NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL & COMPUTER ENGINEERING

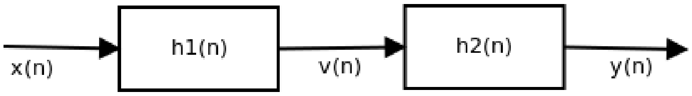
Speech and Language Processing

Spring semester 2021-2022

1^{st} Assignment

Exercise 1

Consider two linear time invariant systems, as shown in the figure below, where the output of the first system is the input to the second.



1. Show that the impulse response of the overall system is

$$h(n) = h_1(n) * h_2(n) \tag{1}$$

2. Show that

$$h_1(n) * h_2(n) = h_2(n) * h_1(n) \tag{2}$$

which implies that the overall impulse response does not depend on the order in which the systems appear.

3. Consider the function

$$H(z) = \left(\sum_{r=0}^{M} b_r z^{-r} \right) \left(\frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \right) = H_1(z) H_2(z) \tag{3}$$

where the two systems are in series. Write the difference equations of the overall system from this perspective.

4. Now consider the two systems from part (3) in the reverse order, where:

$$H(z) = H_2(z) H_1(z) \tag{4}$$

Exercise 2

As shown in the following equation, the autocorrelation function is

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m) w(n-m) x(m+k) w(n-k-m) \tag{5}$$

1. Define

$$R_n(k) = R_n(-k) \tag{6}$$

i.e. $R_n(k)$ is an even function of k.

2. Show that $R_n(k)$ can be expressed as

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)x(m-k)h_k(n-m) \tag{7}$$

where

$$h_k(n) = w(n)w(n+k) \tag{8}$$

3. Assuming that

$$w(n) = \begin{cases} \alpha^n \text{ if } n \geq 0 \\ 0 \text{ if } n < 0 \end{cases} \tag{9}$$

find the impulse response $h_k(n)$.

4. Find the z transform of $h_k(n)$ from the previous part and express $R_n(k)$ recursively based on that.

5. Repeat steps (3) and (4) for

$$w(n) = \begin{cases} n\alpha^n \text{ if } n \geq 0 \\ 0 \text{ if } n < 0 \end{cases} \tag{10}$$

Exercise 3

Consider the finite state machine that is shown below.

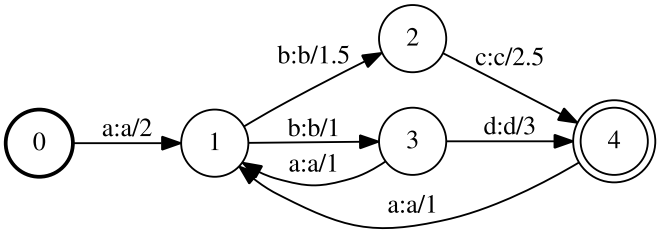
1. What is the regular expression that corresponds to the machine?

2. What is the most probable string that the fsm accepts given that we use the tropical semiring? (collect operation is min, extend operation is +) Note: the cost of the (probable) final states is taken into account only when that state is indeed final. The cost of a non-terminal state is taken into account each time that we pass from that state. The cost of state 3 is 5 and the cost of state 4 is 0.

3. What is the cost of the string abcababd;

4. What is the equivalent deterministic machine without cost?

5. What is the equivalent deterministic machine with cost?



Exercise 4

Consider the alphabet Σ = {A, B, C, D, E, F}.

1. Design the transducer that implements the Levenshtein distance, i.e. d(x, x) = 0 and d(x, ε) = d(ε, x) = d(x, y) = 1 where x and y are different characters of the alphabet Σ.

2. What is the optimal (lowest cost) mapping between the strings EDBAEDC and CDFABEA? How did you use the transducer from part (1)?

3. What is the second best mapping between the strings of part (2)?

Exercise 5

Consider the following sentence segments that use the lexicon

L = {lovely, grand, mother, grandmother}:
... lovely mother grand grandmother lovely grandmother grand mother ... (5 times)
... lovely mother lovely grandmother lovely ... (7 times)
... mother grandmother mother... (2 times)
... lovely grand lovely... (1 time)

1. Compute the bigram language model and the corresponding finite state machine with cost –logP. Use 'back-off' for the bigrams that are not observed in the above sentences.
2. What is the most probable sequence of words for the sentence without spaces: lovelygrandmothergrandmother?

Exercise 6

Consider an all pole model with a transfer function of the form

$$V(z) = \frac{1}{\prod_{k=1}^{q}(1-c_k z^{-1})(1-c_k^* z^{-1})} \tag{11}$$

where

$$c_k = r_k e^{j\theta_k} \tag{12}$$

Show that the corresponding cepstrum is

$$\hat{v}(n) = 2\sum_{k=1}^{q} \frac{(r_k)^n}{n} \cos(\theta_k n) \tag{13}$$

Exercise 7

The distributional hypothesis suggests that words that occur in similar contexts should be similar in meaning. The distributional hypothesis forms the basis for the Skipgram model of Mikolov et. al., which is an efficient way of learning the meaning of words as dense vector representations from unstructured text. The skipgram objective is to learn the probability distribution $P(C \mid T)$ where given a target word w_t, we estimate the probability that a context word w_c lies in the context window of w_t. The distribution of the probabilistic model is parameterized as follows:

$$P(C = w_c \mid T = w_t) = \frac{\exp\left(\mathbf{u}_{w_c}^{\top} \cdot \mathbf{v}_{w_t}\right)}{\sum_{w' \in \mathcal{V}} \exp\left(\mathbf{u}_{w'}^{\top} \cdot \mathbf{v}_{w_t}\right)} \tag{14}$$

where vectors $\mathbf{u}_{w_c}$ and $\mathbf{v}_{w_t}$ represent the context word w_c and the target word w_t respectively. Notice the use of softmax function and how the problem of learning embeddings in this model has been cast as a classification problem. The vectors for all the words in the vocabulary $\mathcal{V}$ can be succinctly represented in two matrices $\mathbf{U}$ and $\mathbf{V}$, where the vector in the j-th column in $\mathbf{U}$ and $\mathbf{V}$ corresponds to the context and target vectors for the j-th word in $\mathcal{V}$. Note that $\mathbf{U}$ and $\mathbf{V}$ are the parameters of the model. Answer the following questions about the Skipgram model.

1. The cross entropy loss between two probability distributions p and q, is expressed as:

$$L_{CE}(p,q) = -\sum_m p_m \log(q_m)\,. \tag{15}$$

   For a given target word w_t, we can consider the ground truth distribution $\mathbf{y}$ to be a one-hot vector of size $|\mathcal{V}|$ with a 1 only at the true context word w_c 's entry, and 0 everywhere else. The predicted distribution $\hat{\mathbf{y}}$ (same length as $\mathbf{y}$) is the probability distribution $P(C \mid T = w_t)$. The j-th entry in these vectors is the probability of the j-th word in $\mathcal{V}$ being a context word. Write a simplified expression of cross entropy loss, $L_{CE}(\mathbf{y}, \hat{\mathbf{y}})$, for the Skipgram model on a single pair of words w_c and w_t. Your answer should be in terms of $P(C = w_c \mid T = w_t)$.

2. Find the gradient of the cross entropy loss calculated in step 1 with respect to the target word vector $\mathbf{v}_{w_t}$. Your answer should be in terms of $\mathbf{y}, \hat{\mathbf{y}}$ and $\mathbf{U}$.

3. Find the gradient of the cross entropy loss calculated in step 1 with respect to each of the context word vectors $\mathbf{u}_{w_c}$. Do this for both cases $C = w_c$ (true context word) and $C \neq w_c$ (all other words). Your answer should be in terms of $\mathbf{y}, \hat{\mathbf{y}}$ and $\mathbf{v}_{w_t}$.