Speech Recognition
Lecture 3: Weighted Transducer Algorithms

Eugene Weinstein
Google, NYU Courant Institute
eugenew@cs.nyu.edu
Slide Credit: Mehryar Mohri
Administrivia

- HW0 due today.
- HW1 out today, due October 17th. Start early!
- Final project presentations: December 19th 5:10-7PM, WWH 512.
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
**Finite-State Transducers**

- **Definition:** A finite-state transducer $T$ over the alphabets $\Sigma$ and $\Delta$ is a 4-tuple where $Q$ is a finite set of states, $I \subseteq Q$ a set of initial states, $F \subseteq Q$ a set of final states, and $E$ a multiset of transitions which are elements of $Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times Q$.

- $T$ defines a relation via the pair of input and output labels of its accepting paths,

$$R(T) = \{(x, y) \in \Sigma^* \times \Delta^*: I \xrightarrow{x:y} F\}.$$
Weight Sets: Semirings

- A **semiring** \((\mathbb{K}, \oplus, \otimes, 0, 1)\) is a grouping two operations, their identity elements, and the set of numbers they operate on.

- “A ring that may lack negation”. Operations:
  - **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

<table>
<thead>
<tr>
<th>Semiring</th>
<th>Set</th>
<th>$\oplus$</th>
<th>$\otimes$</th>
<th>$\bar{0}$</th>
<th>$\bar{1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean</td>
<td>${0, 1}$</td>
<td>$\lor$</td>
<td>$\land$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Probability</td>
<td>$\mathbb{R}_+$</td>
<td>$+$</td>
<td>$\times$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Log</td>
<td>$\mathbb{R} \cup {-\infty, +\infty}$</td>
<td>$\oplus_{\log}$</td>
<td>$+$</td>
<td>$+\infty$</td>
<td>0</td>
</tr>
<tr>
<td>Tropical</td>
<td>$\mathbb{R} \cup {-\infty, +\infty}$</td>
<td>$\min$</td>
<td>$+$</td>
<td>$+\infty$</td>
<td>0</td>
</tr>
</tbody>
</table>

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

A weighted transducer over a semiring: weights belong to the weight set of the semiring and are combined via its $\oplus$ and $\otimes$ operators
Shortest-Distance Problem

- **Definition**: for any regulated weighted transducer $T$, define the **shortest distance** from state $q$ to $F$ as

$$d(q, F) = \bigoplus_{\pi \in P(q, F)} w[\pi].$$

- **Problem**: compute $d(q, F)$ for all states $q \in Q$.

- Generalization of Floyd-Warshall.
Closed Semirings

(Lehmann, 1977)

- **Definition**: a semiring is closed if the closure is well defined for all elements and if associativity, commutativity, and distributivity apply to countable sums.

- **Examples**:
  - Tropical semiring.
  - Probability semiring when including infinity or when restricted to well-defined closures.
All-Pairs Shortest-Distance Algorithm

(Mohri, 2002)

- **Assumption**: closed semiring.

- **Properties**:
  
  - **Time complexity**: $\Omega(|Q|^3(T_\oplus + T\otimes + T\ast))$.
  
  - **Space complexity**: $\Omega(|Q|^2)$ with an in-place implementation.
Pseudocode

\texttt{Gen-All-Pairs}(G)

1. for \( i \leftarrow 1 \) to \(|Q|\) do
2. \hspace{1em} for \( j \leftarrow 1 \) to \(|Q|\) do
3. \hspace{2em} \( d[i, j] \leftarrow \bigoplus_{e \in E \cap P(i, j)} w[e] \)
4. \hspace{1em} for \( k \leftarrow 1 \) to \(|Q|\) do
5. \hspace{2em} for \( i \leftarrow 1 \) to \(|Q|, i \neq k \) do
6. \hspace{3em} for \( j \leftarrow 1 \) to \(|Q|, j \neq k \) do
7. \hspace{4em} \( d[i, j] \leftarrow d[i, j] \oplus (d[i, k] \otimes d[k, k]^* \otimes d[k, j]) \)
8. \hspace{2em} for \( i \leftarrow 1 \) to \(|Q|, i \neq k \) do
9. \hspace{3em} \( d[k, i] \leftarrow d[k, k]^* \otimes d[k, i] \)
10. \hspace{3em} \( d[i, k] \leftarrow d[i, k] \otimes d[k, k]^* \)
11. \hspace{2em} \( d[k, k] \leftarrow d[k, k]^* \)
Single-Source Shortest-Distance Algorithm

(Mohri, 2002)

Assumption: \( k \)-closed semiring.

\[
\forall x \in \mathbb{K}, \quad \bigoplus_{i=0}^{k+1} x^i = \bigoplus_{i=0}^{k} x^i.
\]

Idea: generalization of relaxation, but must keep track of weight added to \( d[q] \) since the last time \( q \) was enqueued.

Properties:

- works with any queue discipline and any \( k \)-closed semiring.
- Classical algorithms are special instances.
**Pseudocode**

**Generic-Single-Source-Shortest-Distance** \((G, s)\)

1. for \(i \leftarrow 1\) to \(|Q|\)
2. do \(d[i] \leftarrow r[i] \leftarrow 0\)
3. \(d[s] \leftarrow r[s] \leftarrow 1\)
4. \(S \leftarrow \{s\}\)
5. while \(S \neq \emptyset\)
6. do \(q \leftarrow \text{head}(S)\)
7. DEQUEUE\((S)\)
8. \(r' \leftarrow r[q]\)
9. \(r[q] \leftarrow 0\)
10. for each \(e \in E[q]\)
11. do if \(d[n[e]] \neq d[n[e]] \oplus (r' \otimes w[e])\)
12. then \(d[n[e]] \leftarrow d[n[e]] \oplus (r' \otimes w[e])\)
13. \(r[n[e]] \leftarrow r[n[e]] \oplus (r' \otimes w[e])\)
14. if \(n[e] \notin S\)
15. then ENQUEUE\((S, n[e])\)
16. \(d[s] \leftarrow 1\)
Notes

 Complexity:

- depends on queue discipline used.
  \[ O(|Q| + (T_\oplus + T_\otimes + C(A))|E| \max_{q \in Q} N(q) + (C(I) + C(E)) \sum_{q \in Q} N(q)) \]
- coincides with that of Dijkstra and Bellman-Ford for shortest-first and FIFO orders.
- linear for acyclic graphs using topological order.
  \[ O(|Q| + (T_\oplus + T_\otimes)|E|) \]
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
Composition

Definition: given two weighted transducers $T_1$ and $T_2$ over a commutative semiring, the composed transducer $T = T_1 \circ T_2$ is defined by

$$(T_1 \circ T_2)(x, y) = \bigoplus_z T_1(x, z) \otimes T_2(z, y).$$

Algorithm:

- Epsilon-free case: matching transitions.
- General case: $\epsilon$-filter transducer.
- Complexity: quadratic, $O(\|T_1\|\|T_2\|)$.
- On-demand construction.
Epsilon-Free Composition

- **States** $Q \subseteq Q_1 \times Q_2$.
- **Initial states** $I = I_1 \times I_2$.
- **Final states** $F = Q \cap F_1 \times F_2$.
- **Transitions**

$$E = \{((q_1, q'_1), a, c, w_1 \otimes w_2, (q_2, q'_2)) : (q_1, a, b, w_1, q_2), (q'_1, b, c, w_2, q'_2) \in Q\}.$$
Illustration

Program:  \texttt{fstcompose A.fst B.fst >C.fst}
Redundant $\varepsilon$-Paths Problem

(Mohri et al. 1996)
Redundant $\epsilon$-Paths Problem

$T_1$

$T_2$

(Mohri et al. 1996)
Redundant $\varepsilon$-Paths Problem

$T_1$

$\tilde{T}_1$

$T_2$

$\tilde{T}_2$

$F$

$T = \tilde{T}_1 \circ F \circ \tilde{T}_2.$
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
Epsilon-Removal

Definition: given weighted transducer \( T \), create equivalent weighted transducer with no epsilon-transition.

Algorithm components:

- Computation of the \( \varepsilon \)-closure at each state:
  \[
  C[p] = \{(q, d_\varepsilon[p, q]) : d_\varepsilon[p, q] \neq 0\}
  \text{ with } d_\varepsilon[p, q] = \bigoplus_{\pi \in P(p, \varepsilon, q)} w[\pi].
  \]
- Removal of \( \varepsilon \)s.
- On-demand construction.
Illustration
Algorithm Main Components

- **Shortest-distance algorithms:**
  - closed semirings: generalization of Floyd-Warshall algorithm.
  - \( k \)-closed semirings: single-source shortest-distance algorithm.

- **Complexity:** shortest-distance and removal.
  - Acyclic \( T_c \): \( O(|Q|^2 + |Q||E|(T_\oplus + T_\otimes)) \).
  - General case, tropical semiring:
    \[
    O(|Q||E| + |Q|^2 \log |Q|).\]
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
Determinization

Definition: given weighted transducer $T$, create equivalent deterministic weighted transducer.

Algorithm (weakly left divisible semirings):

- generalization of subset constructions to weighted labeled subsets: sets of state/residual weight pairs
- complexity: exponential, but lazy implementation.
- not all weighted transducers are determinizable but all acyclic weighted transducers are. Test? For some cases, using the twins property.
Illustration
Illustration

The diagram represents a probabilistic finite-state automaton. Each state is labeled with transitions and probabilities. For example, state 0 transitions to state 2 with an input 'a' and probability 0.1, or to state 1 with an input 'c' and probability 0.5. The transitions are as follows:

- State 0:
  - a: a/0.1
  - a: b/0.2
  - a: c/0.5

- State 1:
  - c: eps/0.7

- State 2:
  - a: a/0.2
  - b: b/0.3
  - b: eps/0.4

- State 3:
  - a: eps/0.5
  - c: eps/0.7

- State 4:
  - a: b/0.6
  - c: c/1.1

The diagram also includes sets of states:

- (0, eps, 0)
- (1, c, 0.4), (2, a, 0)
- (3, b, 0), (4, eps, .1)
- (4, eps, 0)

The automaton transitions through these states based on input symbols and their associated probabilities.
Non-Determinizable Transducer
Twins Property

(Choffrut, 1978; Mohri 1997)

Definition: a weighted transducer $T$ over the tropical semiring has the twins property if for any two states $q$ and $q'$ as in the figure, $c = c'$.

Theorem 1: If $T$ has the twins property, then $T$ is determinizable.

(Choffrut, 1978; Mohri, 1997)
Determinizability

(Choffrut, 1978; Mohri 1997; Allauzen and Mohri, 2002)

- A trim unambiguous weighted automaton over the tropical semiring is determinizable iff it has the twins property.

- Let $T$ be a weighted transducer over the tropical semiring. Then, if $T$ has the twins property, then it is determinizable.

**Algorithm** for testing the twins property:

- unambiguous automata: $O(|Q|^2 + |E|^2)$.

- unweighted transducers: $O(|Q|^2(|Q|^2 + |E|^2))$. 
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
Pushing

Definition: given weighted transducer $T$, create equivalent weighted transducer such that the sum (longest common prefix) of the weights (output strings) of all outgoing paths be $\overline{1}$ $(\varepsilon)$ at all states, modulo initial states.

Algorithm components:

- Single-source shortest-distance computation

\[
d[q] = \bigoplus_{\pi \in P(q,F)} w[\pi].
\]

- Reweighting: $w[e] \leftarrow (d[p[e]])^{-1}(w[e] \otimes d[n[e]])$ for each transition $e$. 
Illustration
Illustration - Label Pushing
This Lecture

- Shortest-distance algorithms
- Composition
- Epsilon-removal
- Determinization
- Pushing
- Minimization
Minimization
(Mohri, 1997, 2000, 2005)

- **Definition**: given deterministic weighted transducer $T$, create equivalent deterministic weighted transducer with the minimal number of states (and transitions).

- **Algorithm components**:
  - apply pushing to create canonical representation.
  - apply unweighted automata minimization after encoding \textit{(input labels, output label, weight)} as a single label.
Algorithm

(Mohri 1997, 2000, 2005)

- **Automata:** pushing and automata minimization, general (Hopcroft, 1971) and acyclic case (Revuz 1992).
  - acyclic case: $O(|Q| + |E|(T_\oplus + T_\otimes))$.
  - general case tropical semiring: $O(|E| \log |Q|)$.

- **Transducers:**
  - acyclic case: $O(S + |Q| + |E| (|P_{max}| + 1))$.
  - general case tropical semiring:
    $$O(S + |Q| + |E| (\log |Q| + |P_{max}|)).$$
Illustration
Illustration
References


References


