

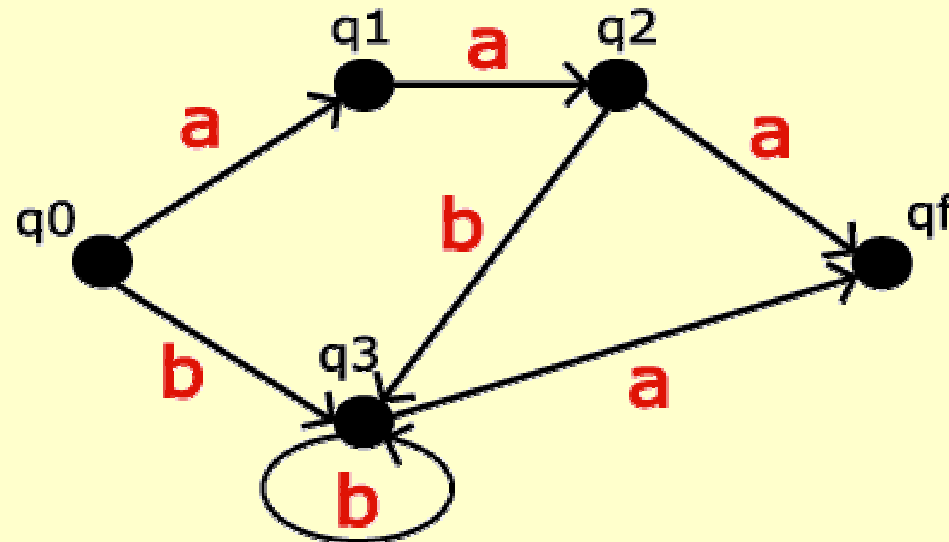
PAC-learnability of PDFFA in terms of Variation Distance

Paul W. Goldberg and Nick Palmer

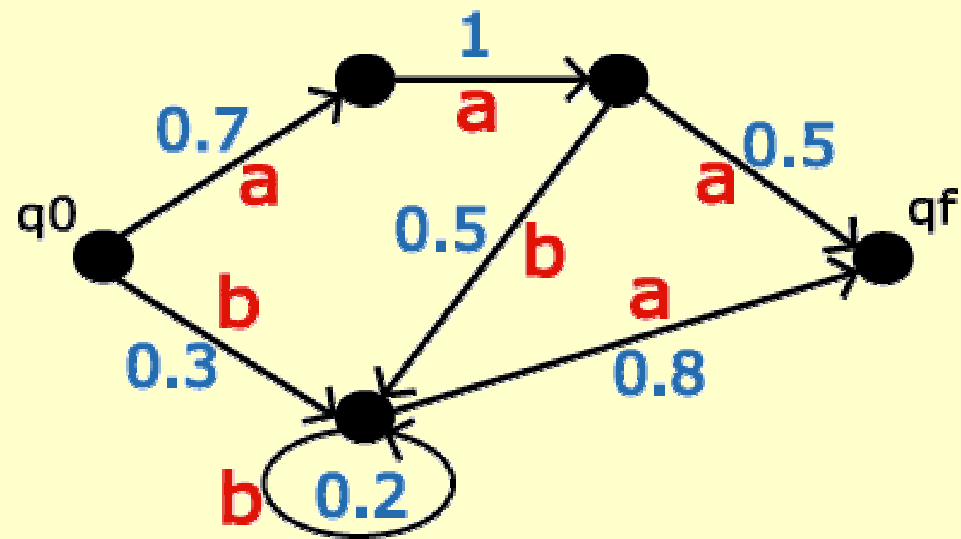
University of Warwick, UK.

PDFFA

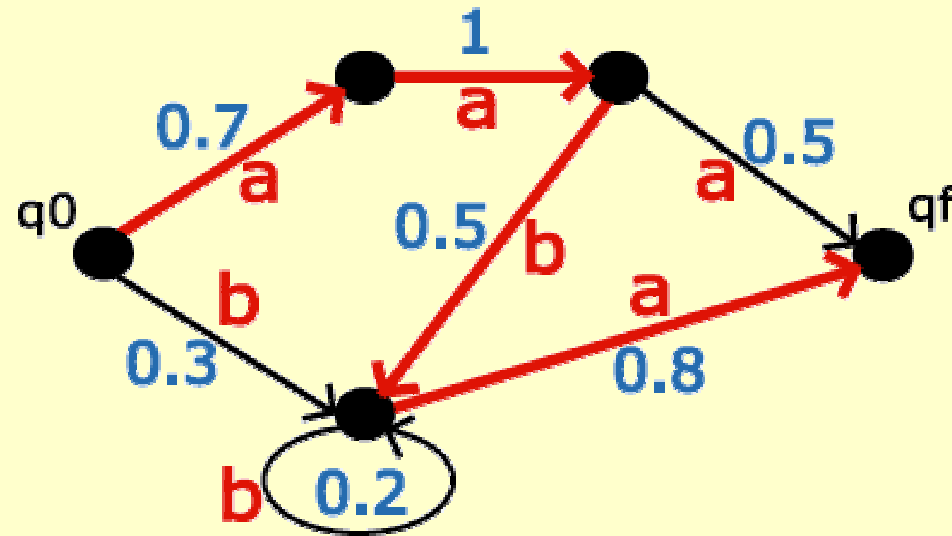
- Q is the set of all states ($|Q| = n$)
- $q_0 \in Q$ is the initial state
- $q_f \in Q$ is the final state



Strings



Strings



String : aaba

$$P(\text{aaba}) = 0.7 \times 1 \times 0.5 \times 0.8 = 0.28$$

PAC-learnability

Unsupervised Learning - positive data.

Given a sample of N strings generated by automaton A , we show that with probability $\geq 1-\delta$ we learn the distribution of A to within accuracy ϵ .

Note : N is polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}, \dots$

KL-Divergence

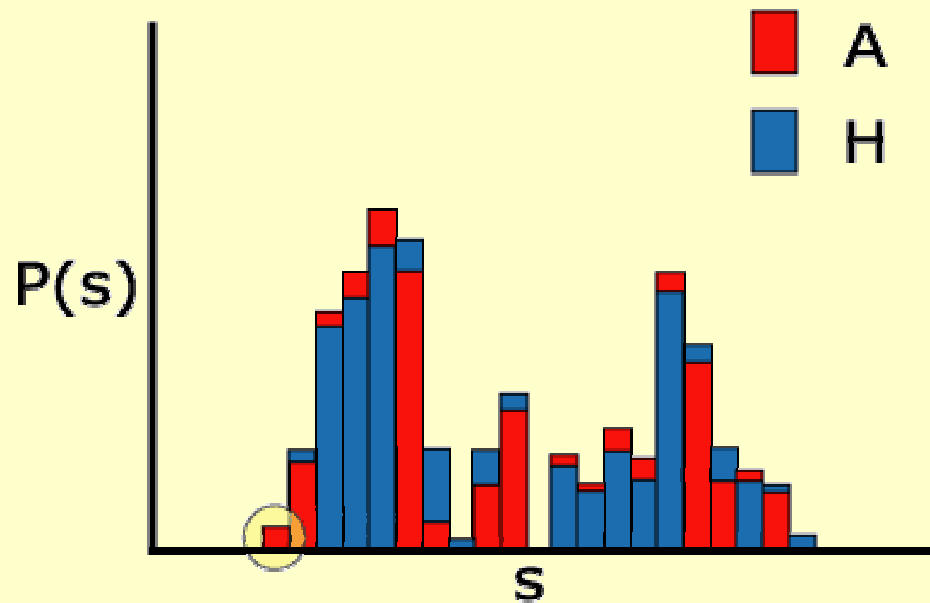
Think of KL-Divergence as being

"How surprised we are to see a particular string given our hypothesis probabilities."

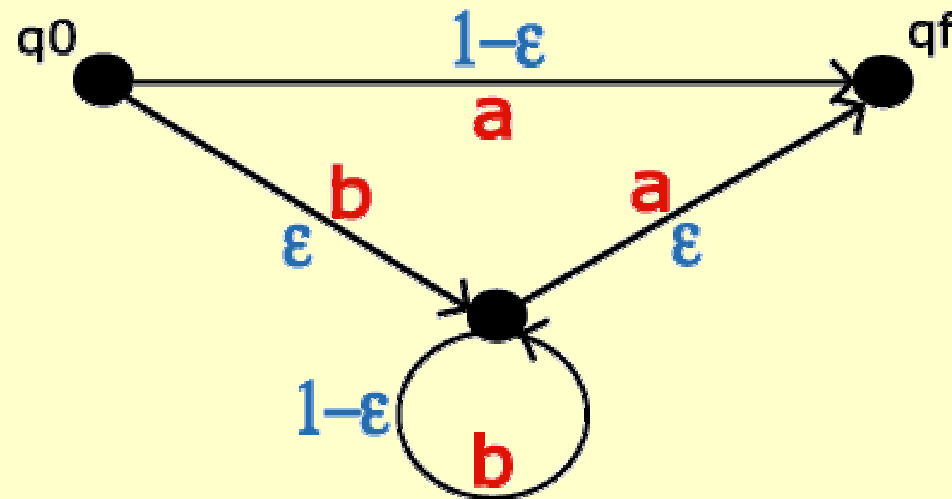
$$KL(D_A, D_H) = \sum_{s \in \Sigma^*} -D_A(s) \cdot \log \left(\frac{D_A(s)}{D_H(s)} \right)$$

Difficulties

If $D_H(s)=0$ for some string s , and $D_A(s)>0$...
this produces unbounded logs.



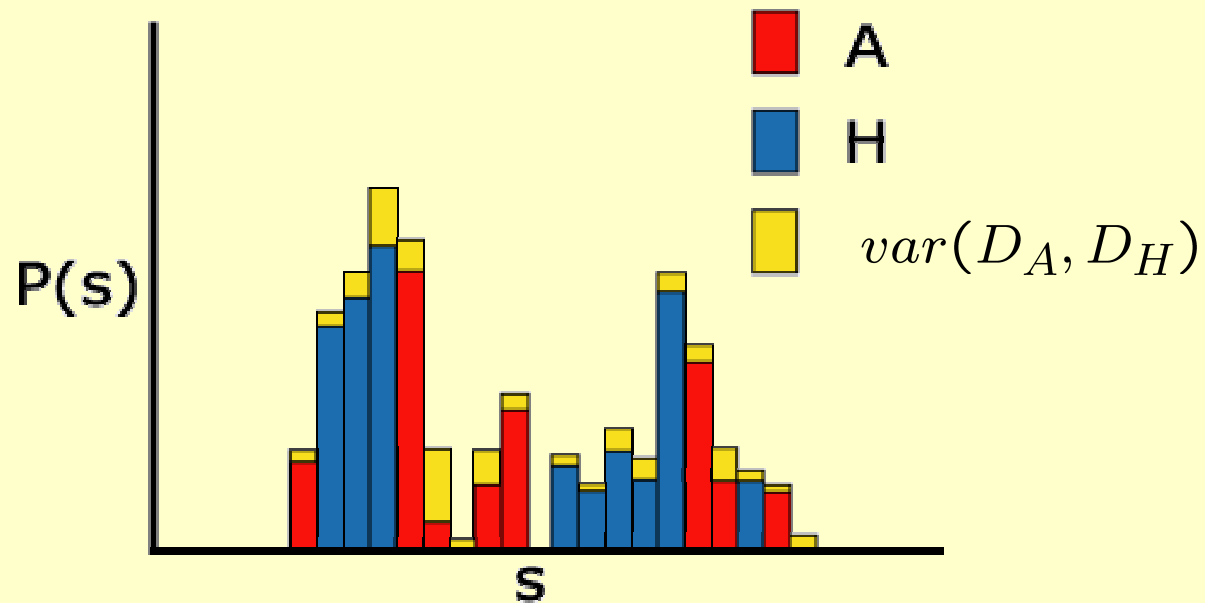
Example



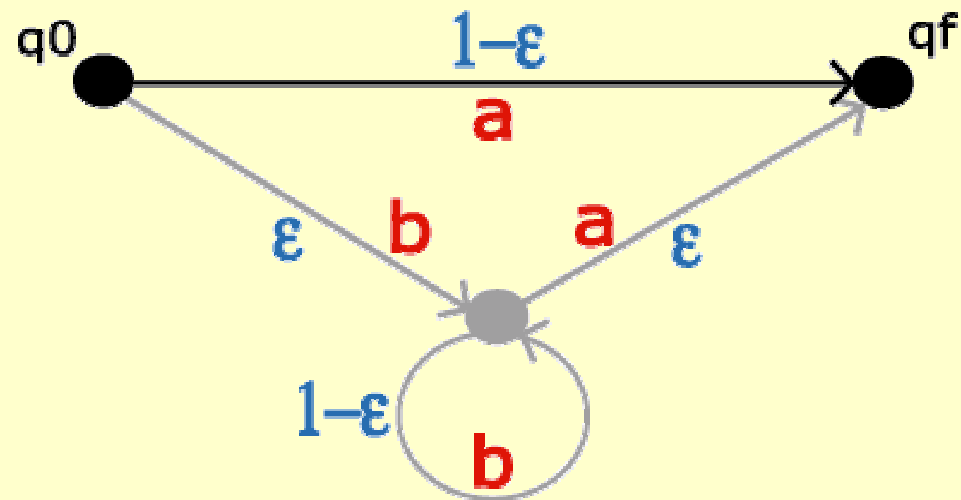
A large number of strings must be sampled in order to be sure of seeing "b...".

Variation Distance

$$\text{var}(D_A, D_H) = \sum_{s \in \Sigma^*} |D_A(s) - D_H(s)|$$



Example



Results

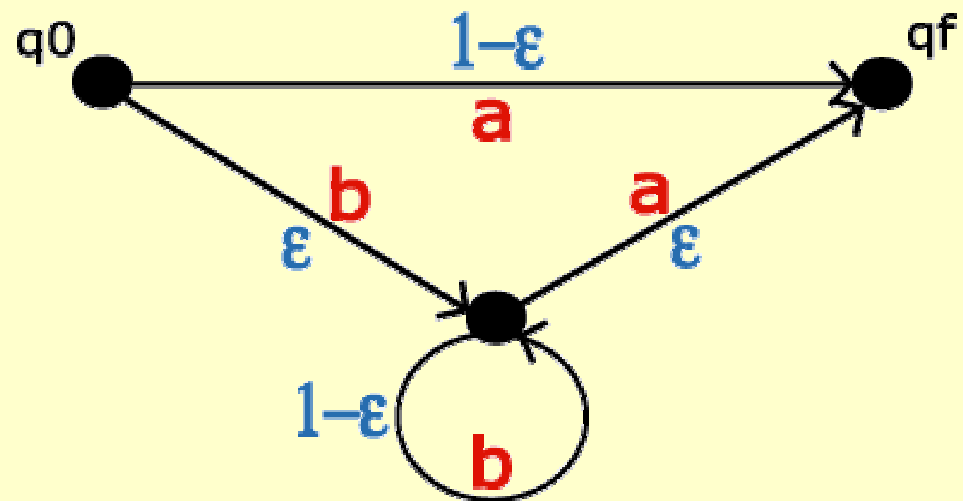
Clark and Thollard (2004)

With probability of at least $1-\delta$,

$$\text{KL}(D_A, D_H) \leq \varepsilon$$

Sample size polynomial in $1/\varepsilon$, $1/\delta$, n , $|\Sigma|$ and μ - also dependent on upper bound of the expected length of a string.

Example



Results

Palmer and Goldberg (2005)

With probability of at least $1-\delta$,

$$\text{var}(D_A, D_H) \leq \varepsilon$$

Sample size polynomial in $1/\varepsilon$, $1/\delta$, $|\Sigma|$, n
and μ .

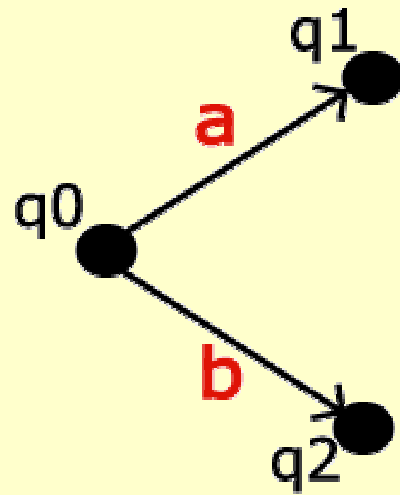
Algorithm to learn DFA

Same approach as Clark and Thollard

- Use of **distinguishability** (μ) to measure similarity of states.
- Use of **candidate nodes** to construct the hypothesis automaton.

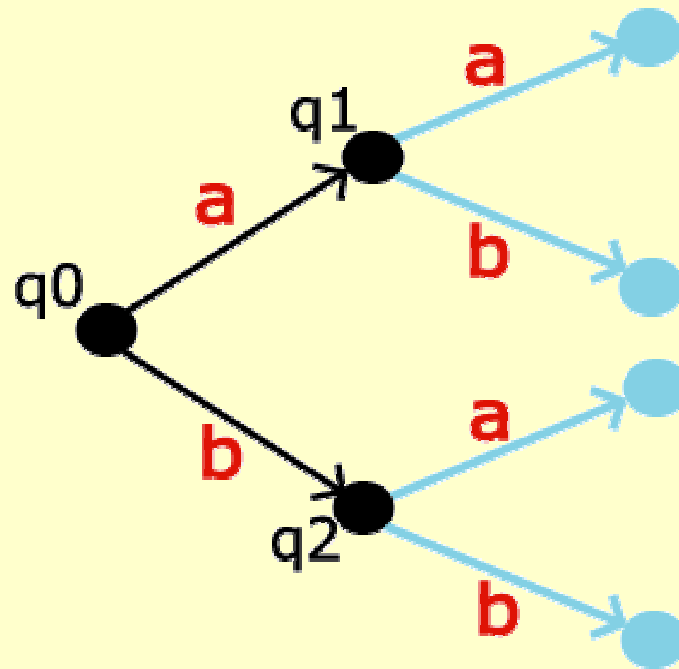
Algorithm to learn DFA

Alphabet $\Sigma = \{a, b\}$

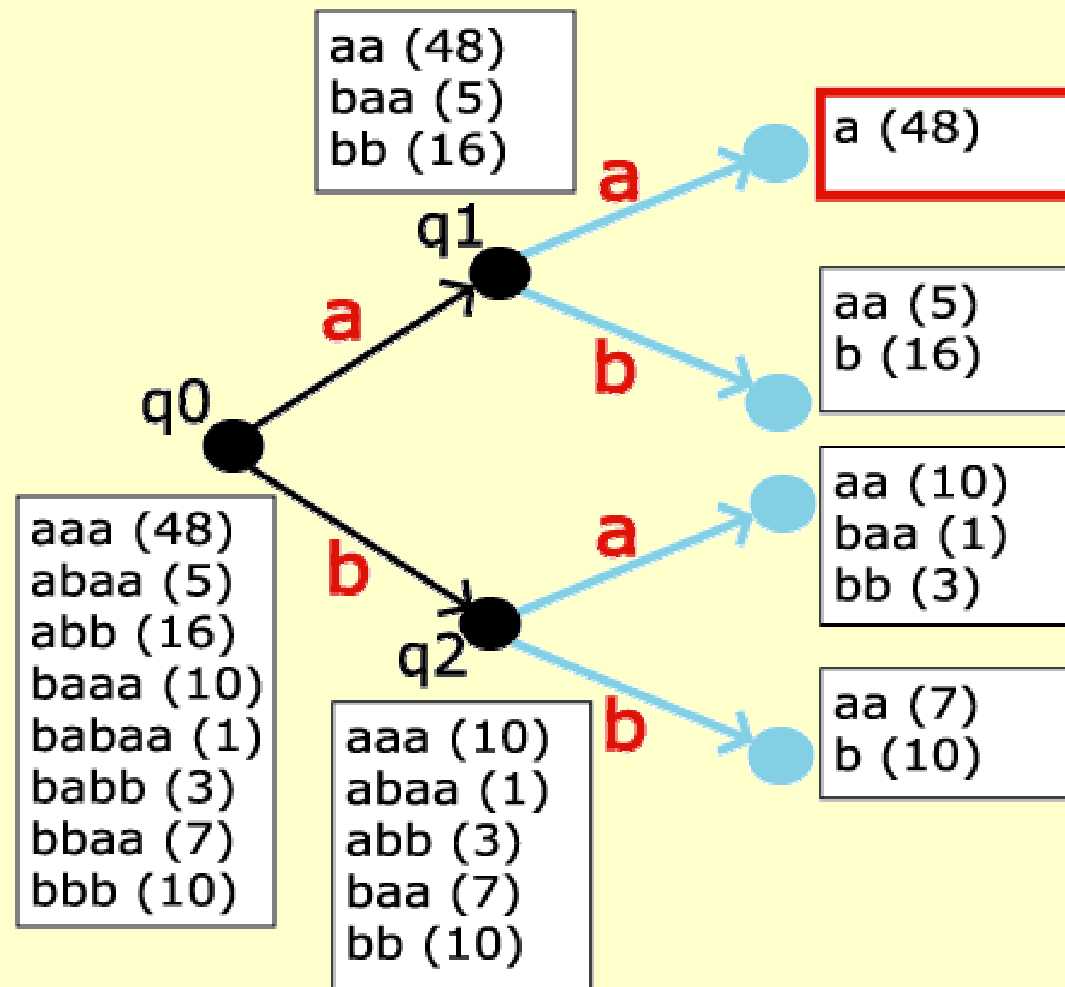


Algorithm to learn DFA

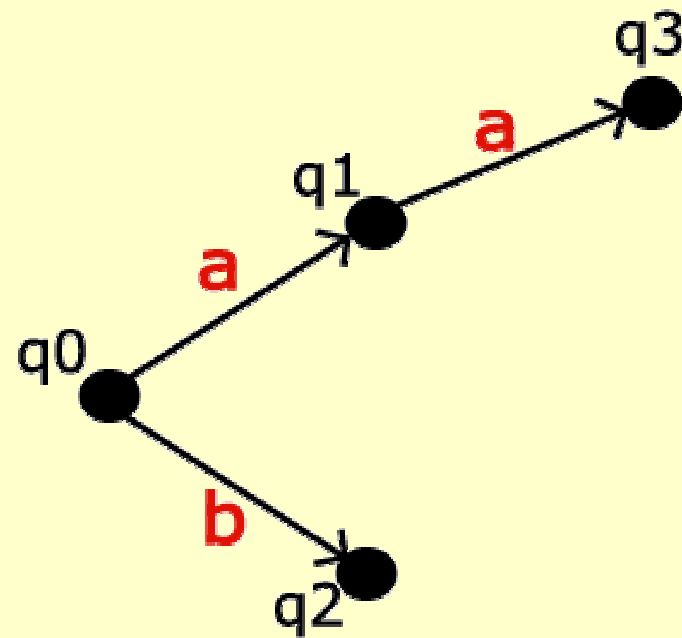
Alphabet $\Sigma = \{a,b\}$



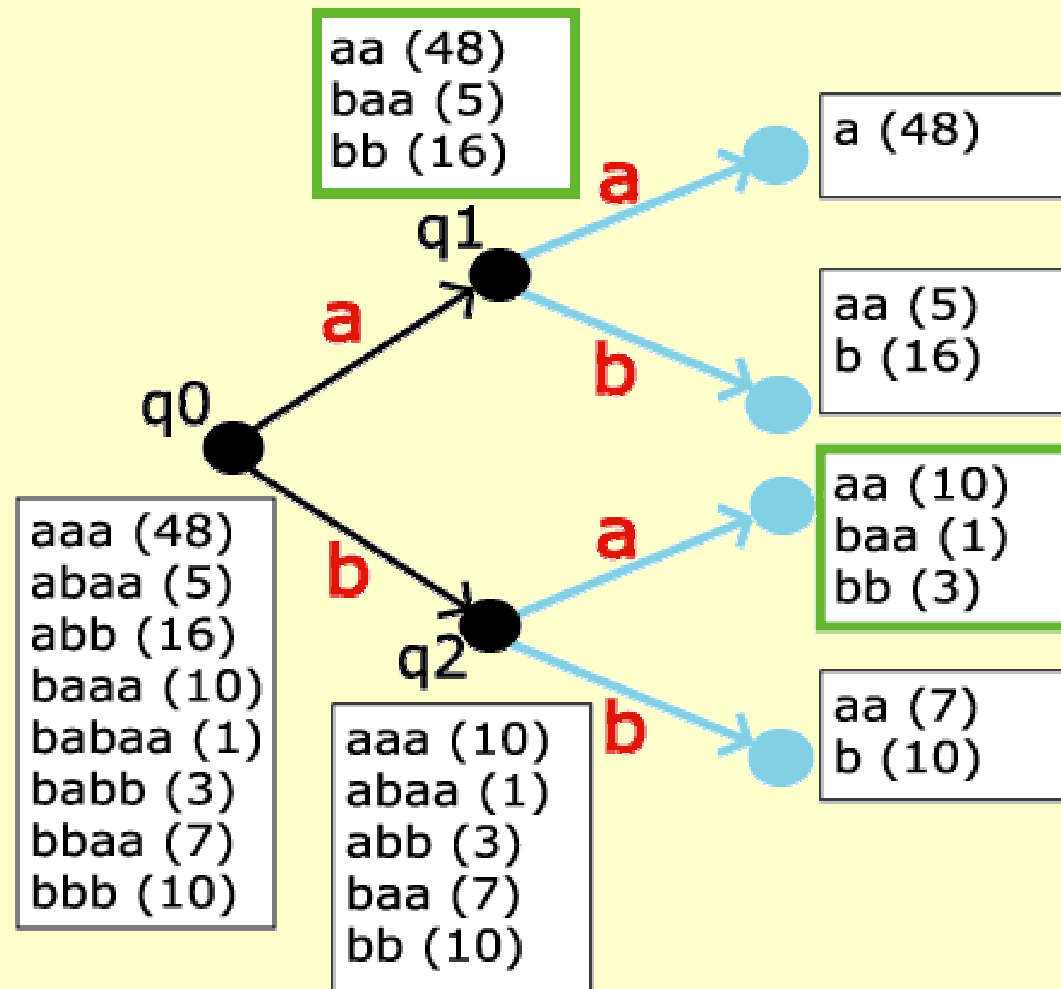
Algorithm to learn DFA



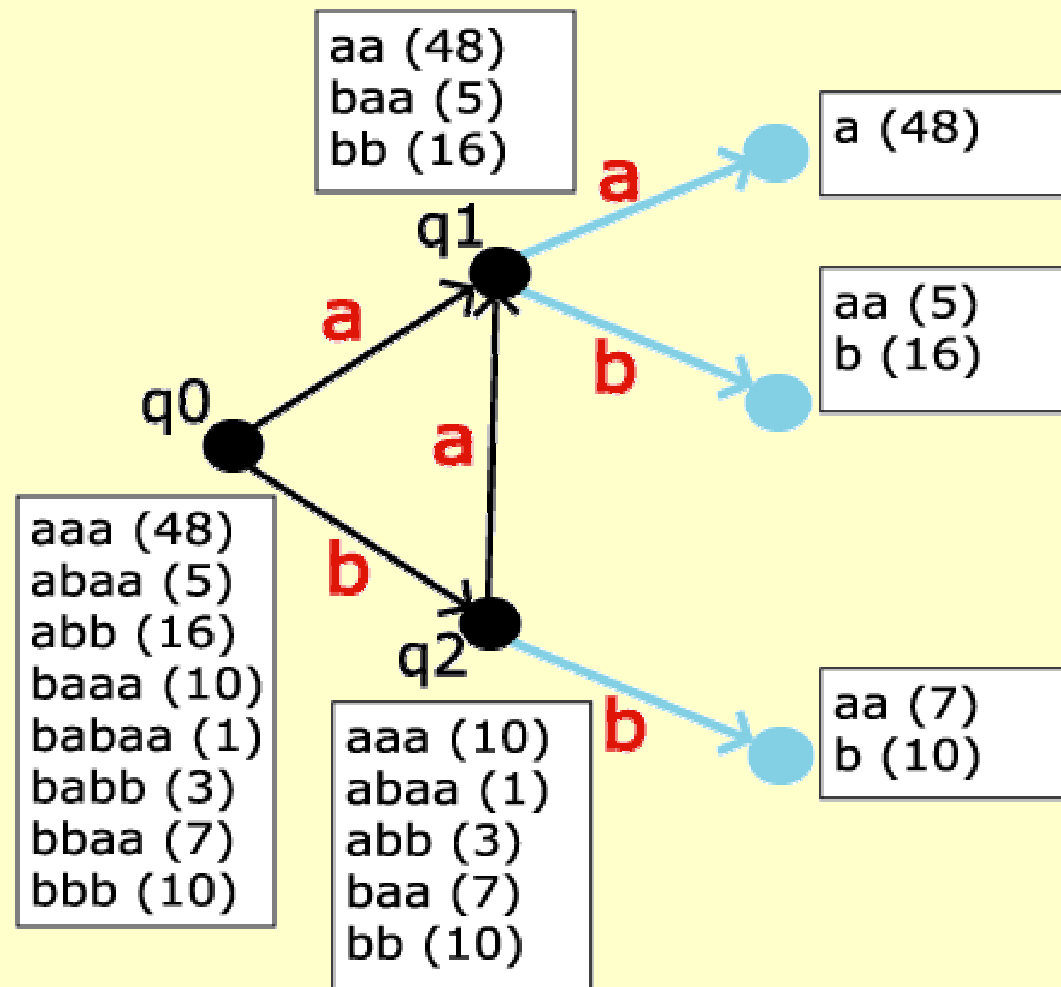
Algorithm to learn DFA



Algorithm to learn DFA



Algorithm to learn DFA



Algorithm to learn DFA

Sample size

$$N = \max \left(\frac{8n^2 |\Sigma|^2}{\epsilon^2} \log \left(\frac{4n^2 |\Sigma|^2}{\delta} \right), \frac{8n^2 |\Sigma|^2}{\epsilon \delta \mu^2} \right)$$

With probability of at least $1 - \frac{\delta}{2}$, the DFA is learnt such that it accepts at least a fraction $1 - \frac{\epsilon}{2}$ of all strings accepted by A.

Assigning Probabilities

- Sample size

$$N' = \left(\frac{2^{17} n^4 |\Sigma|^4}{\delta \epsilon^3} \right) \ln \left(\frac{4n|\Sigma|}{\delta} \right)$$

- Let $S_H \subseteq S_A$ be the set of strings accepted by H ($D_H(s) > 0$). With a probability of at least $1 - \frac{\delta}{2}$:

$$\sum_{s \in S_H} |D_A(s) - D_H(s)| \leq \frac{\epsilon}{2}$$

Assigning Probabilities

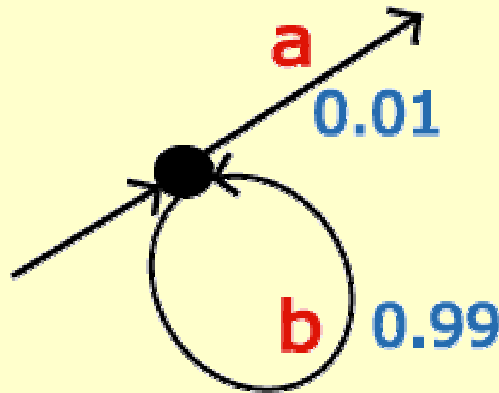
$$\text{var}(D_A, D_H)$$

$$= \left(\sum_{s \in S_H} |D_A(s) - D_H(s)| \right) + \left(\sum_{s \notin S_H} D_A(s) \right)$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

Cumulative Error

Problem with having no bounds on length.



The state is observed a large number of times.

Result

Palmer and Goldberg (2005)

With probability of at least $1-\delta$,

$$\text{var}(D_A, D_H) \leq \varepsilon$$

Sample size polynomial in $1/\varepsilon$, $1/\delta$, $|\Sigma|$, n
and μ .

Complexity

Constructing DFA:

$$N = \max \left(\frac{8n^2|\Sigma|^2}{\epsilon^2} \log \left(\frac{4n^2|\Sigma|^2}{\delta} \right), \frac{8n^2|\Sigma|^2}{\epsilon\delta\mu^2} \right)$$

(at most $n|\Sigma|$ executions)

Complexity

Finding Transition Probabilities:

$$N' = \left(\frac{2^{17} n^4 |\Sigma|^4}{\delta \epsilon^3} \right) \ln \left(\frac{4n|\Sigma|}{\delta} \right)$$

(at most $n|\Sigma|$ executions)

Distinguishability (μ)

States q_1 and q_2 are distinguishable if

$$\max_{s \in \Sigma^*} \left(\left| \frac{|s \in S_{q_1}|}{|S_{q_1}|} - \frac{|s \in S_{q_2}|}{|S_{q_2}|} \right| \right) > \frac{\mu}{2}$$

where S_q is the multiset of suffixes at q .

Properties of KL-Divergence

KL-divergence is *relative entropy*.

It represents the amount of information lost if D_H is used as opposed to D_A .

- It is always non-negative.
- $KL(D_A, D_H) = 0$ only if D_A is identical to D_H .