Speech Recognition
Lecture 3: Weighted Transducer Algorithms

Cyril Alauzen
Google, NYU Courant Institute
allauzen@cs.nyu.edu
Slide Credit: Mehryar Mohri
Finite-State Transducers

Definition: a finite-state transducer $T$ over the alphabets $\Sigma$ and $\Delta$ is 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 \((K, \oplus, \otimes, 0, 1)\) 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><strong>Semiring</strong></th>
<th><strong>Set</strong></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: $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$. 
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: $\Omega(|Q|^3 (T_\oplus + T_\otimes + T_\star))$.
  - Space complexity: $\Omega(|Q|^2)$ with an in-place implementation.
**Pseudocode**

\[ \text{Gen-All-Pairs}(G) \]

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

(Mohri, 2002)

- **Assumption**: $k$-closed semiring.

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

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

\[
E = \left\{ ((q_1, q'_1), a, c, w_1 \otimes w_2, (q_2, q'_2)) : \\
(q_1, a, b, w_1, q_2), (q'_1, b, c, w_2, q'_2) \in Q \right\}.
\]
Program: \( \text{fstcompose A.fst B.fst > C.fst} \)
Redundant \( \varepsilon \)-Paths Problem

(Mohri et al. 1996)

\[ T_1 \]

\[ \left( \begin{array}{cc}
0 & a:d \\
\varepsilon :\varepsilon & 1 \\
\varepsilon :\varepsilon & b : \varepsilon \\
\varepsilon :\varepsilon & 2 \\
\varepsilon :\varepsilon & c : \varepsilon \\
\varepsilon :\varepsilon & 3 \\
\varepsilon :\varepsilon & d : \varepsilon \\
\varepsilon :\varepsilon & 4
\end{array} \right) \]

\[ \tilde{T}_1 \]

\[ \left( \begin{array}{cc}
0 & a:a \\
\varepsilon :\varepsilon & 1 \\
\varepsilon :\varepsilon & b : \varepsilon \\
\varepsilon :\varepsilon & 2 \\
\varepsilon :\varepsilon & c : \varepsilon \\
\varepsilon :\varepsilon & 3 \\
\varepsilon :\varepsilon & d : \varepsilon \\
\varepsilon :\varepsilon & 4
\end{array} \right) \]

\[ T_2 \]

\[ \left( \begin{array}{cc}
0 & a:d \\
\varepsilon :\varepsilon & 1 \\
\varepsilon :\varepsilon & b : \varepsilon \\
\varepsilon :\varepsilon & 2 \\
\varepsilon :\varepsilon & c : \varepsilon \\
\varepsilon :\varepsilon & 3 \\
\varepsilon :\varepsilon & d : \varepsilon \\
\varepsilon :\varepsilon & 4
\end{array} \right) \]

\[ \tilde{T}_2 \]

\[ \left( \begin{array}{cc}
0 & a:d \\
\varepsilon :\varepsilon & 1 \\
\varepsilon :\varepsilon & b : \varepsilon \\
\varepsilon :\varepsilon & 2 \\
\varepsilon :\varepsilon & c : \varepsilon \\
\varepsilon :\varepsilon & 3 \\
\varepsilon :\varepsilon & d : \varepsilon \\
\varepsilon :\varepsilon & 4
\end{array} \right) \]

\[ F \]

\[ \left( \begin{array}{cc}
0 & \varepsilon :\varepsilon \\
\varepsilon :\varepsilon & 1 \\
\varepsilon :\varepsilon & \varepsilon :\varepsilon \\
\varepsilon :\varepsilon & 2 \\
\varepsilon :\varepsilon & \varepsilon :\varepsilon \\
\varepsilon :\varepsilon & 3 \\
\varepsilon :\varepsilon & \varepsilon :\varepsilon \\
\varepsilon :\varepsilon & 4
\end{array} \right) \]

\[ T = \tilde{T}_1 \circ F \circ \tilde{T}_2. \]
Correctness of Filter

**Proposition:** filter $F$ allows a unique path between two states of the following grid.

**Proof:** Observe that a necessary and sufficient condition is that the following sequences be forbidden: $ab$, $ba$, $ac$, and $bc$. 
Correctness of Filter

- **Proof (cont.):** Let $\sigma = \{a, b, c, x\}$, then set of sequences forbidden is exactly

$$L = \sigma^* (ab + ba + ac + bc)\sigma^*.$$ 

- An automaton representing the complement can be constructed by determ. and complementation.
Other Filters

(Pereira and Riley, 1997)

Sequential Filter.
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:
    \[
    C[p] = \{(q, d_\varepsilon[p, q]) : d_\varepsilon[p, q] \neq 0\}
    \]
    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 $T_c: O(|Q|^2 + |Q||E|(T_+ + T_\times))$.
  - 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. Transducers: also output label.
- complexity: exponential, but lazy implementation.
- not all weighted transducers are determinizable but all acyclic weighted transducers are. Test? For some cases, we can use the twins property.
Illustration
Illustration
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, $c = c'$.

\[
c = c'\]

\[
u^{-1}u' = (uv)^{-1}u'v'
\]

![Diagram](image)
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

(Mohri, 1997; 2004)

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 are $\bar{1}$ ($\varepsilon$) 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 $e$. 
Algorithm Main Components

- **Automata**: single-source shortest-distance.
  - **acyclic case**: $O(|Q| + |E|(T_\oplus + T_\otimes))$.
  - **general case tropical semiring**: $O(|Q| \log |Q| + |E|)$.
  - **general case $k$-closed semirings**
    $$O(|Q| + (T_\oplus + T_\otimes + C(A))|E| \max_{q \in Q} N(q) + (C(I) + C(E)) \sum_{q \in Q} N(q))$$
  - **general case closed semirings** $\Omega(|Q|^3 (T_\oplus + T_\otimes + T_\star))$.

- **Transducers**: $O((|P_{max}| + 1) |E|)$. 
Illustration
Ilustration - Label Pushing

\[
\begin{array}{c}
0 \xrightarrow{a:a} 1 \xrightarrow{a:a} 2 \xrightarrow{b:a} 3 \xrightarrow{c:a} 4 \xrightarrow{d:a} 5 \xrightarrow{f:a} 6 \\
0 \xrightarrow{a:a} 1 \xrightarrow{a:a} 2 \xrightarrow{b:a} 3 \xrightarrow{c:a} 4 \xrightarrow{d:a} 5 \xrightarrow{f:a} 6
\end{array}
\]
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 (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:
    $$O(S + |Q| + |E| (\log |Q| + |P_{max}|))$$.
References


References


