levels NORMAL, ESOPHAGEAL and ONESIDED. (기도삽관)
- KINK (kinked tube): a two-level factor with levels TRUE and FALSE. (관꼬임)
- DISC (disconnection): a two-level factor with levels TRUE and FALSE. (관분리)
- LVV (left ventricular end-diastolic volume): a three-level factor with levels LOW, NORMAL and HIGH.
- STKV (stroke volume): a three-level factor with levels LOW, NORMAL and HIGH.
- CCHL (catecholamine): a two-level factor with levels NORMAL and HIGH.
- ERLO (error low output): a two-level factor with levels TRUE and FALSE.
- HR (heart rate): a three-level factor with levels LOW, NORMAL and HIGH.
- ERCA (electrocauter): a two-level factor with levels TRUE and FALSE.
- SHNT (shunt): a two-level factor with levels NORMAL and HIGH.
- PVS (pulmonary venous oxygen saturation): a three-level factor with levels LOW, NORMAL and HIGH.
- ACO2 (arterial CO2): a three-level factor with levels LOW, NORMAL and HIGH.
- VALV (pulmonary alveoli ventilation): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
- VLNG (lung ventilation): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
- VTUB (ventilation tube): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.
- VMCH (ventilation machine): a four-level factor with levels ZERO, LOW, NORMAL and HIGH.

# Example: Patient Monitoring

- The network structure



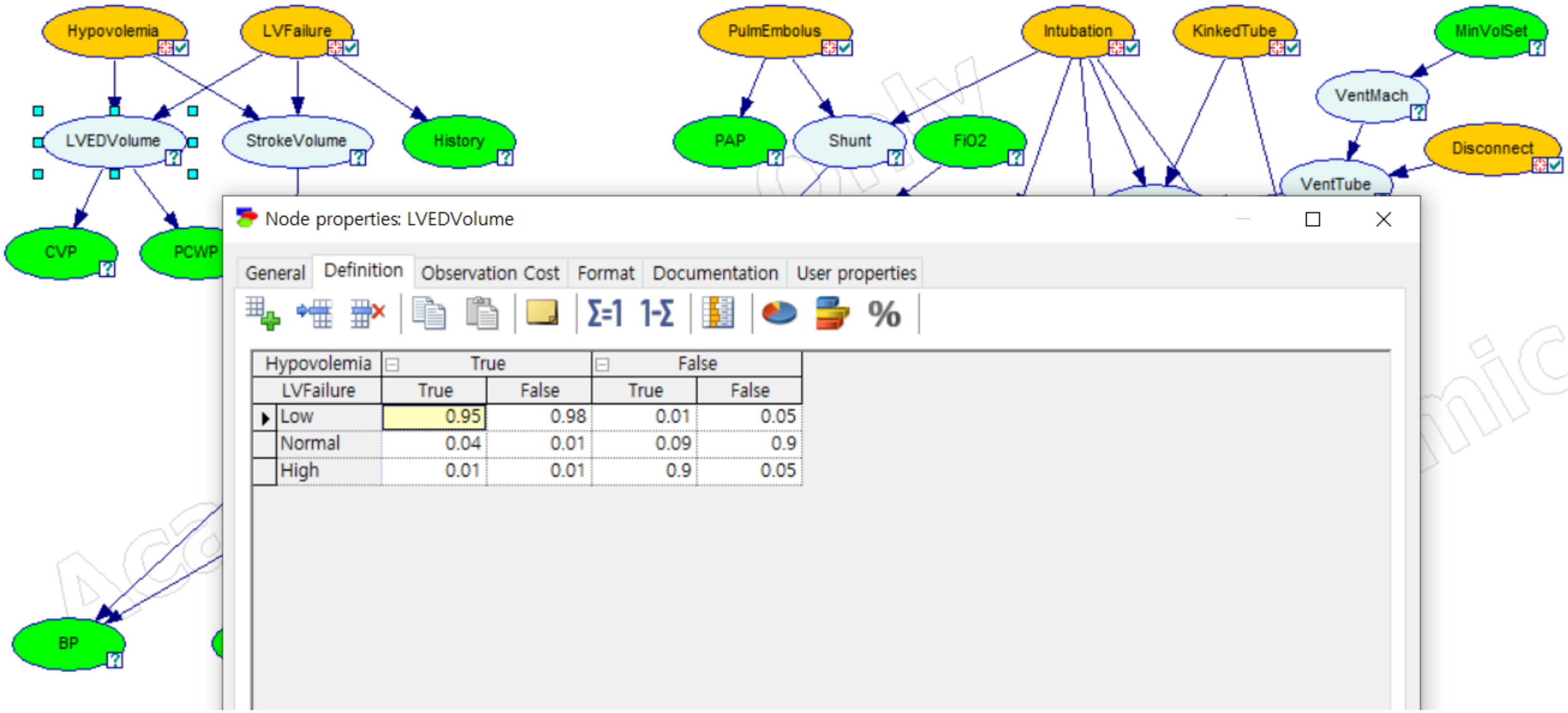
The Alarm network has been developed for on-line monitoring of patients in intensive care units and generously contributed to the community by Ingo Beinlich and his collaborators.

The model has first appeared in the following paper: I.A. Beinlich, H.J. Suermondt, R.M. Chavez and G.F. Cooper, The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks, Proceedings of the Second European Conference on Artificial Intelligence in Medical Care, pages 247-256, Springer-Verlag, Berlin, 1989

🟠 : Diagnostic node 🟢 : Measurement node

# Example: Patient Monitoring

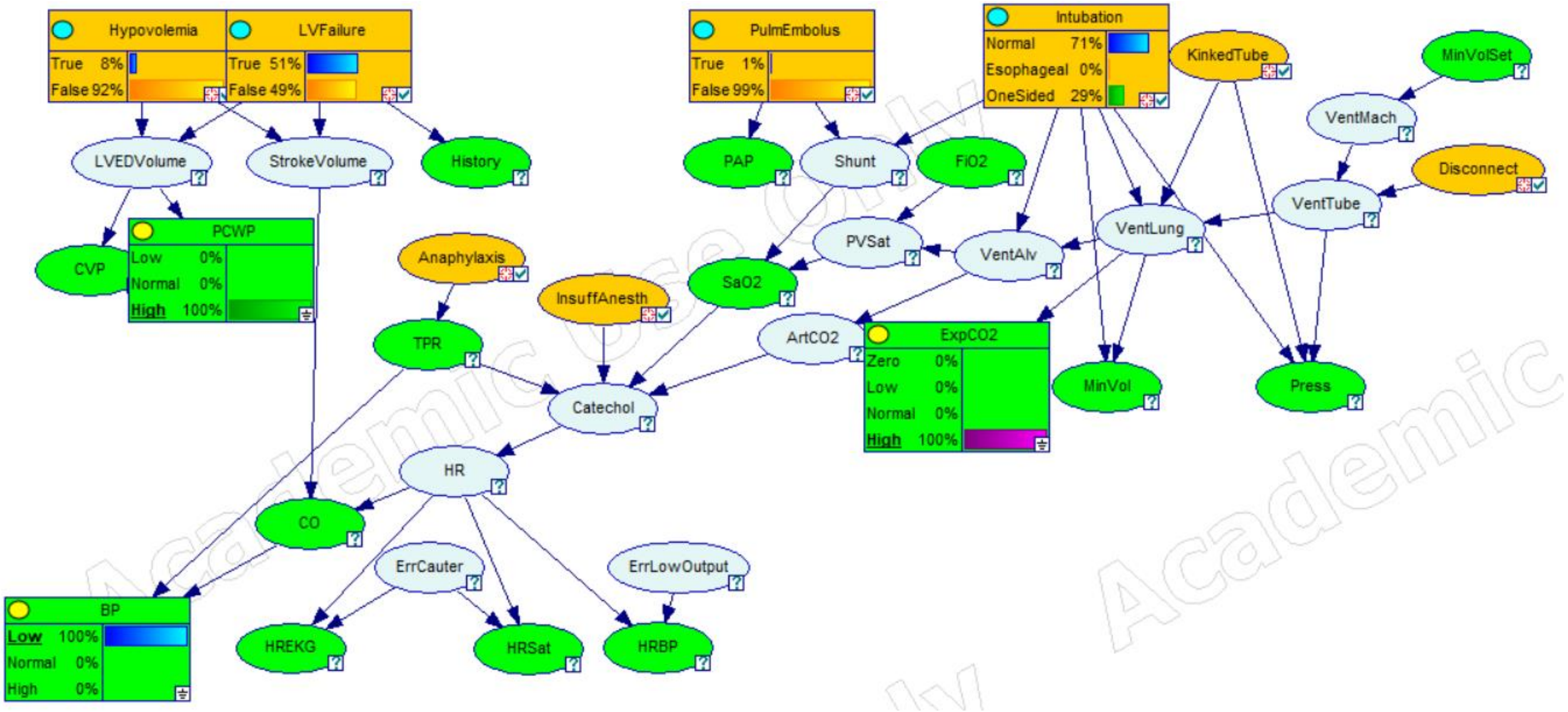
- The node probability



| Hypovolemia | True | | False | |
|---|---|---|---|---|
| LVFailure | True | False | True | False |
| Low | 0.95 | 0.98 | 0.01 | 0.05 |
| Normal | 0.04 | 0.01 | 0.09 | 0.9 |
| High | 0.01 | 0.01 | 0.9 | 0.05 |

# Example: Patient Monitoring

- Inference based on evidences



🟠 : Diagnostic node

🟢 : Measurement node

# Time and Uncertainty

- Inference in static situation
  - Given evidences about a patient → infer the patient state
  - Value of the r.v. remains fixed
- Inference in dynamic situation
  - Given evidences about economy → infer the economic state
  - Given ambiguous sequence of phonemes → infer the spoken word
  - Value of the r.v. changes over time
- $X_t$
  - A state variable at time t
  - Ex> Weather = {Rain, Cloudy, Sunny} vs.

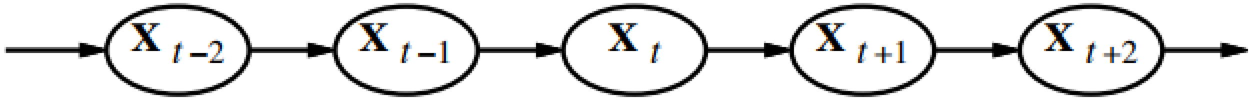
    $\text{Weather}_1$ = {Rain, Cloudy, Sunny} → $\text{Weather}_2$ → $\text{Weather}_3$ → …
  - Probability distribution: Prob. of ($X_t$ | $X_{t-1}$ , $X_{t-2}$ , $X_{t-3}$ , … )

# Markov Process

- Markov process
  - Probability of a state at time t depends on its previous n states

    $P(X_t \mid X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_{t-n})$
- First-order Markov process (Markov chain)
  - Probability of a state at time t depends on its previous 1 state

    $P(X_t \mid X_{t-1})$



- Stationary process
  - State change rule itself does not change over time
  - $P(X_t \mid X_{t-1}) = P(X_{t+k} \mid X_{t+k-1})$

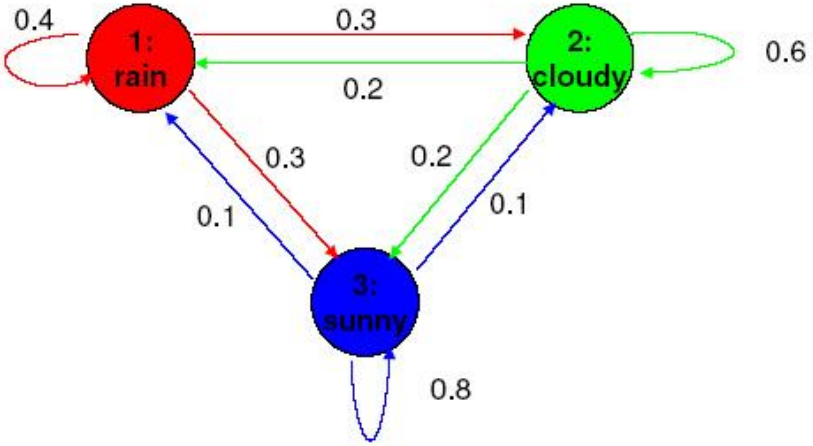
# Markov Process

- Example
  - 3 state values: R(Rain), C(Cloudy), S(Sunny)
  - Transition table and diagram for $P(X_t \mid X_{t-1})$

$X_{t+1}$

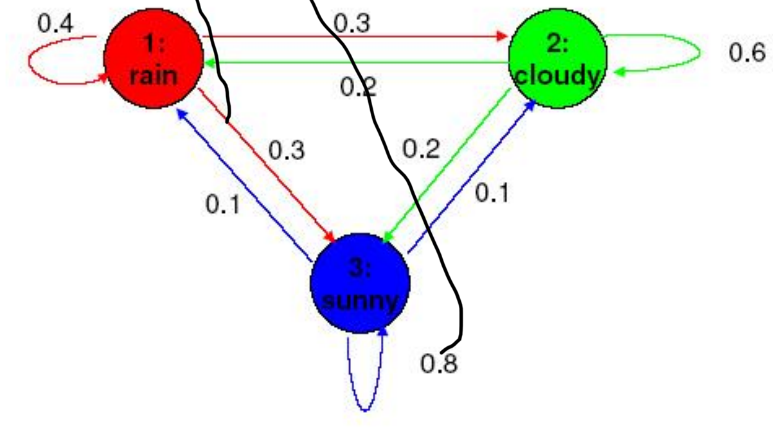
$X_t$

| | Rain | Cloudy | Sunny |
|---|---|---|---|
| Rain | 0.4 | 0.3 | 0.3 |
| Cloudy | 0.2 | 0.6 | 0.2 |
| Sunny | 0.1 | 0.1 | 0.8 |



# Markov Process

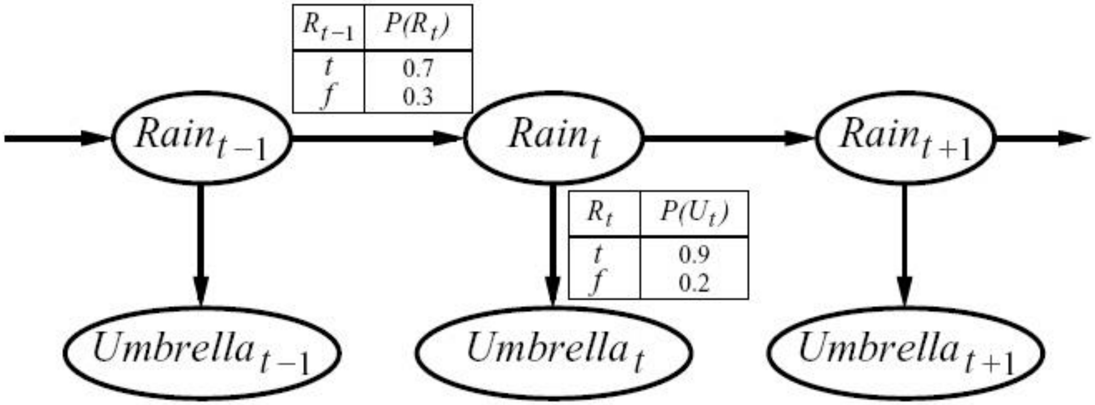
- Probability of a sequence
  - $P(X_1, X_2, \ldots, X_t)$
    $= P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid \cancel{X_1}, X_2) \ldots P(X_t \mid \cancel{X_1}, \cancel{X_2}, \cancel{\ldots}, X_{t-1})$
    $= P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_2) \ldots P(X_t \mid X_{t-1})$
  - Today is rainy. → Prob. of next 2 days are sunny, sunny?
    - $P(R, S, S) = P(R)\, P(S \mid R)\, P(S \mid S)$
      $= 1 * \underbrace{0.3}_{} * \underbrace{0.8}_{} = 0.24$



# Hidden Markov Model (HMM)

## HMM

- States are “hidden” (unobservable)
  - Transition probabilities $P(X_{t+1} \mid X_t)$ are given
- Evidences are observable
  - Probabilities of observation $P(E_t \mid X_t)$ are given
- Example: predict Rain(state) based on Umbrella(evidence)



| $R_{t-1}$ | $P(R_t)$ |
|---|---|
| $t$ | 0.7 |
| $f$ | 0.3 |

| $R_t$ | $P(U_t)$ |
|---|---|
| $t$ | 0.9 |
| $f$ | 0.2 |

# Hidden Markov Model (HMM)

- Computing most likely state sequence based on evidences

$$P(X_1 \dots X_t \mid E_1 \dots E_t)$$

$$= \frac{P(E_1 \dots E_t \mid X_1 \dots X_t)\, P(X_1 \dots X_t)}{P(E_1 \dots E_t)} \rightarrow \beta$$

$$= \boldsymbol{\alpha}\, P(E_1 \dots E_t \mid X_1 \dots X_t)\, P(X_1 \dots X_t)$$

$$= \boldsymbol{\alpha}\, P(E_1 \mid X_1) P(E_2 \mid X_2) \dots P(E_t \mid X_t)\, P(X_1 \dots X_t)$$

$$= \boldsymbol{\alpha}\, P(E_1 \mid X_1) P(E_2 \mid X_2) \dots P(E_t \mid X_t)\, P(X_1) P(X_2 \mid X_1) \dots P(X_t \mid X_{t-1})$$

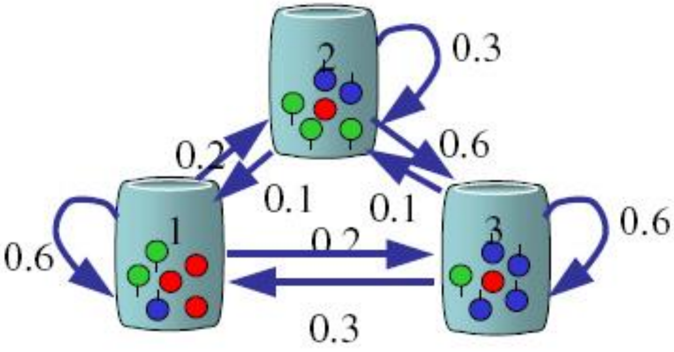
$$\boxed{= \boldsymbol{\alpha} \prod_{i=1..t} P(E_i \mid X_i)\, P(X_i \mid X_{i-1})}$$

# Example

- Markov process P($X_t$ | $X_{t-1}$)

States: S1, S2, S3

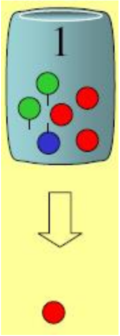
Ex> P(S2 | S1) = 0.2

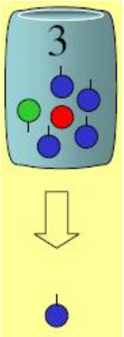


- Output process P($E_t$ | $X_t$)

Evidences: R, B, G

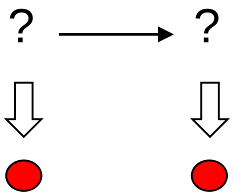
Ex> P(R | S1) = 3/6

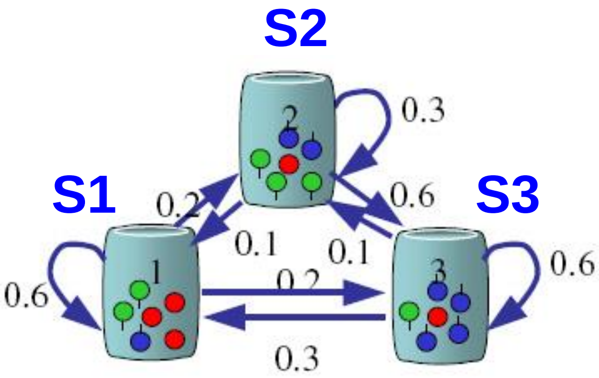




# Example

- Which cup is selected? → hidden
- Only output sequence is observed





- Most likely sequence for evidence (R, R)

$= \text{argmax}_{X}\ P(X_1, X_2, \ldots, X_t \mid E_1, E_2, \ldots, E_t)$

$= \text{argmax}_{X1,X2}\ ( \prod_{i=1..2} P(E_i \mid X_i)\, P(X_i \mid X_{i-1}) )$

$3^2$

$(1/3(\mathbf{S1}) \cdot 3/6 \cdot 0.6(\mathbf{S1}) \cdot 3/6\ ,$ 1 , 1

$1/3(\mathbf{S1}) \cdot 3/6 \cdot 0.2(\mathbf{S2}) \cdot 1/6\ ,$ 1 , 2

$1/3(\mathbf{S1}) \cdot 3/6 \cdot 0.2(\mathbf{S3}) \cdot 1/6\ ,$ 1 , 3

$1/3(\mathbf{S2}) \cdot 1/6 \cdot 0.1(\mathbf{S1}) \cdot 3/6\ ,$ ...

… ) = **S1, S1**

# Viterbi Algorithm

- Finding most likely sequence
  1. Generate all possible sequences
  2. For each sequence, calculate the probability, and pick the best one

     $N$ states, $T$ sequence → $N \text{ x } N \text{ x } \dots \text{ x } N = O(N^T)$

- The Viterbi algorithm
  1. Find the probabilities for all states $X_1$ for $E_1$

     $P(E_1 \mid X_1)\, P(X_1)$ → $\text{score}(X_1)$
  2. For each next states $X_2$ calculate

     $\max P(E_2 \mid X_2)\, P(X_2 \mid X_1)\, \text{score}(X_1)$ → $\text{score}(X_2)$

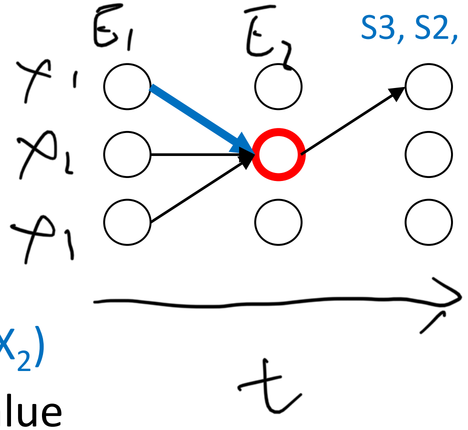
     Record the state $X_1$ which maximizes this value
  3. Repeat the above until the end of the sequence is reached

     $N$ states, $T$ sequence → $N^2 + N^2 + \dots N^2 = O(N^2 \cdot T)$

S1, S2, S1

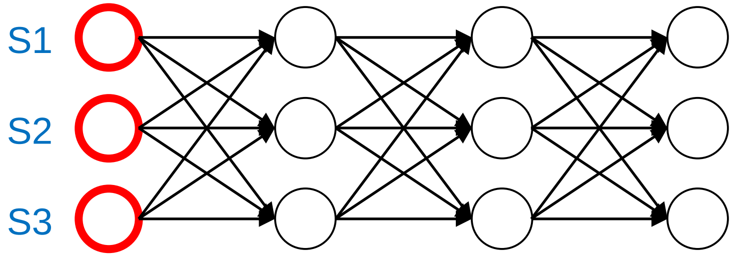
S2, S2, S1

S3, S2, S1

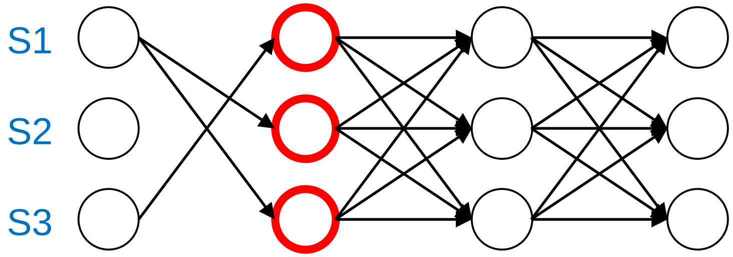


# Viterbi Algorithm

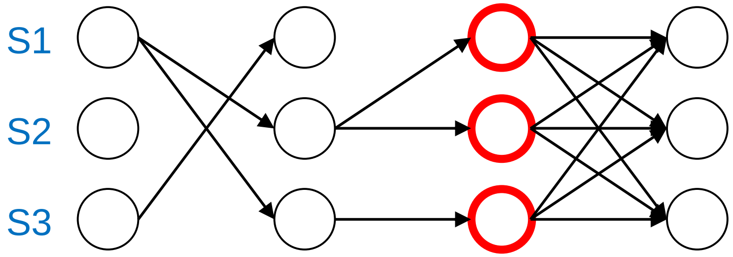
- Example
  - Observation = (R, R, G, B)



score(S1) = P(R | S1) P(S1)



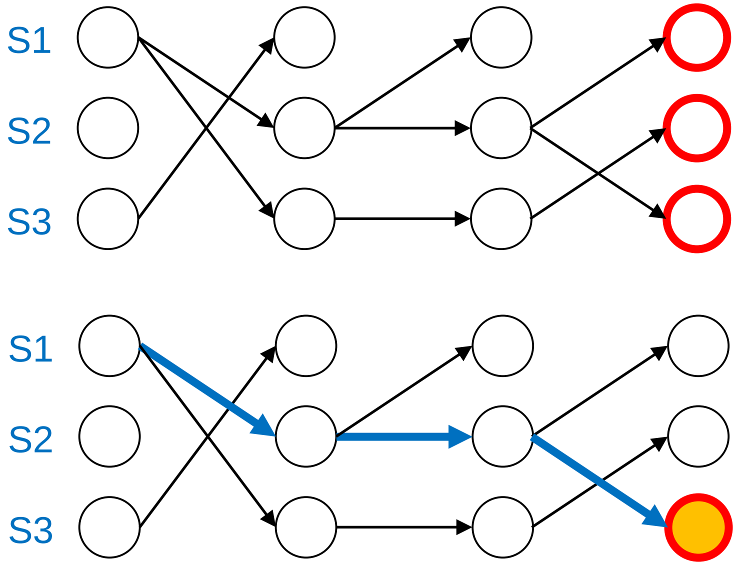
score(S1) = max P(R | S1) P(S1 | $X_1$) score($X_1$)



score(S1) = max P(G | S1) P(S1 | $X_2$) score($X_2$)

# Viterbi Algorithm

- Observation = (R, R, G, B)



score(S1) = max P(B | S1) P(S1 | $X_3$) score($X_3$)

max score

⇨ Most likely sequence = (S1, S2, S2, S3)

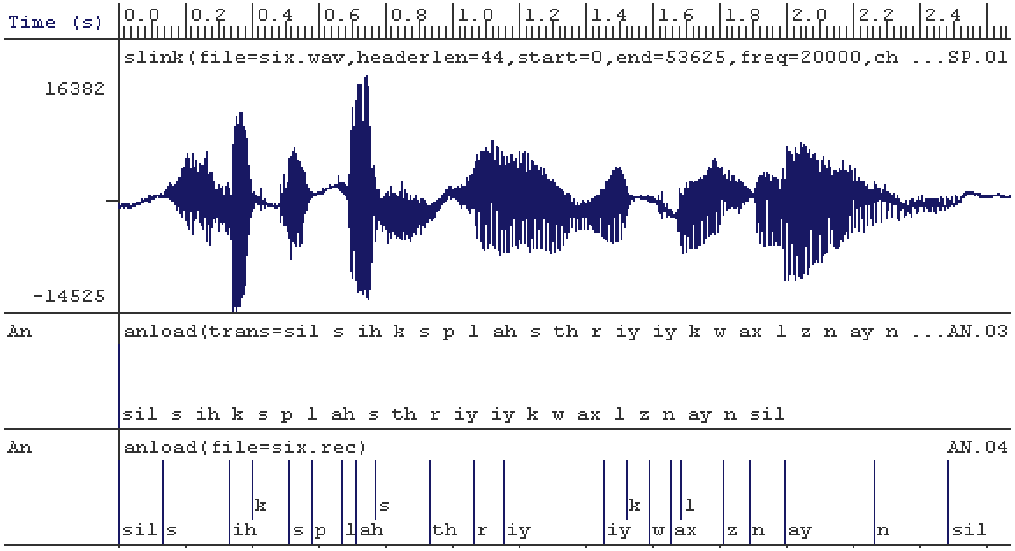
# Speech Recognition

- The problem
  - Observed: sequence of acoustic signals ($c_1 \ldots c_n$ )
  - Determine: which phoneme? → which word? - states ($p_1 \ldots p_n$)

➡ Compute P(states for a phone/word | signal) by using HMM

Find $\operatorname{argmax} P(p_1 \ldots p_n \mid c_1 \ldots c_n)$

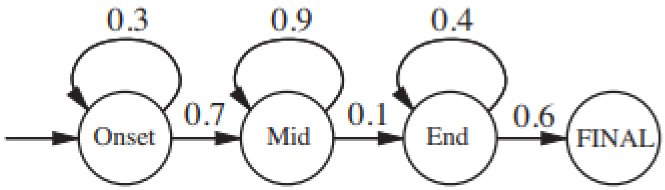
$= \operatorname{argmax} \prod_{i=1..n} P(c_i \mid p_i)\, P(p_i \mid p_{i-1})$


Time (s)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
slink(file=six.wav,headerlen=44,start=0,end=53625,freq=20000,ch ...SP.01
16382
-14525
An
anload(trans=sil s ih k s p l ah s th r iy iy k w ax l z n ay n ...AN.03
sil s ih k s p l ah s th r iy iy k w ax l z n ay n sil
An
anload(file=six.rec)
AN.04
sil
s
k
ih
s
p
l
s
ah
th
r
iy
k
iy
w
l
ax
z
n
ay
n
sil

# Speech Recognition

- Phone model
  - Phoneme : smallest unit of sound that has a distinct meaning
    - [b] (bet), [ch] (chat), [d] (dog), … [iy] (beat), [ih] (bit), [eh] (bet), …
  - Sound signal → frames (sampling 8kHz)
    → feature labels $C_1$, $C_2$, … (vector quantization)
  - Three states phone model

Phone HMM for [m]:



*state transition probability*

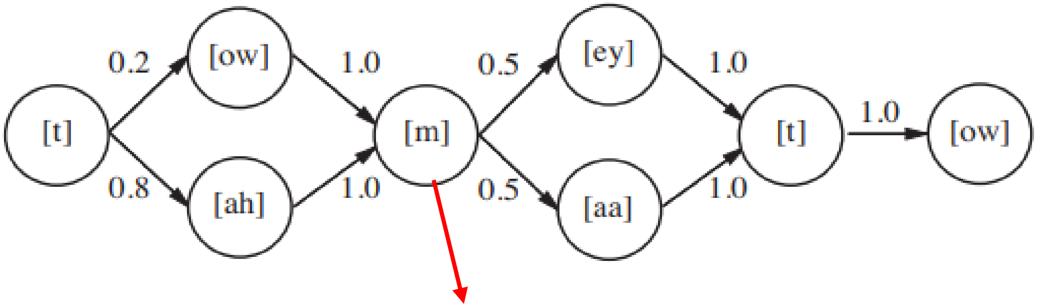
Output probabilities for the phone HMM:

| Onset: | Mid: | End: |
|---|---|---|
| $C_1$: 0.5 | $C_3$: 0.2 | $C_4$: 0.1 |
| $C_2$: 0.2 | $C_4$: 0.7 | $C_6$: 0.5 |
| $C_3$: 0.3 | $C_5$: 0.1 | $C_7$: 0.4 |

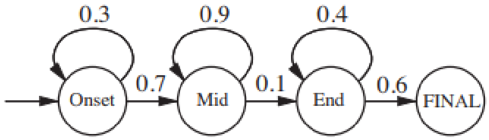
*observation probability*

# Speech Recognition

- Word model
  - Word is represented by sequence of phonemes
    - [t ow m aa t ow], [t ow m ey t ow], …
  - Word pronunciation model



Phone HMM for [m]:

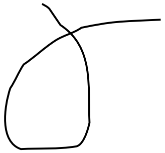


Output probabilities for the phone HMM:

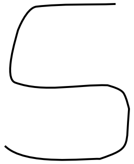
| Onset: | Mid: | End: |
|---|---|---|
| $C_1$: 0.5 | $C_3$: 0.2 | $C_4$: 0.1 |
| $C_2$: 0.2 | $C_4$: 0.7 | $C_6$: 0.5 |
| $C_3$: 0.3 | $C_5$: 0.1 | $C_7$: 0.4 |

# Handwriting Recognition

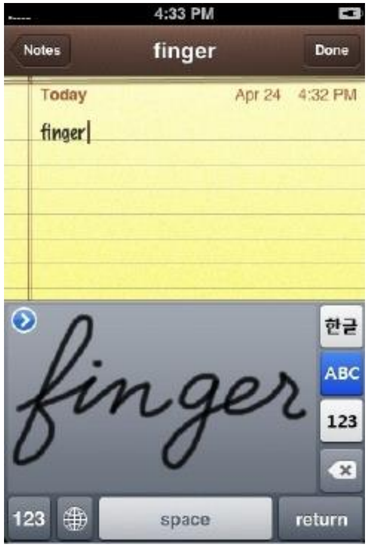
- Hand-written character recognition



α or 6 ?



5 or S ?

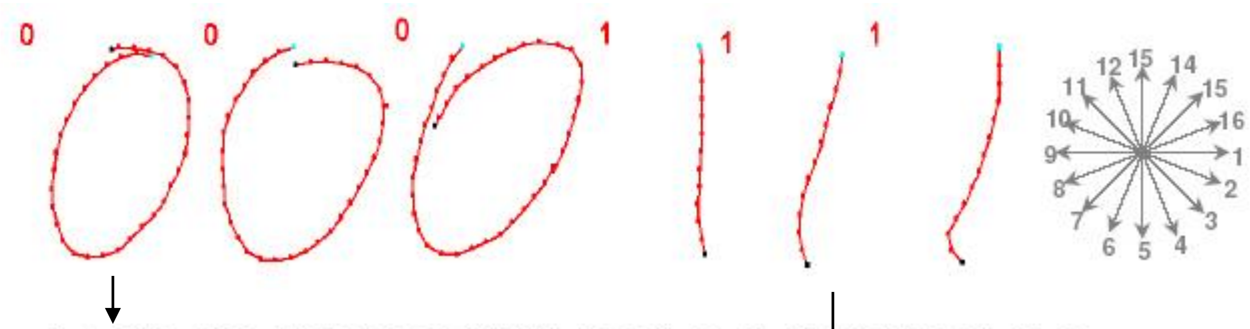


# Handwriting Recognition

- The problem
  - Observed: sequence of moving directions ($d_1$.. $d_n$ )
  - Determine: which character? - states ($s_1$.. $s_n$)

⇨ Find argmax $P(s_1 \dots s_n \mid d_1 \dots d_n)$

$= \text{argmax} \prod_{i=1..n} P(d_i \mid s_i) \, P(s_i \mid s_{i-1})$

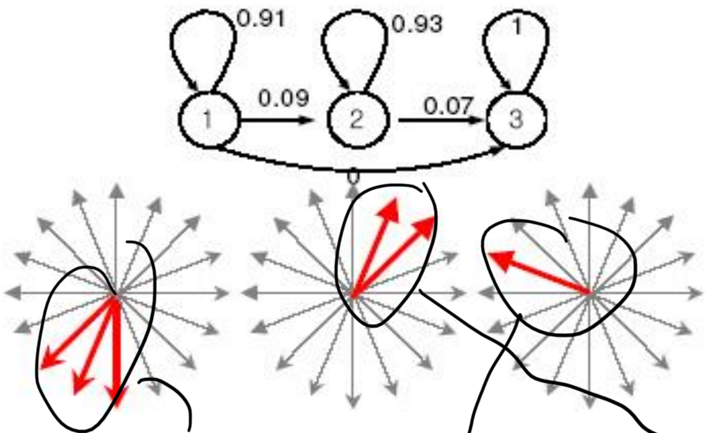


data0{1} = [9 8 8 7 7 7 6 6 6 5 5 5 4 4 3 2 1 16 15 15 15 15 14 14 14 13 13 13 12 12 11 10 9 9 8 ]

data1{2} = [5 6 6 6 6 6 6 6 6 5 5 4 ]

# Handwriting Recognition

- The character model
  - HMM for ‘0’:



| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.33 | 0.21 | 0.12 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.09 | 0.07 | 0.07 | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.28 | 0.11 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.06 | 0.16 | 0.23 | 0.13 | 0.10 | 0.10 | 0.16 | 0.00 | 0.00 | 0.00 |

*state transition probability*

*observation probability*

  - Observation: [8, 8, 7, 7, 7, 6, 6, 5, 5, … ]

⇨ P(states of ‘0’ | 8, 8, 7, 7, 7, 6, 6, 5, 5, … ) is max.