The objective of this lab is for you to program in Matlab a discrete Markov process, train it on sequences of characters, and sample sequences from it. The TA will first demonstrate the results of this on several datasets of sequences, and then you will program them, replicate those results, and further explore the datasets. You can use the textbook, lecture notes and your own notes.

I Datasets

Use sequences of characters defined on N states, say $N = 5$ with states $\{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}\}$, and sequences such as `cbccd`, `edcbeddeea`, etc. Once this works, use English sentences as training sequences, where the states correspond to the English letters and punctuation, e.g. `a`–`z`,`.`,`;`, etc. You can also consider as states whole words rather than individual characters, and have the discrete Markov process generate sequences of words.

II Implementing and using discrete Markov processes

Write Matlab code to implement the following:

- *The learning problem*: given a dataset of training sequences over a set of N states, estimate the parameters:
  - The transition matrix $\mathbf{A}$ of $N \times N$.
  - The initial distribution $\boldsymbol{\pi}$ of $N \times 1$.
- *The sampling problem*: given parameters $(\mathbf{A}, \boldsymbol{\pi})$, generate a sequence of length T by sampling from the discrete Markov process.

The following Matlab functions will be useful (among others): `randi rand cumsum unique cell2mat`.

To visualize the results, create the following plots:

- Plot the parameters $(\mathbf{A}, \boldsymbol{\pi})$ as an image.
- Print the sequences sampled from the discrete Markov process.

Consider the following questions:

- Given the parameters $(\mathbf{A}, \boldsymbol{\pi})$ learnt from a set of English sequences:
  - Inspect their values and try to guess what kind of sequences will be generated with high probability.
  - How will sequences sampled from these parameters $(\mathbf{A}, \boldsymbol{\pi})$ differ from:
    - True English text sequences.
    - Random sequences of English characters. (We call "random sequences" sequences generated uniformly at random, i.e., picking each element independently from each other at random from the states.)
- If you train a discrete Markov process on a set of random sequences, how will $(\mathbf{A}, \boldsymbol{\pi})$ look like?

III What you have to submit

We provide you with a script `lab10.m` which sets up the problem (toy dataset) and plots the figures mentioned earlier. You have to code the training and sampling for a discrete Markov process and explore its behavior.

Follow these instructions strictly. Email the TA the following packed into a **single** file (`lab10.tar.gz` or `lab10.zip`) and with email subject `[CSE176] lab10`:

- Matlab code for the functions `dmptrain.m` and `dmpsample.m`. Use the templates provided. Read them carefully to understand what the function should do. They should work when called from the script `lab10.m` listed above. Note: you are not allowed to use any functions from the Matlab Toolboxes (in particular, the Statistics and Machine Learning Toolbox, or the Neural Network Toolbox). You can only use basic Matlab functions.
- A brief report (1 page) in PDF format describing your experience with the algorithms. The more extensive and insighful your exploration, the higher the grade. Be concise. Don't include code or figures, we can recreate them by running your functions. Indicate the part that each member of the group did.