Learning Languages with Rational Kernels

Mehryar Mohri
Courant Institute and Google Research
mohri@cs.nyu.edu

Joint work with
Corinna Cortes and Aryeh (Leo) Kontorovich

State Merging/Splitting Paradigm

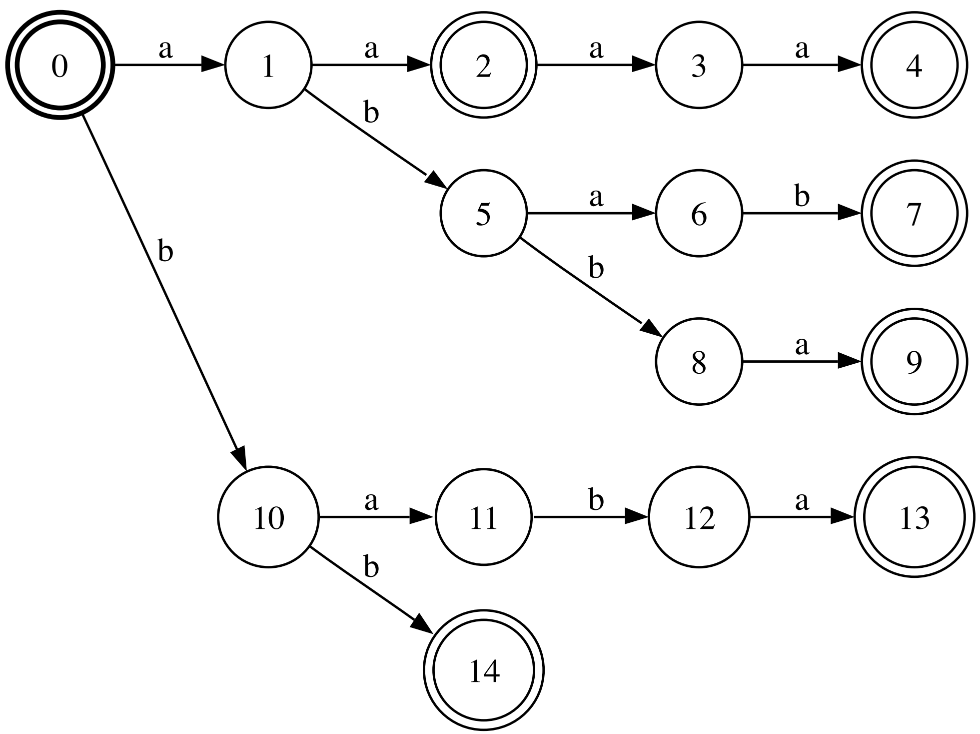
[Angluin 1982; Oncina et al. 1993; Ron et al. 1997; ...]

- Start with automaton or tree accepting all examples (finest partition).
- Iteratively merge states (partition blocks) while preserving some congruence.
- Return resulting automaton when no more merging is possible while preserving congruence.

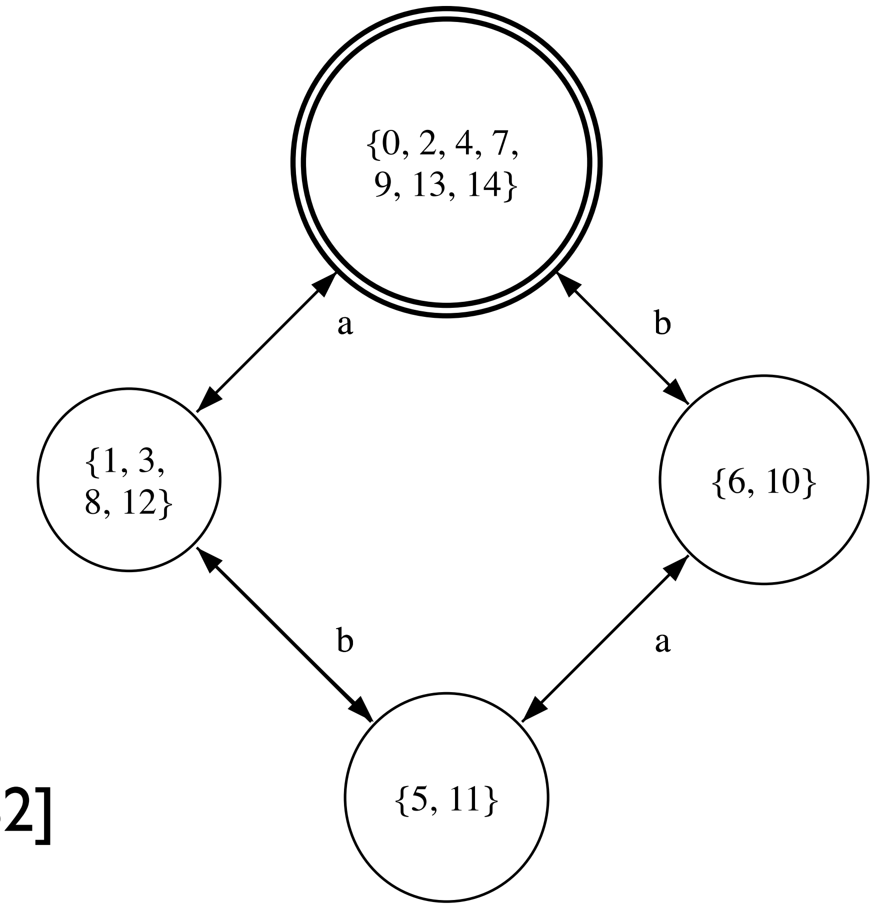
→ choice of congruence fully determines the algorithm.

Example

- Example: $L = \{\epsilon, aa, bb, aaaa, abab, abba, baba\}$.



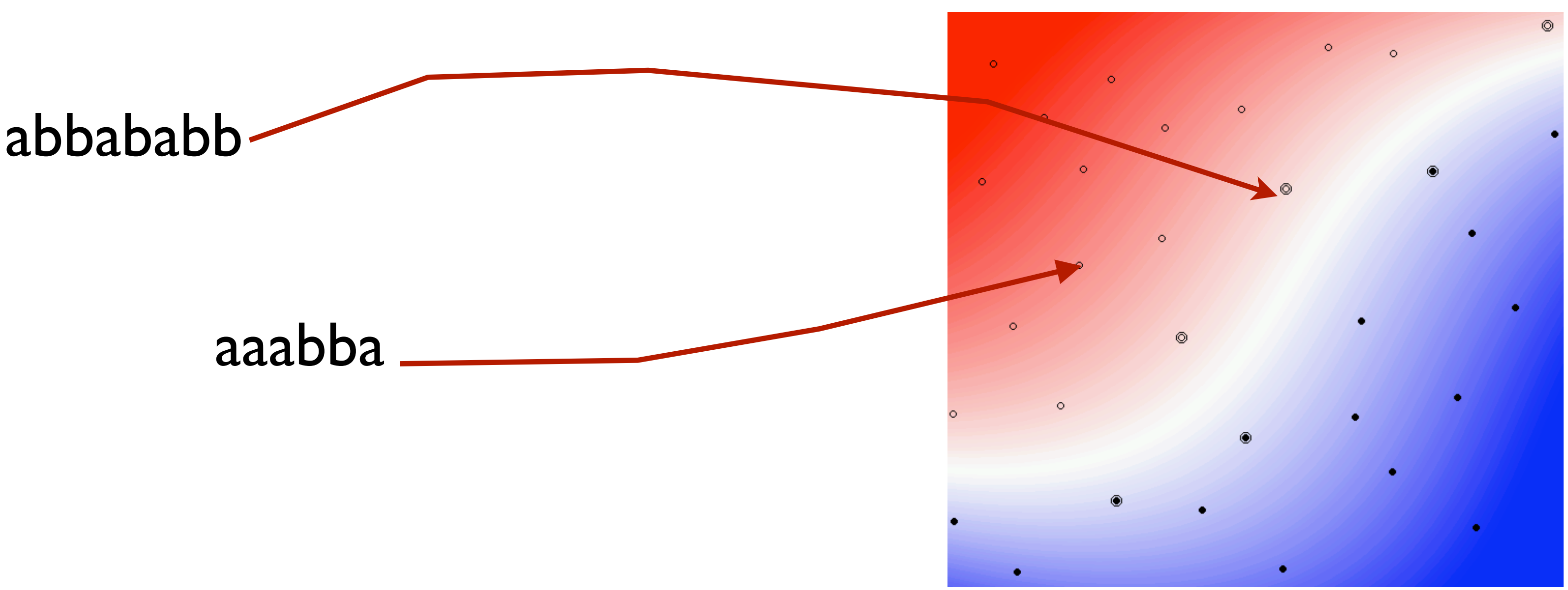
Zero reversible languages [Angluin, 1982]



Even number of as and bs.

New Language Learning Paradigm

- Map strings to a high-dimensional feature space $\Phi\colon \Sigma^* \rightarrow F$.
- Learn separating hyperplane in that space.
- Mappings can be implicitly defined via PDS kernels.



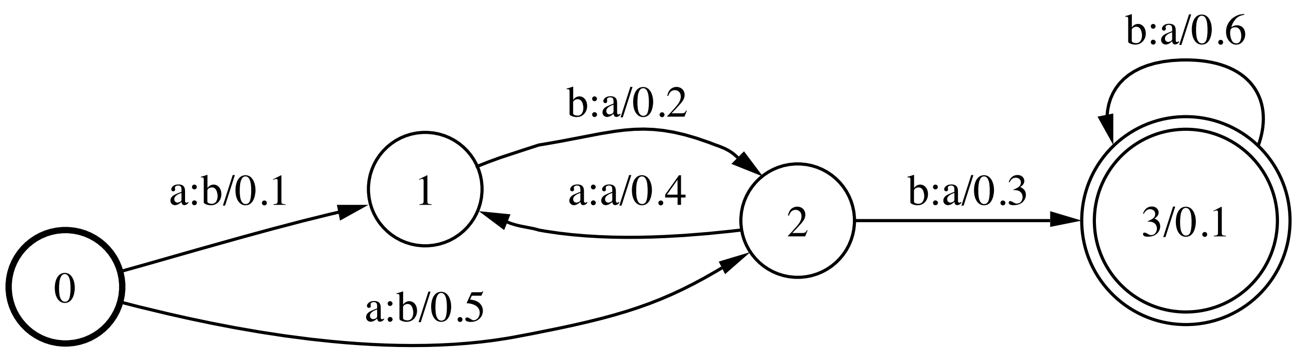
Which Sequence Kernels?

- The string kernels used in NLP and bioinformatics are all special instances of rational kernels:
  - n-gram kernel, gappy n-gram kernels (Lodhi et al., 2001).
  - tree kernels (Collins and Duffy, 2002).
  - moment kernels (Cortes and Mohri, 2005).
  - locality-improved kernels (Zien et al., 2000).
  - mismatch kernels (Leslie et al., 2003).

This Talk

- Weighted transducers
- Rational kernels
- Linear separability with rational kernels

Weighted Finite-State Transducers



$T(x, y) =$ Sum of the weights of all successful paths with input *x* and output *y*

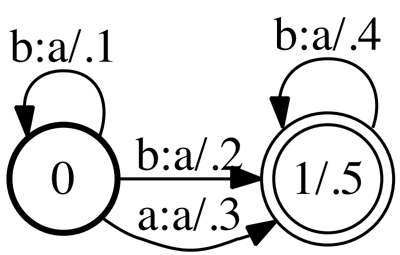
$$T(abb, baa) = .1 \times .2 \times .3 \times .1 + .5 \times .3 \times .6 \times .1.$$

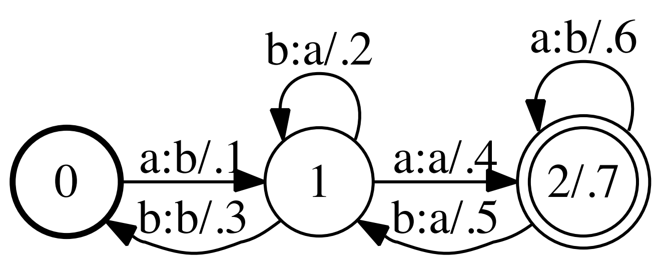
Sum

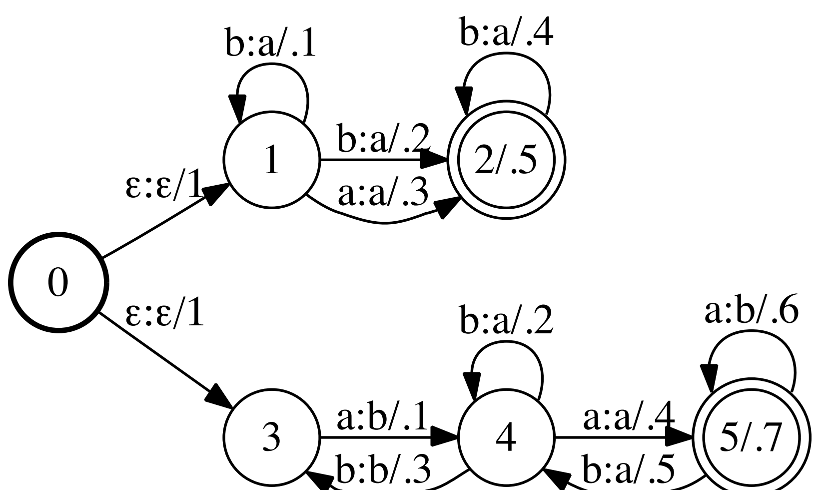
- Definition:

$$(T_1 + T_2)(x, y) = T_1(x, y) + T_2(x, y).$$

- Illustration:





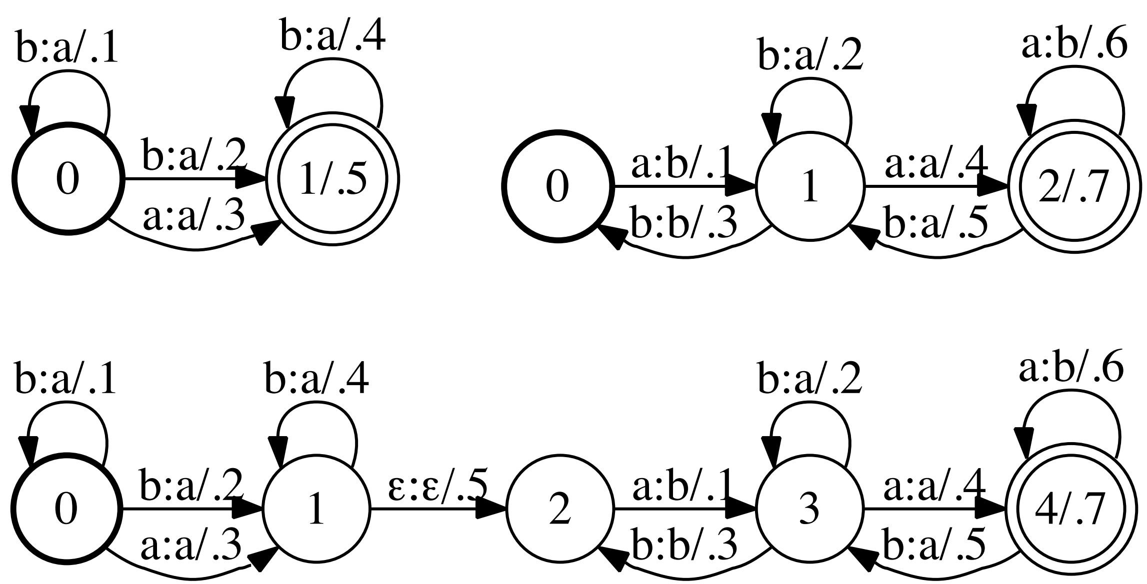


Product

- Definition:

$$(T_1 \cdot T_2)(x, y) = \sum_{x_1 x_2 = x,\ y_1 y_2 = y} T_1(x_1, y_1) \cdot T_2(x_2, y_2).$$

- Illustration:

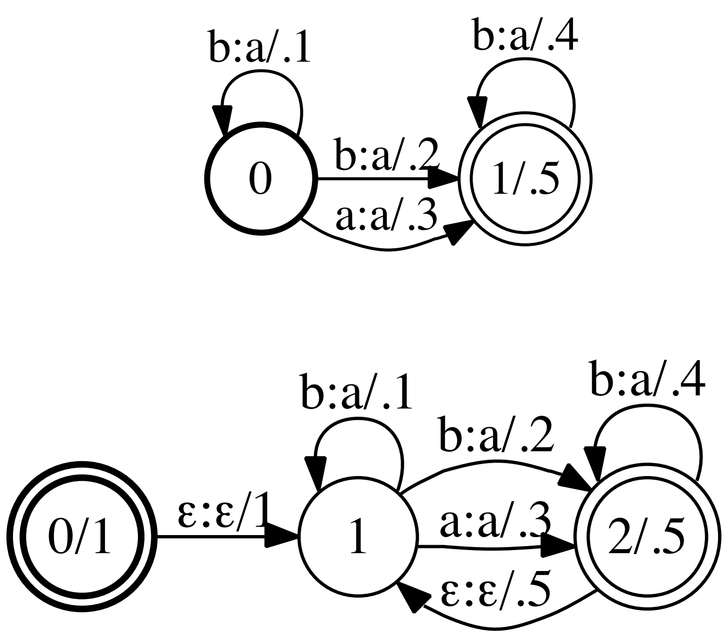


Closure

- Definition:

$$T^*(x, y) = \sum_{n=0}^{+\infty} T^n(x, y).$$

- Illustration:

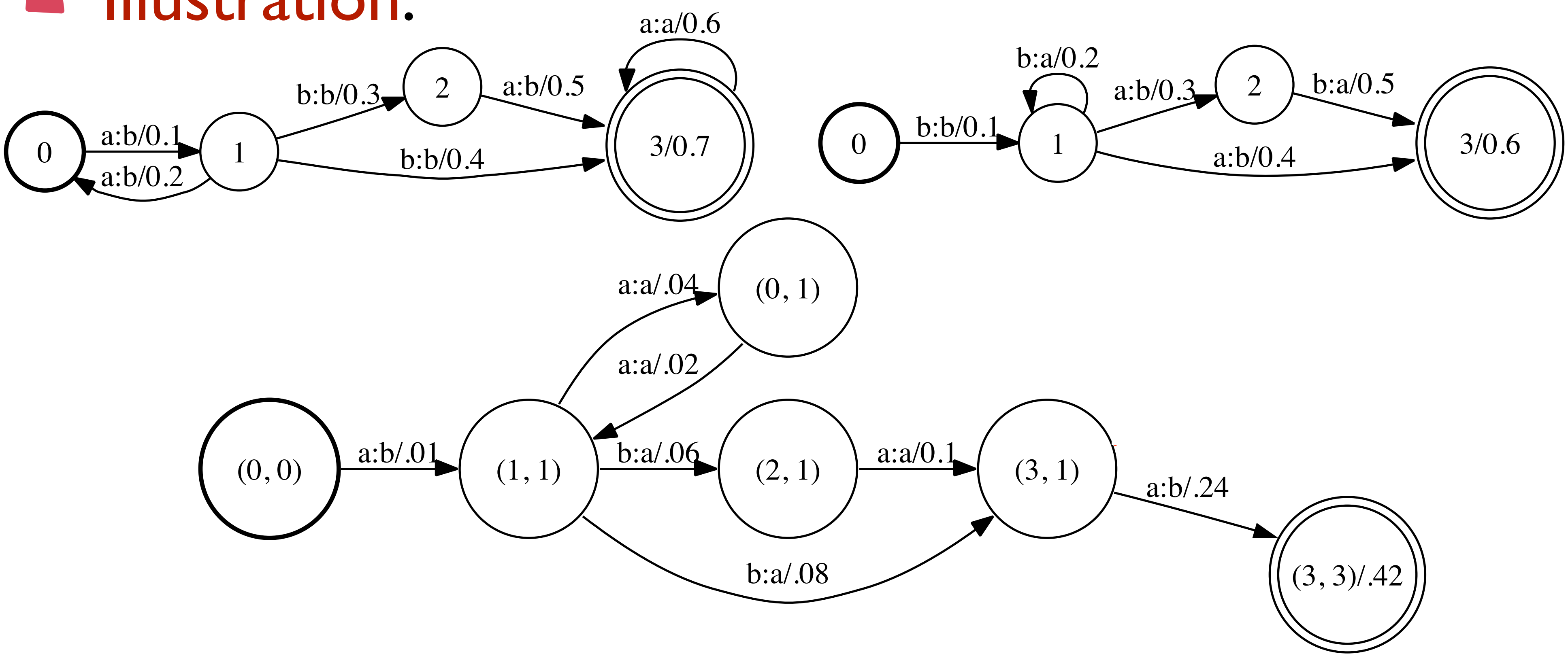


Composition

- **Definition:**

$$(T_1 \circ T_2)(x, y) = \sum_{z \in \Sigma^*} T_1(x, z)\, T_2(z, y).$$

- **Illustration:**



Sum of Path Weights

- The sum $S(T)$ of the weights of all accepted paths of a transducer T is

$$S(T) = \sum_{\pi \in P(I,F)} \underbrace{w[\pi]}_{\text{path weight}} \; \underbrace{\rho(n[\pi])}_{\text{final weight}} .$$

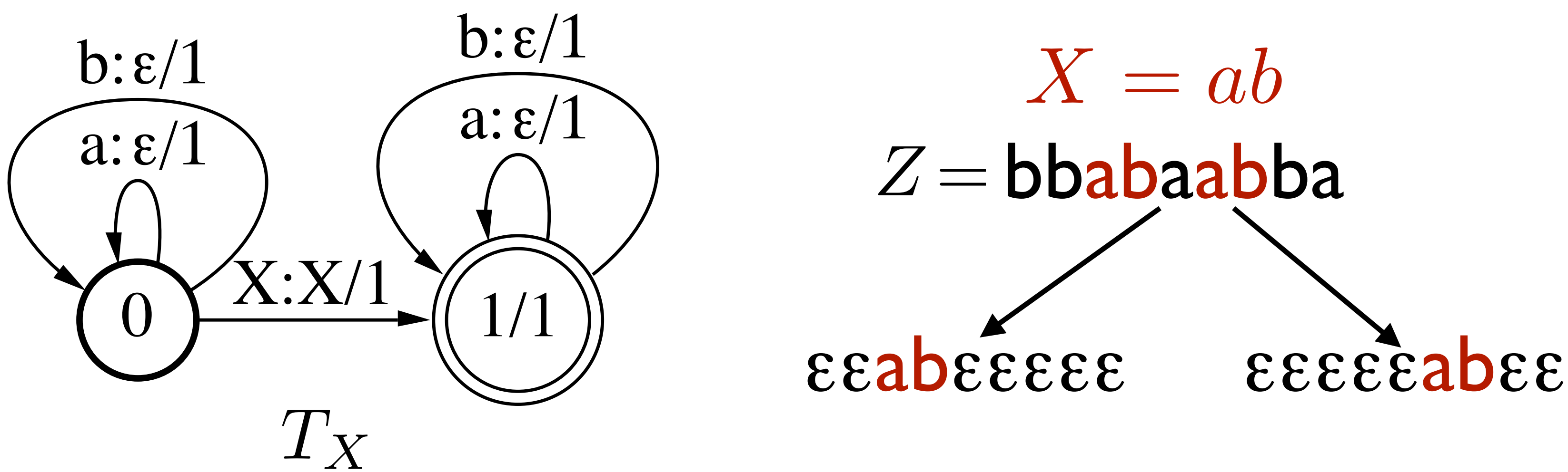
- Properties:
  - linearity: $S(T_1 + T_2) = S(T_1) + S(T_2).$

$$S(\lambda T) = \lambda S(T).$$

$$S((\lambda T_1) \circ T_2) = \lambda S(T_1 \circ T_2).$$

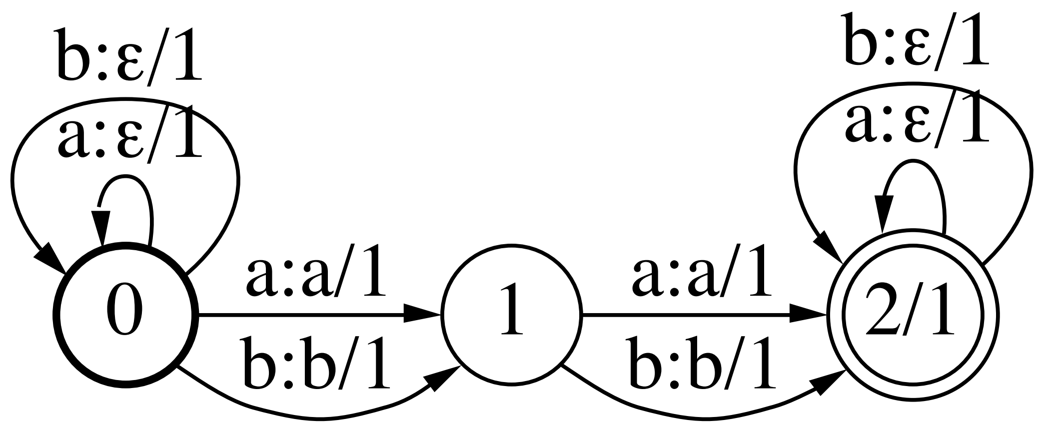
  - computation: general shortest-distance algorithm (extension to (+, x) of standard (min, +)) (MM, 1998).

Counting Transducers



- X may be a string or an automaton representing a regular expression.
- Counts: $\text{count}(Z, X) = S(Z \circ T_X)$.

Transducer Counting Bigrams



T_{bigram}

$$\text{count}(Z, ab) = S(Z \circ T_{\text{bigram}} \circ ab).$$

This Talk

- Weighted transducers
- Rational kernels
- Linear separability with rational kernels

Rational Kernels over Strings

(Cortes, Haffner, and MM 2004)

- Definition: a kernel K is rational if there exists a weighted transducer U such that for all strings x and y:

$$K(x, y) = U(x, y).$$

- Computation: composition and shortest-distance algorithm using $K(x, y) = S(Aut(x) \circ U \circ Aut(y))$.
  - complexity: $O(|x||y|)$ in general.
  - better complexity in specific cases, using more efficient composition.

Rational Kernels over Strings

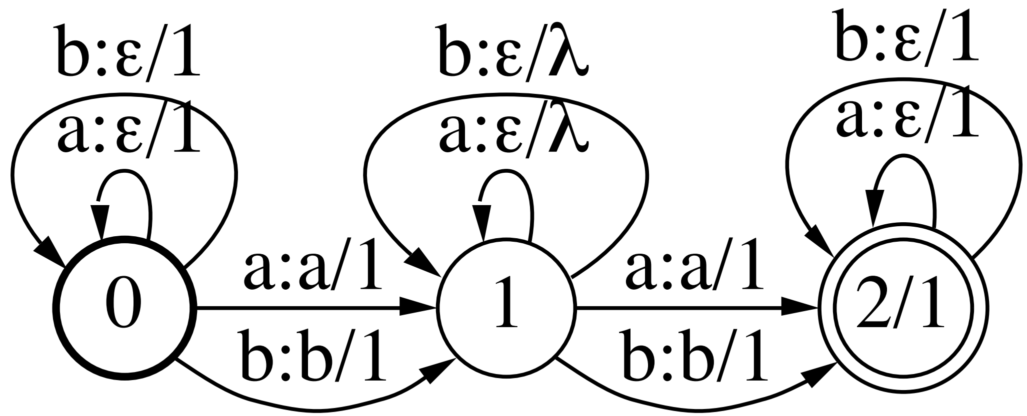
(Cortes, Haffner, and MM 2004)

- **Definition**: a kernel *K* is rational if there exists a weighted transducer *U* such that for all strings *x* and *y*:

$$K(x, y) = U(x, y).$$

- **Theorem**: let T^{-1} denote T with input and output labels swapped. Then, $U = T \circ T^{-1}$ defines a positive definite symmetric rational kernel.

Gappy Bigram Kernel



$T_{\text{gappy-bigram}}$

$Z \circ T_{\text{gappy-bigram}}$ computes the expected count of all gappy bigram with gap penalty factor λ.

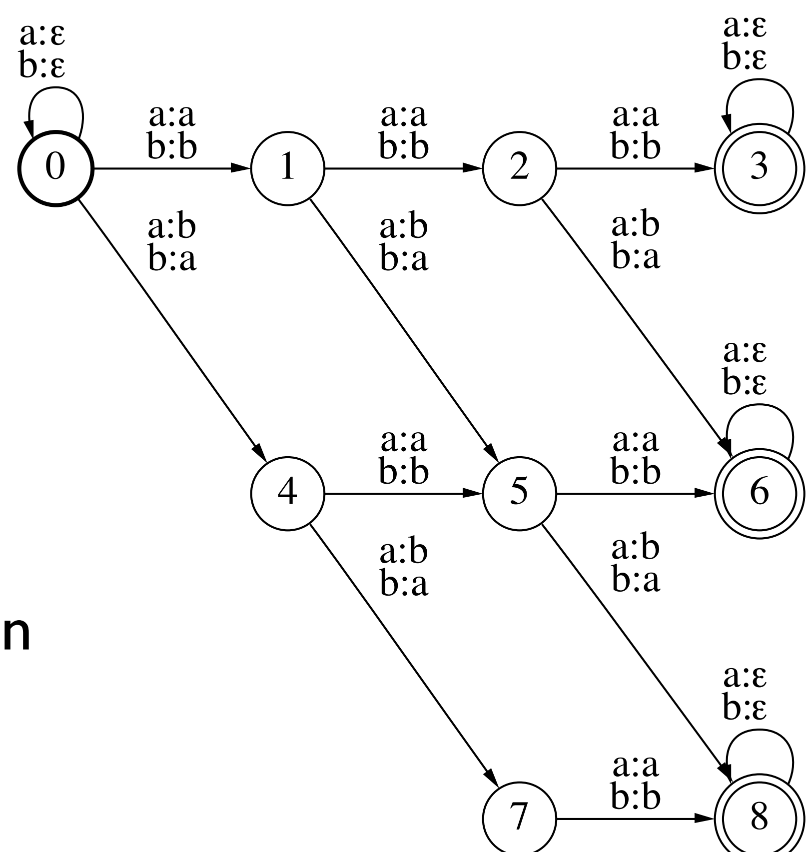
Mismatch Kernel

- **Definition**: for sequences *x* and *y*,

$$K_{(k,m)}(x,y) = \sum_{z_1 \in F_k(x),\, z_2 \in F_k(y),\, z \in \Sigma^k} d_m(z_1, z)\; d_m(z, z_2)$$

- **Representation**:

$K_{(3,2)}$



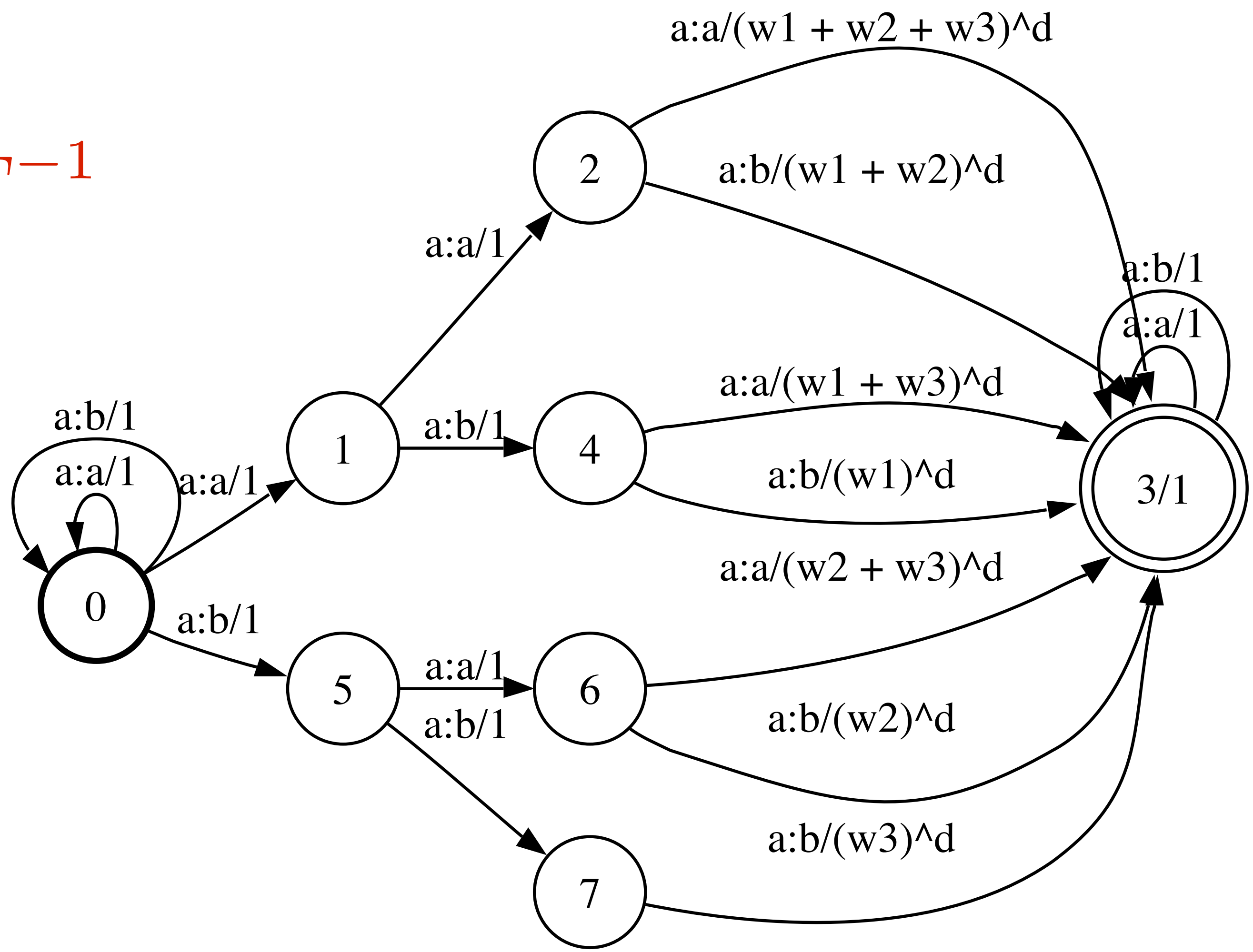
Remote homology detection
(Leslie et al., 2002)

Locality-Improved Kernel

- Representation $(l = 1)$:

Recognition of translation initiation sites (Zien et al., 2000)

$T \circ T^{-1}$



Questions

- Linear separation with RKs
  - what is the set of languages separable with RKs?
  - what languages are separable with a given RK?
  - when does linear sep. guarantee positive margin?
  - how do we create transducers with finite range?

This Talk

- Weighted transducers
- Rational kernels
- Linear separability with rational kernels

Probabilistic Automata

(Paz, 1971; Rabin, 1963)

- Definition: a weighted automaton over $\mathbb{R}$ is probabilistic if
  - it has no negative weight.
  - the weights of outgoing transitions with the same label sum to one at every state.

- Definition: a language L is $\mathbb{R}$-stochastic iff there exists a probabilistic automaton A and $\lambda > 0$ such that

$$L = \{x \colon A(x) > \lambda\}.$$

Turakainen's Theorem

- **Theorem** (Turakainen, 1969): Let S be a weighted automaton over $\mathbb{R}$ with n states. A probabilistic automaton B over $\mathbb{R}$ with n + 3 states can be constructed from S such that:

$$\forall x \in \Sigma^+, \; S(x) = c^{|x|} \left(B(x) - \frac{1}{n+3} \right),$$

where c is a large number.

→ $L = \{x \colon S(x) > 0\}$ for some weighted automaton S is necessarily stochastic.

Languages Linearly Separable by RKs

- Theorem (characterization): L is linearly separable by a RK $K = T \circ T^{-1}$ iff it is stochastic.
- Proof: assume that L is stochastic.
  - We can assume that there exists a S such that $L = \{x : S(x) > 0\}$.
  - Let $x_0 \in L$. Then, $L = \{x : K(x, x_0) > 0\}$, where $K = T \circ T^{-1}$, and T is the transducer derived from S by adding output ϵs, since

$$(T \circ T^{-1})(x, x_0) = T(x, \epsilon)T^{-1}(\epsilon, x_0) = S(x)S(x_0).$$

Languages Linearly Separable by RKs

- Proof: conversely, assume that L is linearly separable by $K = T \circ T^{-1}$.

$$\sum_{i=1}^{m} \alpha_i K(x_i, x) = \sum_{i=1}^{m} \alpha_i S\big(Aut(x_i) \circ T \circ T^{-1} \circ Aut(x)\big)$$

$$= S\big(\sum_{i=1}^{m} \alpha_i Aut(x_i) \circ T \circ T^{-1} \circ Aut(x)\big)$$

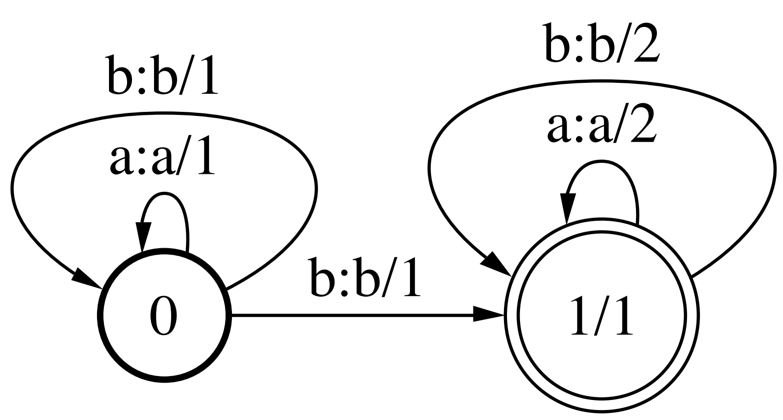
$$= S\big(A \circ T \circ T^{-1} \circ Aut(x)\big)$$

$$= S\big(R \circ Aut(x)\big) = R(x).$$

- Let $U = R + b$, then $L = \{x \colon S(x) > 0\}$, and, by Turakainen's theorem, L is stochastic.

Examples

- $L = \{a^m b^n c^p : m > n > p\}$ is stochastic.
- $L = \{a^m b^{mn}\}$ is not stochastic.
- The language of palindromes is stochastic.



Computes the integer value of binary numbers (a = 0, b = 1), (Cortes and Mohri, 2000).

Linearly separable *L* for a Fixed RK

- Definition: a rational kernel $K = T \circ T^{-1}$ has finite range if $\{T(x, y) \colon x, y \in \Sigma^*\}$ is finite.
- Theorem: let $K = T \circ T^{-1}$ be a RK with finite range. If L is linearly separable by K, then L is a finite Boolean combination of preimage languages

$$\{x : T(x, y) = v\}.$$

Examples

- For subsequence kernels, each preimage language is a set of sequences admitting a string *y* as a subsequence (the shuffle ideal of *y*):

$$\Sigma^* y_1 \Sigma^* \cdots \Sigma^* y_n \Sigma^*.$$

- For factor (or mismatch) kernels, each preimage language is a set of sequences admitting *y* as a substring:

$$\Sigma^* y \Sigma^*.$$

Margin Property

- In general, linear separation does not guarantee a positive margin.
- Theorem: let *K* be a finite range RK and let *L* be a language linearly separable with *K*. Then, the separation margin is positive.
- Proof (sketch): hyperplane $\langle w, \Phi(x) \rangle + b = 0$.
  - w has finite support, thus can use Φ' instead.
  - since *K* has finite range,

$$\rho = \inf_{x \in X} \frac{|\langle w, \Phi(x) \rangle + b|}{\|w\|} = \min_{x \in X} \frac{|\langle w, \Phi'(x) \rangle + b|}{\|w\|} > 0.$$

Margin Bound

- Theorem: let C be a finitely linearly separable concept class for the rational kernel $K = T \circ T^{-1}$, then for any concept class $c \in C$ there exists $\rho_0 > 0$ such that with probability at least $1 - \delta$, there exists a linear separator h with generalization error at most

$$O\left(\frac{(\log^2 m)R^2/\rho_0^2 + \log(1/\delta)}{m}\right).$$

Piecewise Testable Languages

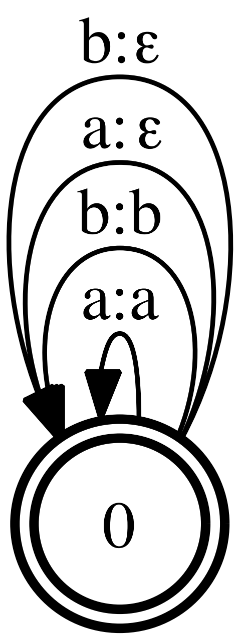
(Kontorovich, Cortes, MM, 2006)

- Piecewise testable languages are linearly separable using subsequence kernels.
- Subsequence kernels are rational kernels (Kontorovich, Cortes, MM, 2007) and are efficient to compute.
- Linear separation with a positive margin.

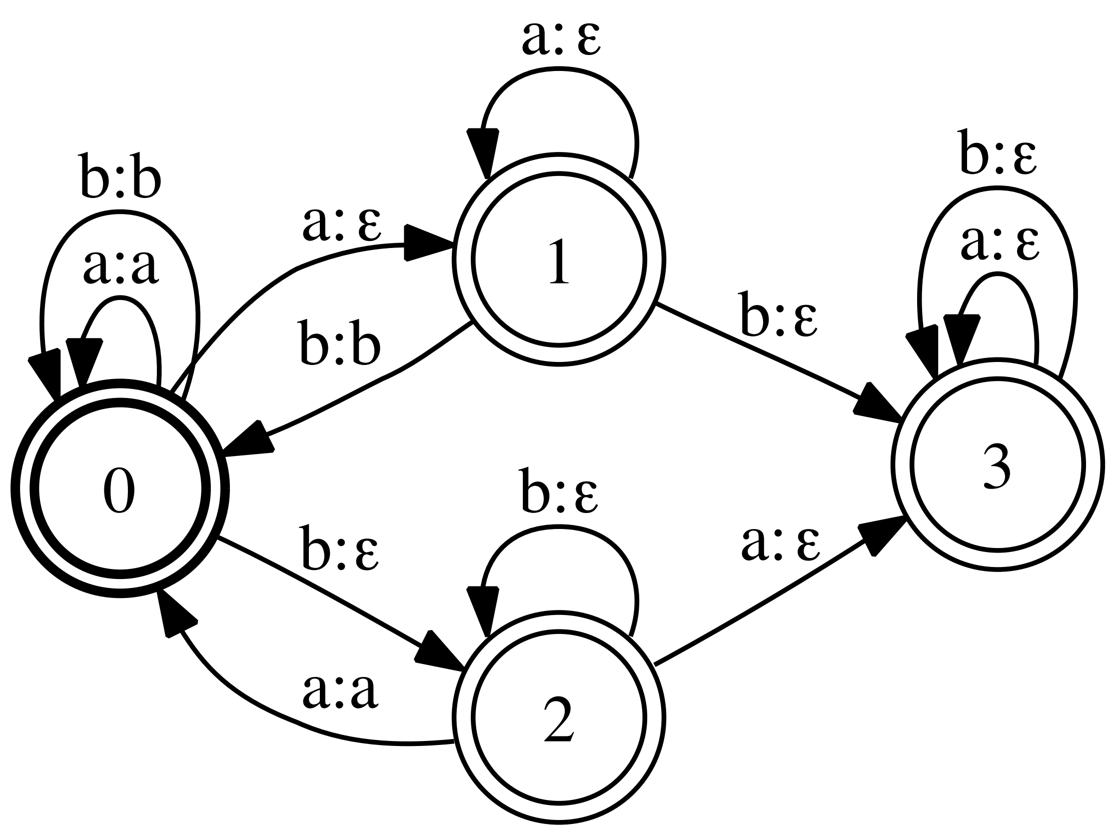
Double-Tape Disambiguation

- Objective: given transducer T, create unambiguous transducer T', that is for any (x, y) labeling a path in T, unique path labeled with (x, y) in T'.

Double-Tape Disambig. - Example

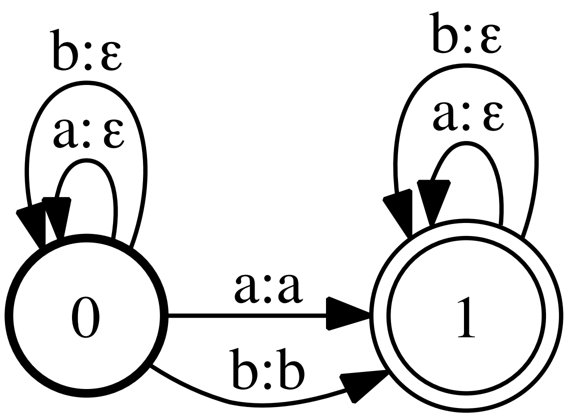


Subsequence transducer

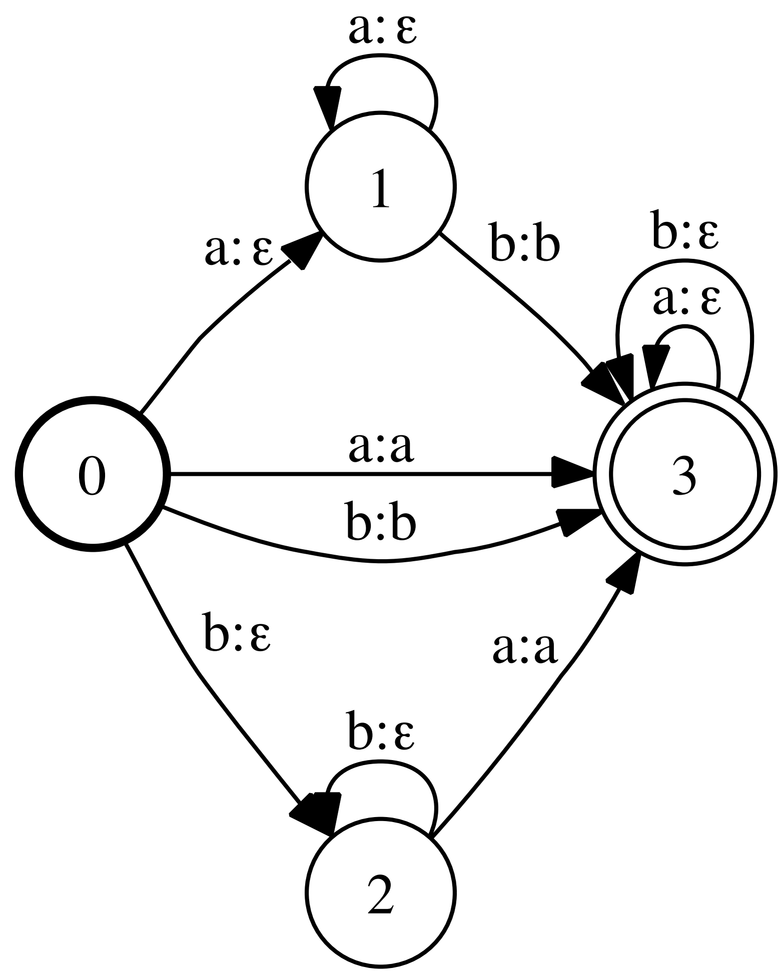


Double-tape disambiguated subsequence transducer

Double-Tape Disambig. - Example



Unigram transducer



Double-tape disambiguated unigram transducer

Conclusion

- Characterization of languages linearly separable by RKs: stochastic languages.
- RKs with finite range have remarkable properties:
  - separable languages are finite Boolean combinations of preimages by T.
  - guarantee positive margin.
  - double-tape disambiguation possible in some cases to design finite range RKs.