Rational Kernels 

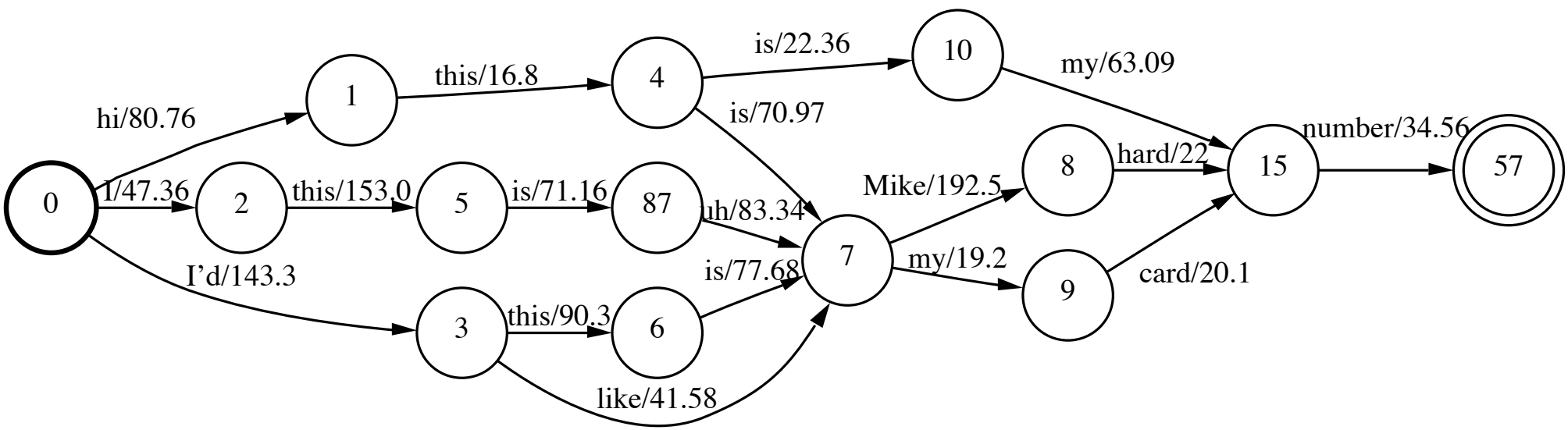
Mehryar Mohri
Courant Institute
mohri@cs.nyu.edu

Joint work with Corinna Cortes (Google Research)
and Patrick Haffner (AT&T Labs)

With contributions from Fernando Pereira and Michael Riley

Initial Motivation

- Problem: assign a category (e.g., *referral*, *pre-certification*) to each speech utterance
- Example:
  - Spoken utterance: *"Hi this is my number"*
  - Speech recognizer's output ('word lattice'):



Computational Biology: Similar Situation

- Problem: decide which class (e.g., *protein families*, *CpG islands*) a biosequence, or a group of biosequences, belongs to
- Objects to classify:
  - Single protein sequence
  - Protein clusters (represented or modeled by weighted automata)

General Problem

- Spoken-dialog classification
- Computational biology
- Information extraction
- Text mining
- Document classification
- Database queries

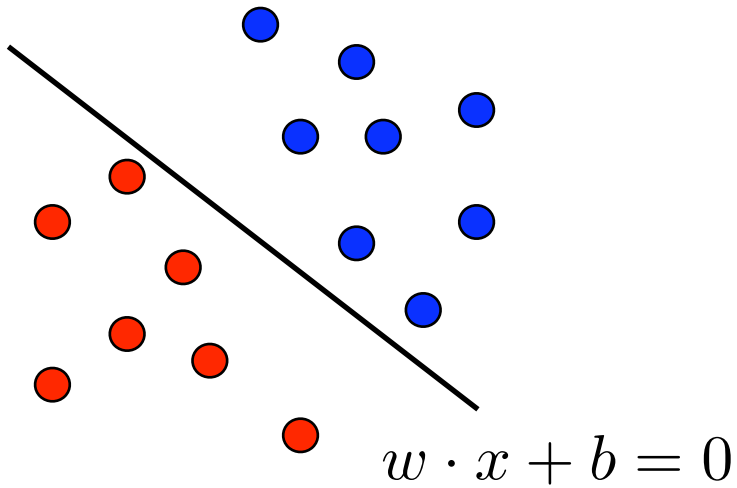
Question

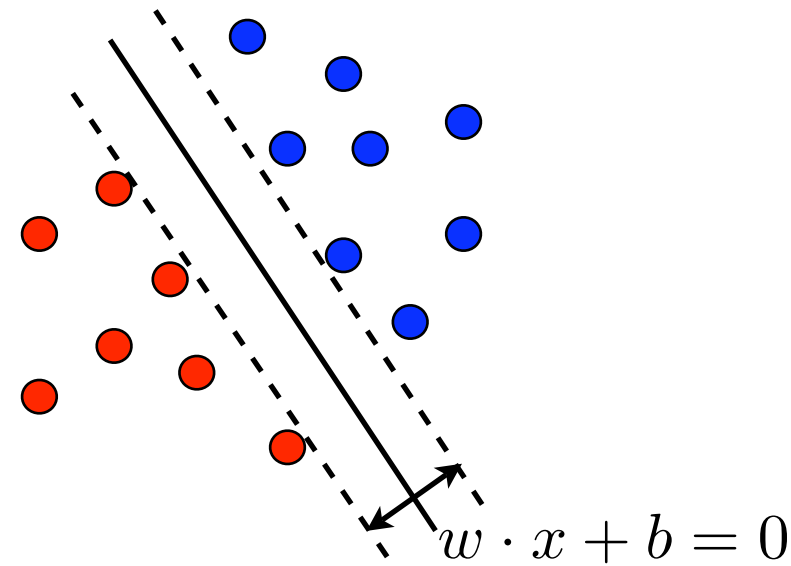
- The objects to analyze in many modern applications are:
  - Variable-length sequences
  - Distributions over variable-length sequences, typically weighted automata
- But, most classification algorithms are designed for fixed-size vectors
- How do we handle real-world inputs?

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

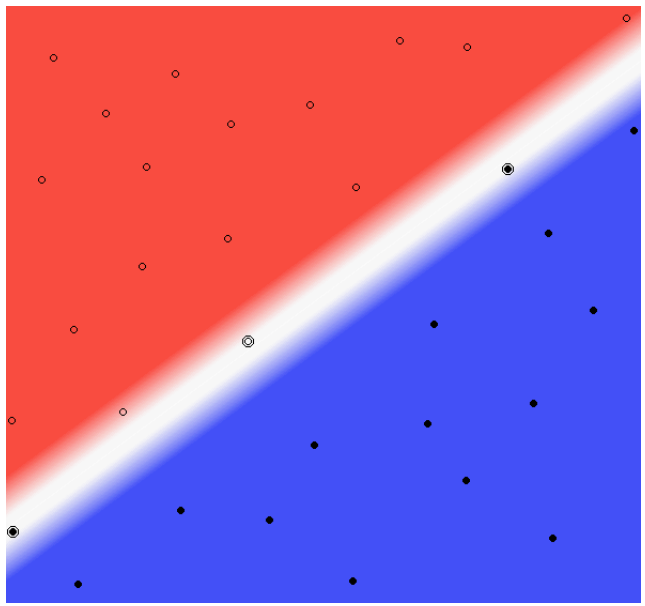
Large-Margin Linear Classifiers

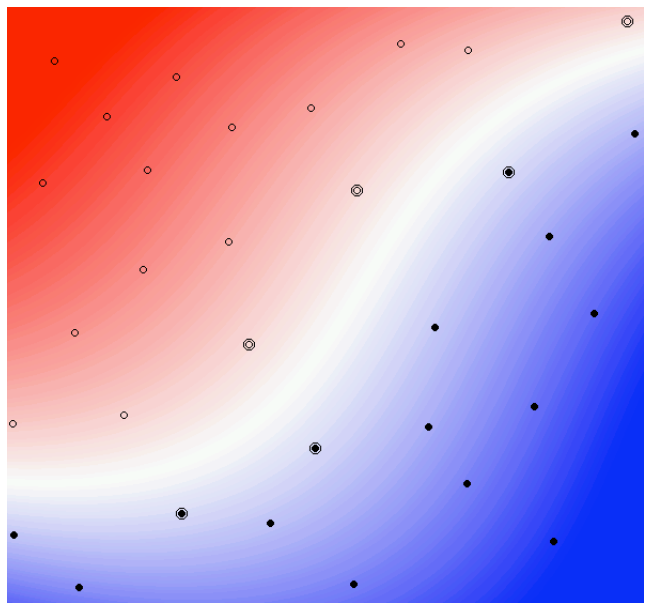




$$f(x) = w \cdot x + b = \sum_{i=1}^{N} \alpha_i (x_i \cdot x) + b$$

Non-Linear case





- Non-linear separation necessary in most problems.
- Non-linear mapping from input space to high-dimensionnel feature space: $\Phi : X \to F$
- Generalization ability: independent of the dimension of *F*, depends only on margin and the number of examples

Kernel Methods

- Idea:
  - Define K (called *kernel*) such that:

$$\Phi(x) \cdot \Phi(y) = K(x, y)$$

  - K often interpreted as a 'similarity measure'
- Benefits:
  - Efficiency: K may be much more efficient to compute than Φ and the dot product
  - Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (Mercer's condition).

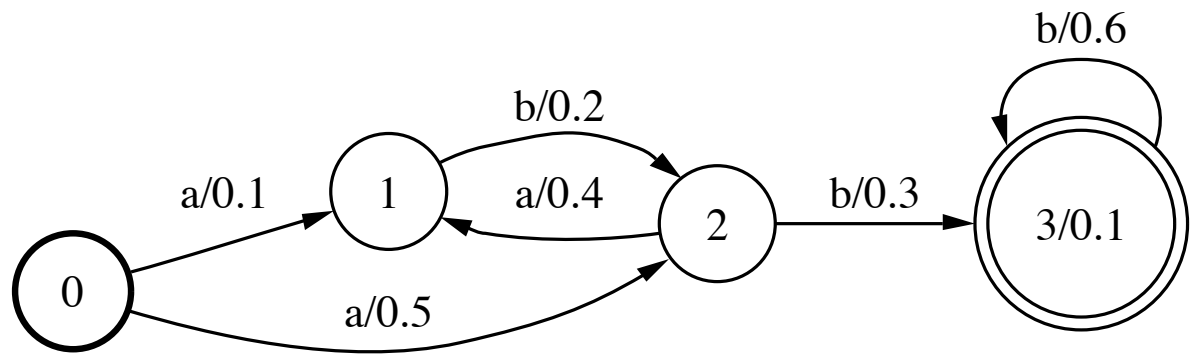
Positive Definite Symmetric Kernels

- Conditions equivalent to Mercer's condition: for all $\{x_1, \ldots, x_n\} \subseteq X$, matrix $K(x_i, x_j)_{i,j \leq n}$ is symmetric and:
  - is semi-definite positive
  - has non-negative eigenvalues
- How do we define a positive definite symmetric kernel for weighted automata?

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

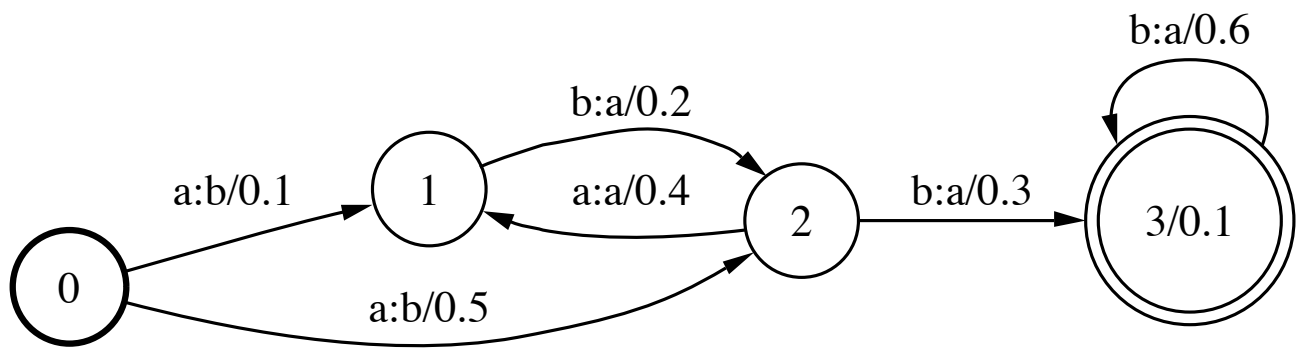
Weighted Automata



$[[A]](x) =$ Sum of the weights of all successful paths labeled with *x*

$$[[A]](abb) = .1 \times .2 \times .3 \times .1 + .5 \times .3 \times .6 \times .1$$

Weighted Transducers



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

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

'Similarity Measures' for Sequences

- Idea: two sequences are related when they share some common factors/subsequences
- Example: sum of the product of the counts of common factors
- Question: how do we do that in a general way and for weighted automata?

Rational Kernels Over Strings

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

$$K(x, y) = [[T]](x, y)$$

Rational Kernels Over Weighted Automata

- Definition: a kernel K is *rational* if there exists a weighted transducer T such that for all weighted automata A and B:

$$K(A, B) = \sum_{x,y} [[A]](x) \cdot [[T]](x, y) \cdot [[B]](y)$$

This definition can be generalized to the case of an arbitrary *semiring* (general operations)

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

Algorithm: Outline

- Observation:

$$\sum_{x,y} [[A]](x) \cdot [[T]](x,y) \cdot [[B]](y) = \sum_{x,y} [[A \circ T \circ B]](x,y)$$

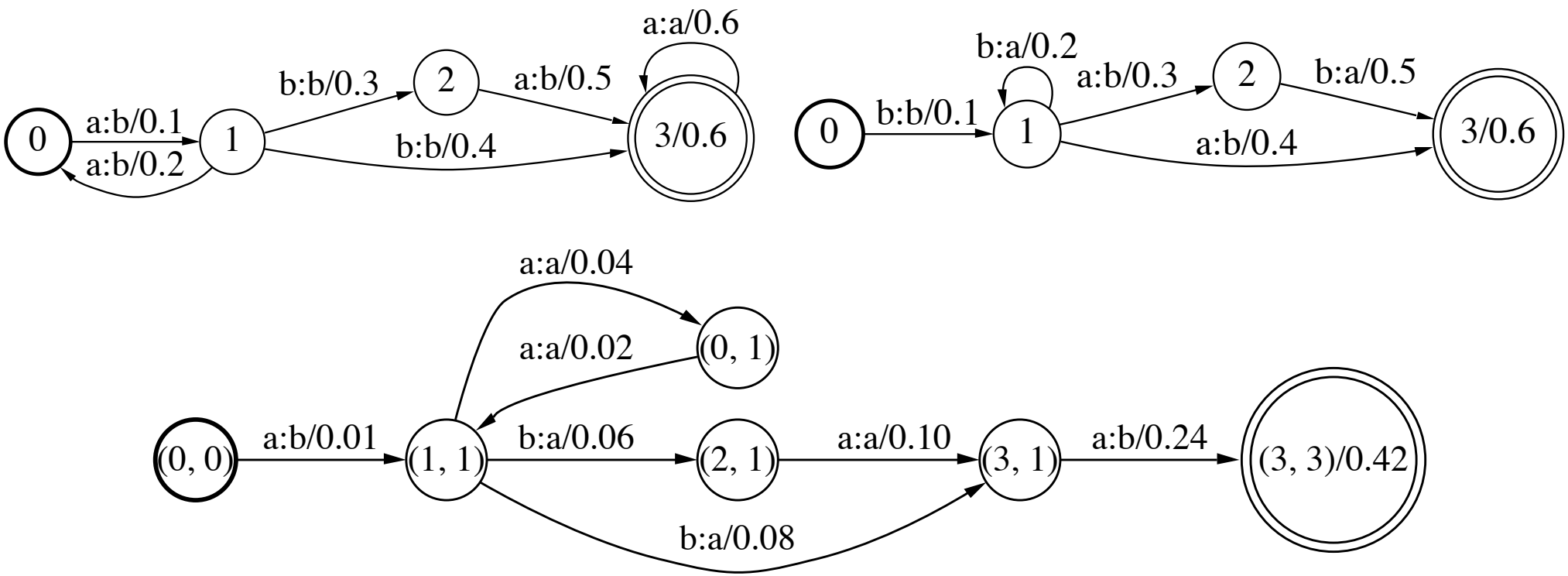
- Compute composed weighted transducer:

$$U = A \circ T \circ B$$

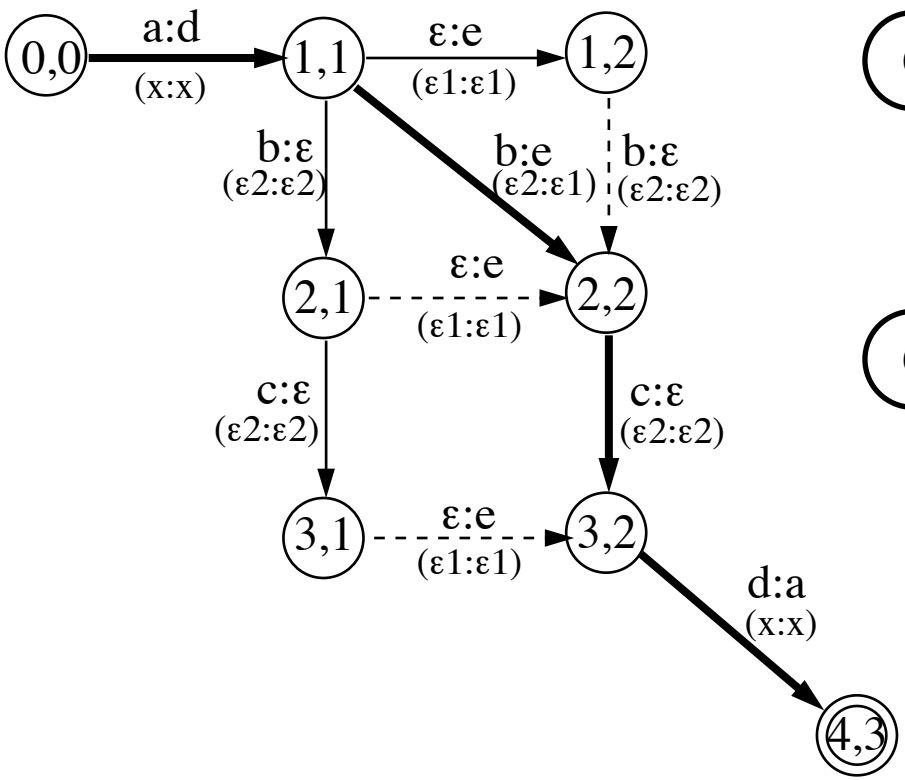
- Use single-source shortest-distance algorithm to compute the sum of the weights of all successful paths of *U*.

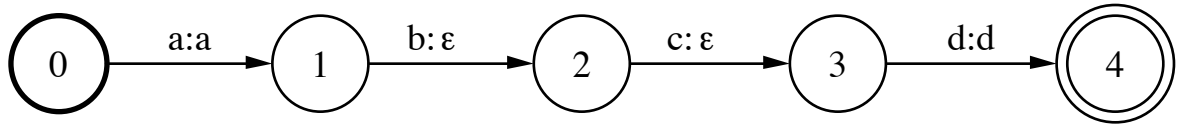
Composition

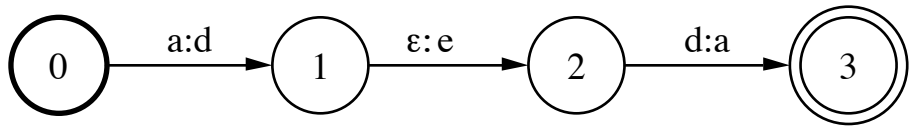
$$\left.\begin{array}{l} q_1 \xrightarrow{a:b/w_1} q_1' \\ q_2 \xrightarrow{b:c/w_2} q_2' \end{array}\right\} \Longrightarrow (q_1, q_2) \xrightarrow{a:c/w_1 w_2} (q_1', q_2')$$



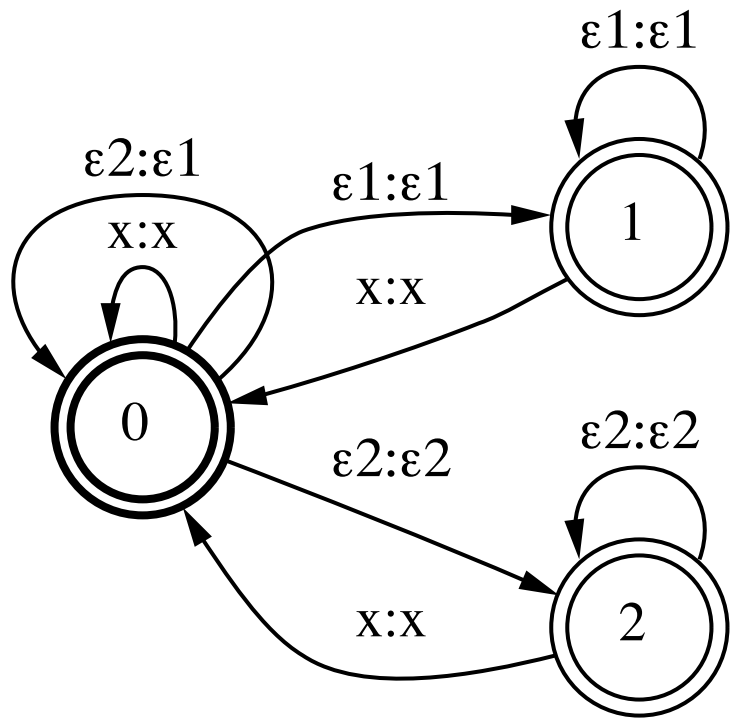
Multiplicity Problem







Composition Filter *F*



$$T_1 \circ T_2 \text{ is replaced by } \tilde{T}_1 \circ F \circ \tilde{T}_2$$

(Mohri et al. 1996)

Algorithm: Complexity

- Automata case: $O(|T||A||B|)$
- String case:
  - General: $O(|T||x||y|)$
  - Specific, using failure functions: $O(|x| + |y|)$

Weight Sets: Semirings

- A *semiring* $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$ is a ring that may lack negation.
  - sum: to compute the weight of a sequence (sum of the weights of the paths labeled with that sequence).
  - product: to compute the weight of a path (product of the weights of constituent transitions).

Semirings - Examples

Semiring	Set	$\oplus$	$\otimes$	$\overline{0}$	$\overline{1}$
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1
Log	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\oplus_{\log}$	$+$	$+\infty$	0
Tropical	$\mathbb{R} \cup \{-\infty, +\infty\}$	$\min$	$+$	$+\infty$	0

with $\oplus_{\log}$ defined by: $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$.

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
  - General computation of rational kernels
  - Counts (expected counts)
- Application to text and speech processing
- Theory
- Application to computational biology

Expected Counts

$|u|_x$: number of occurrences of *x* in *u*

- Expected count of sequence *x* in weighted automaton *A*:

$$c(x) = \sum_{u \in \Sigma^*} |u|_x \, [[A]](u)$$

Count-Based Similarity Measures

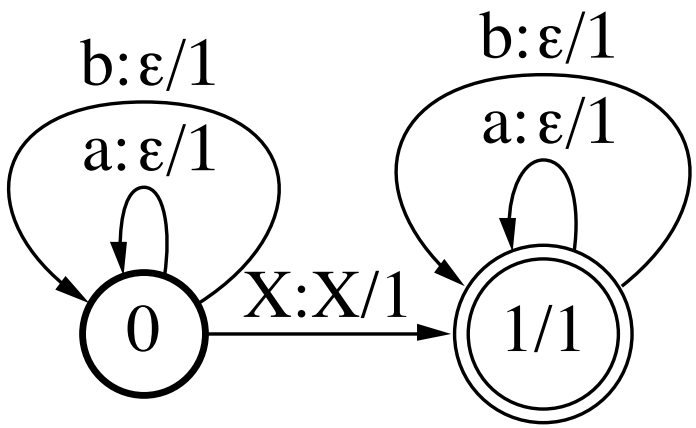
$(A \circ T)$: expected counts of sequences in A

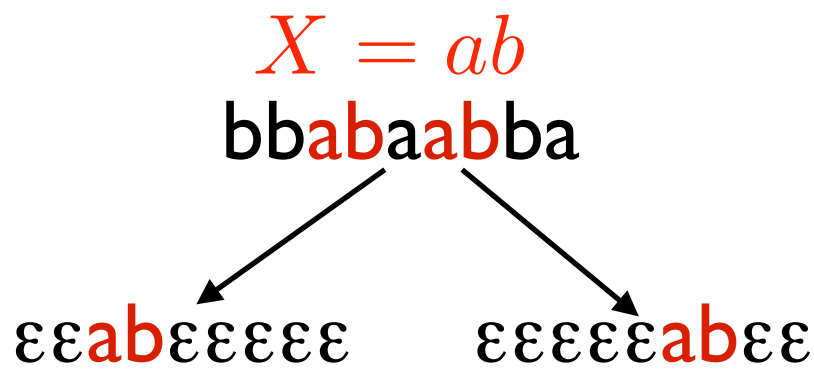
$(T^{-1} \circ B)$: expected counts of sequences in B

$(A \circ T \circ T^{-1} \circ B)$: sum of the expected counts of matching sequences

Rational kernel $(T \circ T^{-1})$

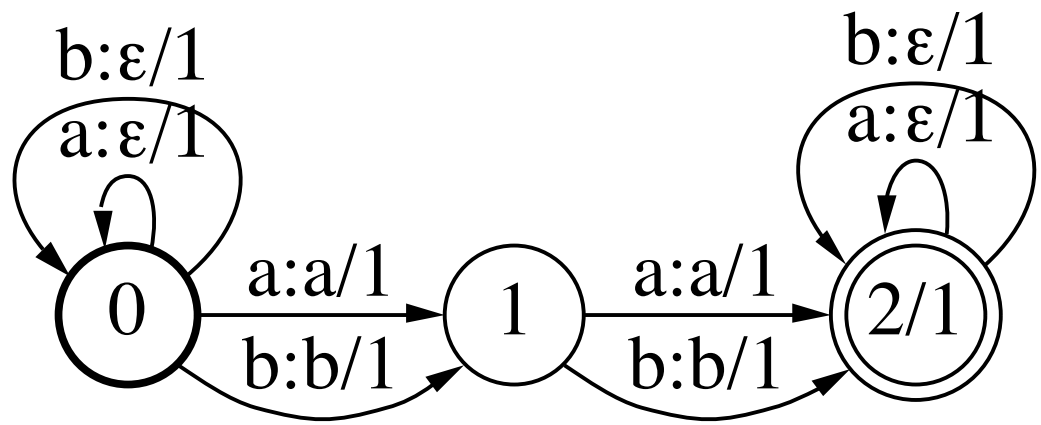
Transducer for Expected Counts





- X may be a string or an automaton representing a regular expression
- Alphabet $\Sigma = \{a, b\}$

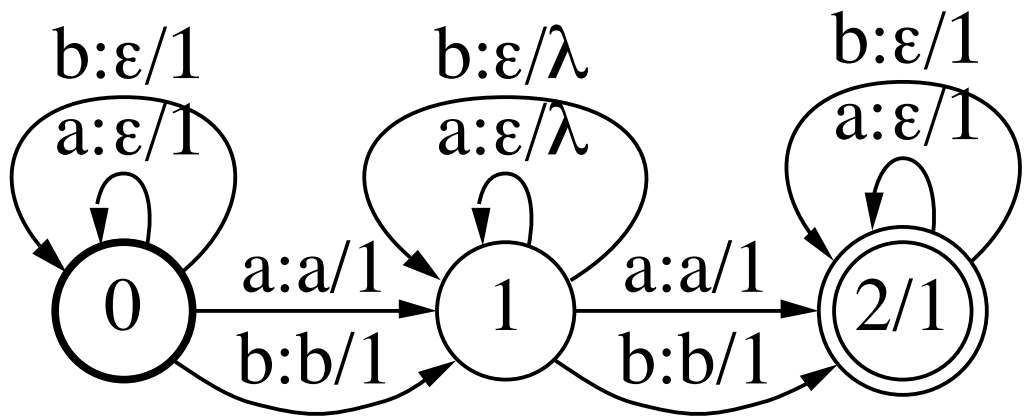
Transducer for Bigram Counts



Weighted Transducer $T, \Sigma = \{a, b\}$

(*A* o *T*) computes the expected count of each bigram (*aa*, *ab*, *ba*, *bb*) in *A*

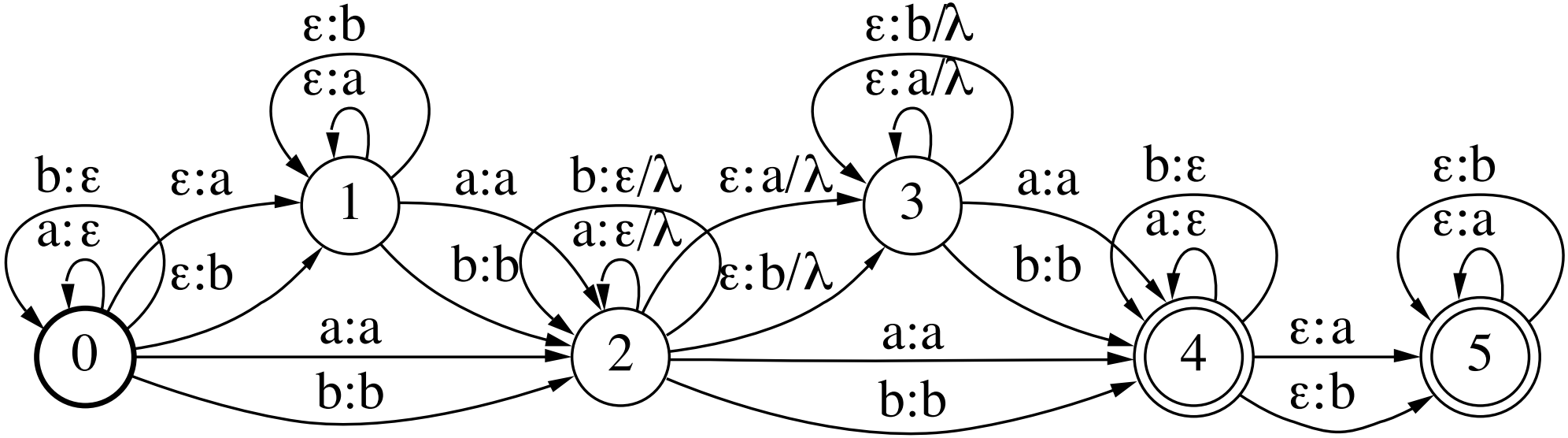
Transducer for Gappy Bigram Counts



Weighted Transducer $T', \Sigma = \{a, b\}$

(*A* o *T*) computes the expected count of each 'gappy bigram' (*aa*, *ab*, *ba*, *bb*) in *A*, with gap penalty factor λ

Gappy Bigram Kernel



Rational kernel $(T' \circ T'^{-1})$

Representation of the kernel of (Lodhi et al., 2002).

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

Application: Spoken-Dialog Classification

- Problem: assign a category (out of a finite set) to user's speech utterance
- Categories (or call-type): *Billing Services*, *Calling Plans*, *Credit*, etc.
- Data: Deployed spoken-dialog systems

Dataset	Nb of classes	Training size	Test size	Word Accuracy
HMIHY 0300	64	35,551	5,000	72.5%
VoiceTone 1	97	29,561	5,537	70.5%
VoiceTone 2	82	9,093	5,172	68.8%

Difficult Classification Tasks

- Efficiency (real-time)
- Classification
- Speech recognition

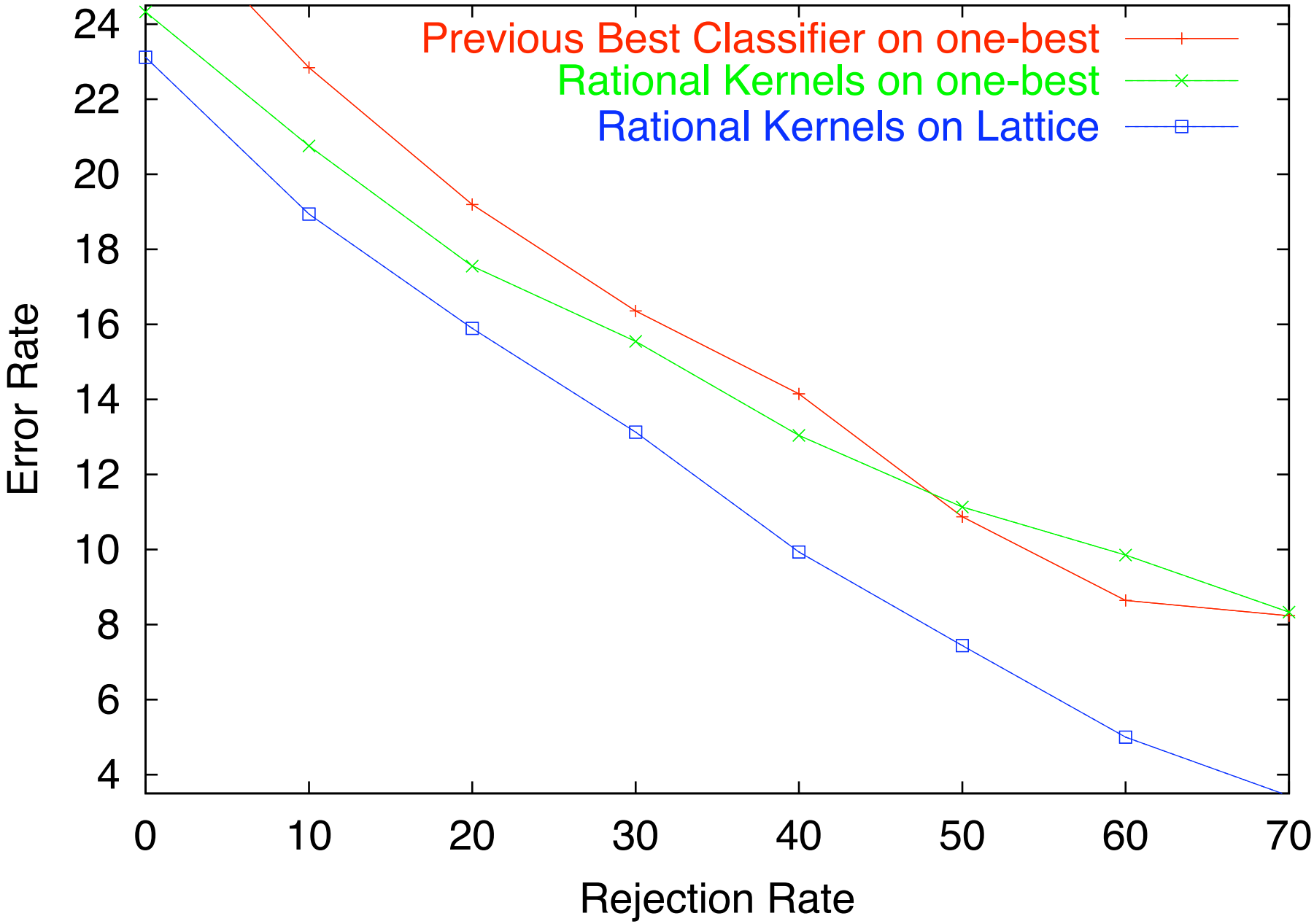



hello is question is I need somebody to know my question on the phone number speak to an address get the bill is in May and they said an eight hundred number and I don't know and then the question is I know maybe what in the long distance get the charges they maintain that I'm not letting they say

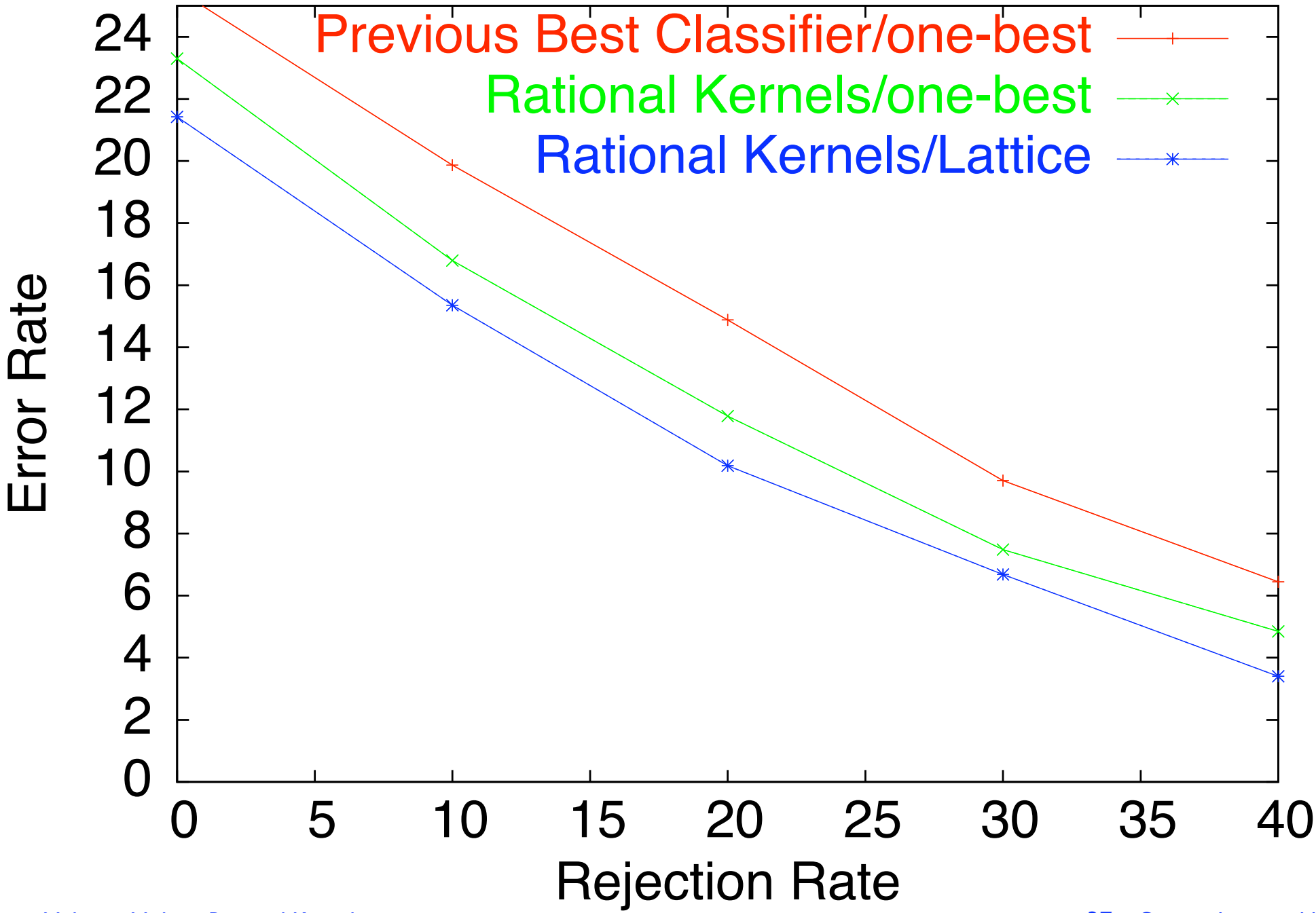
Spoken-Dialog Classification: Experiments

- Input: speech recognition word lattices
- Algorithm: rational kernels with SVMs
  - Trigram kernels (expected counts)
  - Variance kernels (trigram + variance)
  - Rejection based on top classifier output
- Implementation: FSM library and GRM Library

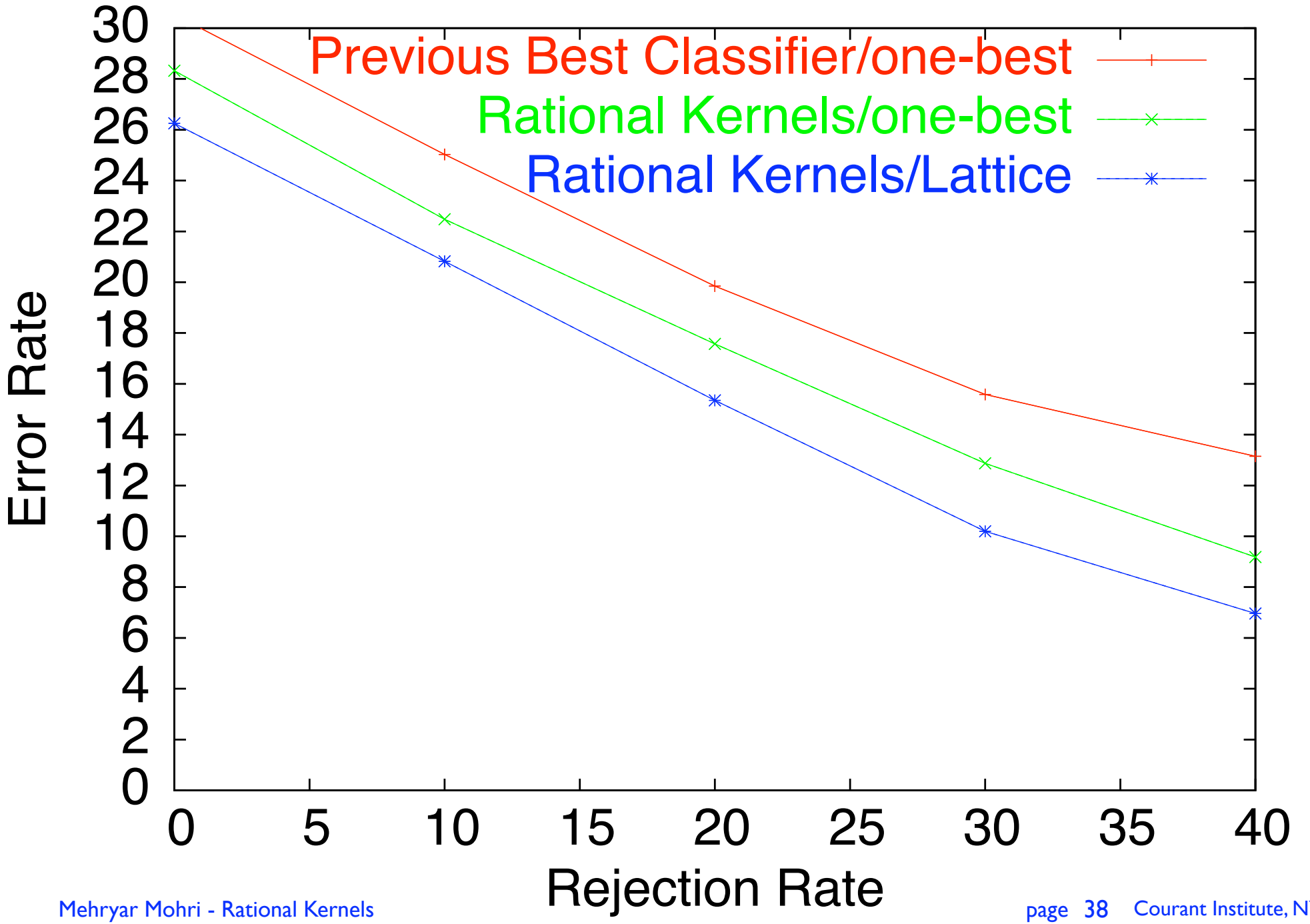
HMIHY 0300


Previous Best Classifier on one-best
Rational Kernels on one-best
Rational Kernels on Lattice
Error Rate
Rejection Rate
24
22
20
18
16
14
12
10
8
6
4
0
10
20
30
40
50
60
70

VoiceTone 1


Previous Best Classifier/one-best
Rational Kernels/one-best
Rational Kernels/Lattice
Error Rate
Rejection Rate
0
2
4
6
8
10
12
14
16
18
20
22
24
0
5
10
15
20
25
30
35
40

VoiceTone 2


Previous Best Classifier/one-best
Rational Kernels/one-best
Rational Kernels/Lattice
Error Rate
Rejection Rate
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
0
5
10
15
20
25
30
35
40

Extension: Other Moments of The Counts

$|u|_x$: number of occurrences of *x* in *u*

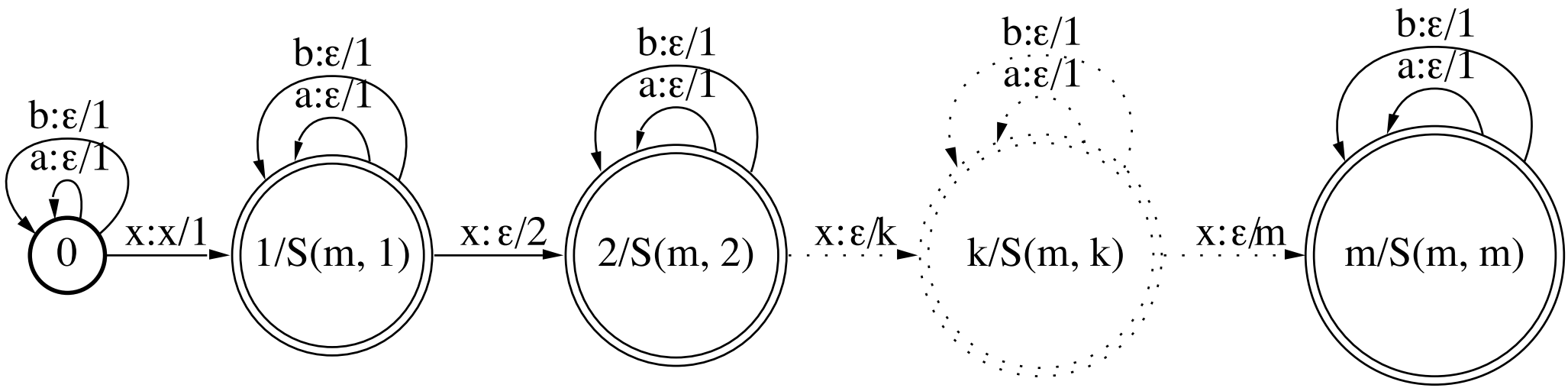
- Expected count of sequence *x* in weighted automaton *A*:

$$c(x) = \sum_{u \in \Sigma^*} |u|_x \, [[A]](u)$$

- *m*-th moment of the count of sequence *x* in weighted automaton *A*:

$$c_m(x) = \sum_{u \in \Sigma^*} |u|_x^m \, [[A]](u)$$

Transducer for *m*-th Moment of The Count



where $S(m,k)$, $k = 1, \ldots, m$, are the Stirling numbers of the second kind

($A \circ T$) computes the *m*-th moment of the count of an aperiodic sequence x in A in $O(m|x||A|)$.

Spoken-Dialog Classification: Results

- Rational kernels: best classification accuracy
- Lattice vs. one-best: significant reduction of classification error rate
- Variance kernels: accuracy gain of 1% absolute value at 15% rejection rate

Emotion Data: Samples

- Negative 
- Non-negative 

Application: Emotion Detection

- Problem: detect speaker's emotion based on speech utterance
- Categories: negative vs. non-negative (originally: positive-neutral, very-angry, somewhat-angry, very-frustrated, somewhat-frustrated, other-negative)
- Data: deployed spoken-dialog application (HMIHY 0300)

Emotion Detection: Results

- Input: 650 word lattices
- Algorithm: rational kernels with SVMs
  - 4-gram kernels (expected counts)
- Implementation: FSM library, GRM Library, DCD Library

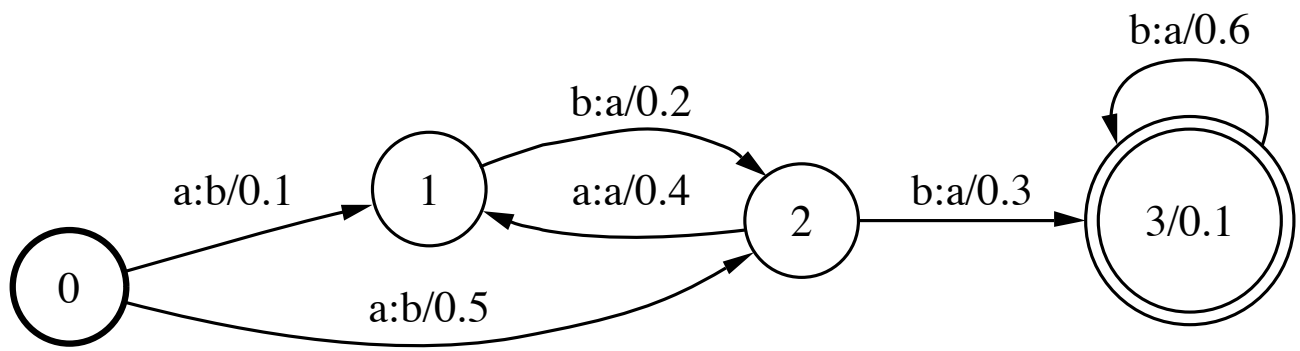
HMM-based classifiers on cepstra	76.8%
Rational Kernels on word lattices	80.6%
Rational Kernels on exact transcriptions	81.7%

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

`http://www.cs.nyu.edu/~mohri/rational.html`

Weighted Transducers



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

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

Rational Kernels Over Strings

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

$$K(x, y) = [[T]](x, y)$$

Rational Kernels Over Weighted Automata

- Definition: a kernel K is *rational* if there exists a weighted transducer T such that for all weighted automata A and B:

$$K(A, B) = \sum_{x,y} [[A]](x) \cdot [[T]](x, y) \cdot [[B]](y)$$

This definition can be generalized to the case of an arbitrary *semiring* (general operations)

Kernel Methods

- Idea:
  - Define K (called *kernel*) such that:
  - $$\Phi(x) \cdot \Phi(y) = K(x, y)$$
  - K often interpreted as a 'similarity measure'
- Benefits:
  - Efficiency: K may be much more efficient to compute than Φ and the dot product
  - Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (Mercer's condition).

Positive Definite Symmetric Kernels

- Condition equivalent to Mercer's condition in the discrete case: for all $\{x_1, \ldots, x_n\} \subseteq X$, matrix $K(x_i, x_j)_{i,j \leq n}$ is symmetric and:
  - is semi-definite positive: for all $\{c_1, \ldots, c_n\} \subseteq \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

  - or, equivalently, has non-negative eigenvalues.

Positive Definite Symmetric (PDS) Rational Kernels: Theory

- How to construct a PDS rational kernel?
- Can we combine PDS rational kernels?
- Is there a characterization of PDS rational kernels?

PDS Rational Kernels: General Construction

- T arbitrary weighted transducer
- Theorem: $T \circ T^{-1}$ defines a PDS rational kernel
- Proof ideas:
  - Kernel: $K(x,y) = \sum_{z \in \Delta^*} [[T]](x,z) \cdot [[T]](y,z)$
  - Pointwise limit of:

$$K_n(x,y) = \sum_{|z| \leq n} [[T]](x,z) \cdot [[T]](y,z)$$

  - Matrix $M_n = (K_n(x_i, x_j))_{i \leq l, j \leq l} = AA^t$
  - With: $A = [[T]](x_i, z_j)_{i \leq l, j \leq m}$

PDS Rational Kernels: Closure Properties

- Theorem: PDS rational kernels are closed under sum, product, and Kleene-Closure.
- Proof ideas (product case):

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

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

  - $(T_1 \odot T_2)$: tensor product is PDS. Thus, there exists a Hilbert Space *H* and a mapping $u \rightarrow \phi_u$ such that:
  - Thus, $K_{T_1} \odot K_{T_2}(u,v) = \langle \phi_u, \phi_v \rangle$

$$[[T_1 \cdot T_2]](x,y) = \left\langle \sum_{x_1 x_2 = x} \phi_{(x_1,x_2)}, \sum_{y_1 y_2 = y} \phi_{(y_1,y_2)} \right\rangle$$

PDS Rational Kernels: Characterization?

- Theorem: in the acyclic case, PDS rational kernels are of the form $T \circ T^{-1}$
- Theorem: PDS rational kernels $S = T \circ T^{-1}$ are closed under sum, product, and Kleene-closure
- Conjecture: all PDS rational kernels are of the form $T \circ T^{-1}$

This Tutorial

- Introduction to kernel methods
- Rational kernels
- Algorithms
- Application to text and speech processing
- Theory
- Application to computational biology

`http://www.cs.nyu.edu/~mohri/rational.html`

String Kernel Applications: Edit-Distance, Smith-Waterman

Negative Definite Kernels

- Definition: Let X be a non-empty set. A function $K: X \times X \to \mathbb{R}$ is said to be a *negative definite symmetric (NDS) kernel* if it is symmetric ($K(x, y) = K(y, x)$) and for all $\{x_1, \ldots, x_n\} \subseteq X, \{c_1, \ldots, c_n\} \subseteq \mathbb{R}$, with $\sum_{i=1}^{n} c_i = 0$,

$$\sum_{i=1}^{n} c_i c_j K(x_i, x_j) \leq 0.$$

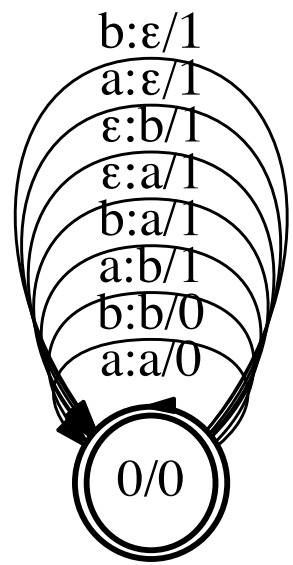
- Clearly, if K is a PDS kernel, then $-K$ is a NDS kernel but the converse is not true in general.

PDS and NDS Kernels

- Theorem: let X be a non-empty set and let $K : X \times X \to \mathbb{R}$ be a symmetric kernel, then:
  - *K* is a NDS kernel iff *exp(-tK)* is a PDS kernel for all $t > 0$;

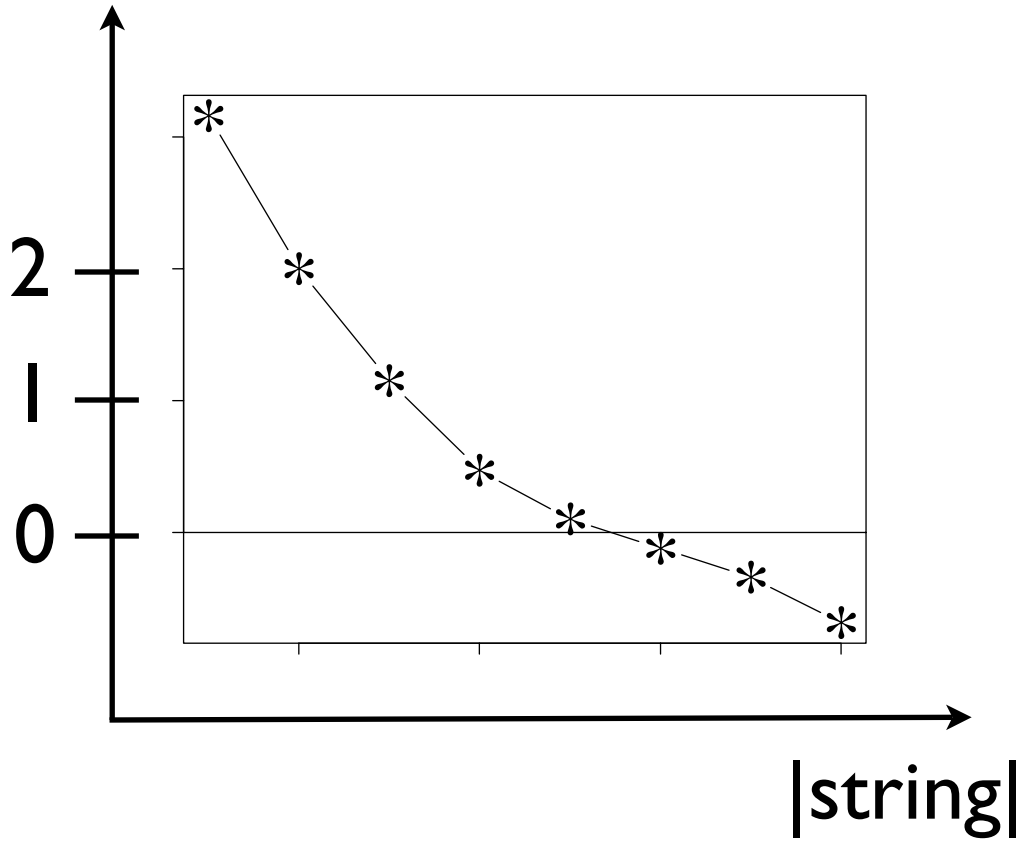
Edit-Distance

- Proposition: Let Σ be a non-empty finite alphabet. *d* is a symmetric rational kernel. But,
  - *d* is not a PDS kernel.
  - *d* is not *negative definite* if $|\Sigma|>1$



Edit-Distance

Smallest eigenvalue of *exp(-d)* for $|\Sigma| = 2$



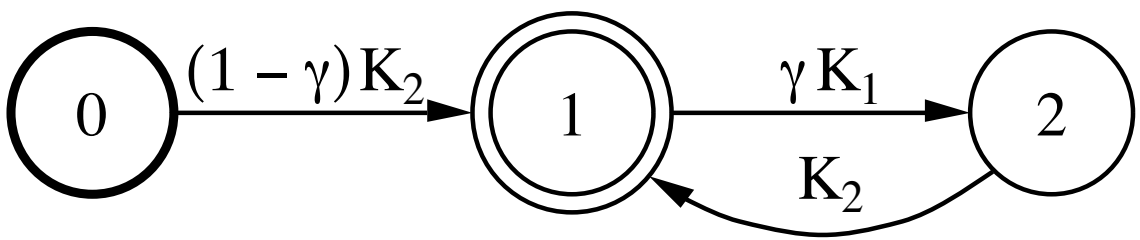
String Kernel Applications: Convolution Kernels

Convolution Kernels for Strings

- Definition (Haussler, 1999): K_1 a PDS rational transduction modeling substitutions, K_2 modeling insertions.

$$K_H = (1 - \gamma)[K_2(\gamma K_1 K_2)^*]$$

- Representation:



Relationship with Other Kernels

- Edit-distance
- Convolution kernels (Haussler, 1999)
- Mismatch kernels (Leslie *et al.*, 2003)
- Path kernels (Takimoto and Warmuth, 2004)
- Many other string kernels

Software Libraries

- FSM Library: Finite-State Machine Library. General software utilities for building, combining, optimizing, and searching weighted automata and transducers.

  http://www.research.att.com/sw/tools/fsm/

- GRM Library: Grammar Library. General software collection for constructing and modifying weighted automata and transducers representing grammars and statistical language models.

  http://www.research.att.com/sw/tools/grm/