Full Name: _____ **Lab Section:** _____

EEL 4750 / EEE 5502 (Fall 2018) – Code #01 **Due Date: Sept. 6, 2018**

Question #1: *(1 pts)* How many hours did you spend on this homework?

Question #2: *(10 pts)* *The Inner Product*

One of the most important operations in signal processing, statistics, and machine learning is the inner product. In signals notation, the inner product between length-N signals $x[n]$ and $y[n]$ is

$$s = \sum_{n=0}^{N-1} x[n]y[n] \; . \tag{1}$$

In a linear algebra notation, the inner product of two length-N vectors is

$$s = \mathbf{x}^T \mathbf{y} \; , \tag{2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are real-valued (i.e., not complex) vectors. In MATLAB, this is expressed as

```
s = x'*y                % Compute the inner product of x and y
```

We will be using the inner product throughout the course. In this coding problem, we will use the inner product to create a simple search engine.

Before we do that, let's establish some underlying theory.

(a) Show that when $y[n] = x[n]$, the inner product is the energy of $x[n]$.

(b) Show that when $\sum_{n=0}^{N-1} |x[n]|^2 = 1$ and $\sum_{n=0}^{N-1} |y[n]|^2 = 1$, then

$$s = \sum_{n=0}^{N-1} x[n]y[n] \; . \tag{3}$$

is maximized when $x[n] = y[n]$. What is this maximum value?

Hint: use the Cauchy-Schwarz Inequality:

$$\left| \sum_{n=0}^{N-1} x[n]y[n] \right|^2 \leq \sum_{n=0}^{N-1} |x[n]|^2 \sum_{n=0}^{N-1} |y[n]|^2 \tag{4}$$

(c) Based on the above results, consider

$$c = \frac{\sum_{n=0}^{N-1} x[n]y[n]}{\sqrt{\sum_{n=0}^{N-1} |x[n]|^2} \sqrt{\sum_{n=0}^{N-1} |y[n]|^2}} \; . \tag{5}$$

What is the maximum of value of c? What is the minimum value of c?

(d) Based on the above results, determine the value of c when $x[n] = ay[n]$. Determine the value of c when $y[n] = -ax[n]$. Assume a is a real number.

(e) Based on the above results, describe why the value c is a good "similarity" metric for comparing $x[n]$ and $y[n]$. What are some of the strengths of this metric?

Side Note: The value c is often referred to as the *correlation coefficient* between $x[n]$ and $y[n]$. You may know this as the R-value that is often measured when computing a linear regression. In that context, you are computing the correlation coefficient between the line (i.e., $x[n]$) and each data point at the same horizontal locations (i.e., $y[n]$).

Question #3: *(10 pts) Creating a Search Engine*

In this problem, we will create a simple search engine using the inner product and its properties, discussed in Question # 1. From the downloaded zip file, retrieve the file called `2018_eee5502_code01_q2.mat`. The file contains three variables: a cell `vocabulary`, a cell `documents`, and matrix `counts`.

The cell `vocabulary` is a list of 4436 English words from the given documents. The cell `documents` is a list of 1734 text fragments from several old 1980's text-based adventure games. The matrix `counts` has a size of 1734×4436 and contains the frequency of each word across 1734 text fragments.

(a) From this data, write a MATLAB script that uses the *correlation coefficient* to find the text fragment in `documents` that best "match" given search terms. Accomplish this by finding the document that exhibits the largest correlation coefficient between the counts in `counts` and the chosen search terms. Submit your MATLAB script for achieving this.

(b) Submit the three matched documents and the three corresponding correlation coefficients for the following search terms (not all of the words are in `vocabulary` – ignore these):

welcome to zork
we found the crown of lord dimwit flathead in the canyon
the wizard of frobozz cast a spell

(c) The zip file contains the MATLAB function: `get_search_term`. Run `get_search_term` with your UFID as a parameter to retrieve unique search terms for you. Submit the matched text fragment and the corresponding correlation coefficient for the search terms.

(d) Use the correlation coefficient as a similarity metric to determine the two most similar texts in `documents`. Submit the two matched documents and the corresponding correlation coefficient.