Programming Hidden Markov Models (60 P)

In this exercise, you will experiment with hidden Markov models, in particular, applying them to modeling character sequences, and analyzing the learned solution. As a starting point, you are provided in the file `hmm.py` with a basic implementation of an HMM and of the Baum-Welch training algorithm. The names of variables used in the code and the references to equations are taken from the HMM paper by Rabiner et al. downloadable from ISIS. In addition to the variables described in this paper, we use two additional variables: Z for the emission probabilities of observations O, and ψ (i.e. psi) for collecting the statistics of Equation (40c).

Question 1: Analysis of a small HMM (30 P)

We first look at a toy example of an HMM trained on a binary sequence. The training procedure below consists of 100 iterations of the Baum-Welch procedure. It runs the HMM learning algorithm for some toy binary data and prints the parameters learned by the HMM (i.e. matrices A and B).

Question 1a: Qualitative Analysis (15 P)

- *Run* the code several times to check that the behavior is consistent.
- *Describe* qualitatively the solution A, B learned by the model.
- *Explain* how the solution $\lambda = (A, B)$ relates to the sequence of observations O that has been modeled.

```
In [1]: import numpy,hmm

        O = numpy.array([1,0,1,0,1,1,0,0,1,0,0,0,1,1,1,0,1,0,0,0,1,1,0,1,1,0,0,1,1,
                         0,0,0,1,0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,0,1,0,1,0,0,0,1,0,
                         0,0,1,0,1,0,1,0,0,0,1,1,1,0,1,0,0,0,1,0,0,0,1,0,1,0,1,0,0,
                         0,1,1,1,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,1,0,0,1,0,1,1,
                         1,0,0,0,1,1,0,0,1,0,1,1,1,0,0,1,1,0,0,0,1,1,0,0,1,1,0,0,1,
                         0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,1,0,1,0,0,0,1,0,0,0,1,0,
                         0,0,1,0,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,0,0,0,1,1,0,0])

        hmmtoy = hmm.HMM(4,2)

        for k in range(100):
            hmmtoy.loaddata(O)
            hmmtoy.forward()
            hmmtoy.backward()
            hmmtoy.learn()

        print('A')
        print("\n".join([" ".join(['%.3f'%a for a in aa]) for aa in hmmtoy.A]))
        print(' ')
        print('B')
        print("\n".join([" ".join(['%.3f'%b for b in bb]) for bb in hmmtoy.B]))
        print(' ')
        print('Pi')
        print("\n".join(['%.3f'%b for b in hmmtoy.Pi]))
```

```
A
0.000 1.000 0.000 0.000
0.000 0.000 0.000 1.000
```

```
1.000 0.000 0.000 0.000
0.000 0.000 1.000 0.000

B
0.720 0.280
0.800 0.200
0.000 1.000
0.880 0.120

Pi
0.000
0.000
1.000
0.000
```

Question 1b: Finding the best number N of hidden states (15 P)

For the same sequence of observations as in Question 1a, we would like to determine automatically what is a good number of hidden states $N = \mathrm{card}(S)$ for the model.

- *Split* the sequence of observations into a training and test set (you can assume stationarity).
- *Train* the model on the training set for several iteration (e.g. 100 iterations) and for multiple parameter N.
- *Show* for each choice of parameter N the log-probability $\log p(O|\lambda)$ for the test set. (If the results are unstable, perform several trials of the same experiment for each parameter N.)
- *Explain* in the light of this experiment what is the best parameter N.

```
In [2]: import solutions
        solutions.question1b(O,hmm.HMM)
```

```
N=2
trial 0 logptrain= -56.241 logptest= -61.575
trial 1 logptrain= -65.013 logptest= -66.999
trial 2 logptrain= -56.241 logptest= -61.575
trial 3 logptrain= -56.241 logptest= -61.575

N=4
trial 0 logptrain= -37.774 logptest= -36.301
trial 1 logptrain= -56.207 logptest= -61.542
trial 2 logptrain= -37.774 logptest= -36.301
trial 3 logptrain= -37.774 logptest= -36.301

N=8
trial 0 logptrain= -36.686 logptest= -34.798
trial 1 logptrain= -36.887 logptest= -38.785
trial 2 logptrain= -32.591 logptest=-129.229
trial 3 logptrain= -36.861 logptest= -38.792

N=16
trial 0 logptrain= -30.521 logptest=-161.304
trial 1 logptrain= -27.836 logptest=-131.112
trial 2 logptrain= -28.311 logptest=-150.817
trial 3 logptrain= -27.520 logptest=-173.900
```

Question 2: Text modeling and generation (30 P)

We would like to train an HMM on character sequences taken from English text. We use the 20 newsgroups dataset that is accessible via scikits-learn http://scikit-learn.org/stable/datasets/twenty_newsgroups.html. (For this, you need to install scikits-learn if not done already.) Documentation is available on the website. The code below allows you to (1) read the dataset, (2) sample HMM-readable sequences from it, and (3) convert them back into string of characters.

```
In [3]: from sklearn.datasets import fetch_20newsgroups

        # Download a subset of the newsgroup dataset
        newsgroups_train = fetch_20newsgroups(subset='train',categories=['sci.med'])
        newsgroups_test  = fetch_20newsgroups(subset='test' ,categories=['sci.med'])

        # Sample a sequence of T characters from the dataset
        # that the HMM can read (0=whitespace 1-26=A-Z).
        #
        # Example of execution:
        # O = sample(newsgroups_train.data)
        # O = sample(newsgroups_test.data)
        #
        def sample(data,T=50):
            i = numpy.random.randint(len(data))
            O = data[i].upper().replace('\n',' ')
            O = numpy.array([ord(s) for s in O])
            O = numpy.maximum(O[(O>=65)*(O<90)+(O==32)]-64,0)
            j = numpy.random.randint(len(O)-T)
            return O[j:j+T]

        # Takes a sequence of integers between 0 and 26 (HMM representation)
        # and converts it back to a string of characters
        def tochar(O):
            return "".join(["%s"%chr(o) for o in (O+32*(O==0)+64*(O>0.5))])
```

Question 2a (15 P)

In order to train the HMM, we use a stochastic optimization algorithm where the Baum-Welch procedure is applied to randomly drawn sequences of $T = 50$ characters at each iteration. The HMM has 27 visible states (A-Z + whitespace) and 200 hidden states. Because the Baum-Welch procedure optimizes for the sequence taken as input, and no necessarily the full text, it can take fairly large steps in the parameter space, which is inadequate for stochastic optimization. We consider instead for the parameters $\lambda = (A, B, \Pi)$ the update rule $\lambda^{new} = (1-\gamma)\lambda + \gamma\bar{\lambda}$, where $\bar{\lambda}$ contains the candidate parameters obtained from Equations 40a-c. A reasonable value for γ is 0.1.

- *Create* a new class `HMMChar` that extends the class `HMM` provided in `hmm.py`.
- *Implement* for this class a new method `HMMchar.learn(self)` that overrides the original methods, and implements the proposed update rule instead.
- *Implement* the stochastic training procedure and run it.
- *Monitor* $\log p(O|\lambda)$ on the test set at multiple iterations for sequences of same length as the one used for training. (Hint: for less noisy log-probability estimates, use several sequences or a moving average.)

```
In [4]: import solutions

        hmmchar = solutions.HMMChar(200,27)
        trainsample = lambda: sample(newsgroups_train.data)
        testsample  = lambda: sample(newsgroups_test.data)

        solutions.question2a(hmmchar,trainsample,testsample)
```

```
it=   0  logptrain=-165.177  logptest=-159.105
it= 100  logptrain=-142.394  logptest=-141.412
it= 200  logptrain=-135.318  logptest=-135.677
it= 300  logptrain=-132.016  logptest=-132.288
it= 400  logptrain=-130.017  logptest=-130.060
it= 500  logptrain=-128.958  logptest=-128.601
it= 600  logptrain=-128.005  logptest=-127.525
it= 700  logptrain=-127.320  logptest=-126.869
it= 800  logptrain=-126.735  logptest=-126.397
it= 900  logptrain=-126.576  logptest=-125.928
```

Question 2b (15 P)

In order to visualize what the HMM has learned, we would like to generate random text from it. A well-trained HMM should generate character sequences that have some similarity with the text it has been trained on.

- *Implement* a method `generate(self,T)` of the class `HMMChar` that takes as argument the length of the character sequence that has to be generated.
- *Test* your method by generating a sequence of 250 characters and comparing it with original text and a purely random sequence.
- *Discuss* how the generated sequences compare with written English and what are the advantages and limitations of the HMM for this problem.

```
In [5]: print("original:\n"+tochar(sample(newsgroups_test.data,T=250)))
        print("\nlearned:\n"+tochar(hmmchar.generate(250)))
        print("\nrandom:\n" +tochar(solutions.HMMChar(200,27).generate(250)))
```

```
original:
OW OF PAST FAILURES THAT SHARED THESE CHARACTERISTICS WITH THE NOTION OF HIDDEN CANDIDA INFECTION  THER

learned:
R A FAIM O HEO HORS AXCRINFOS ANLUNTUST THE NANDIRTERFELIOBEL DTET OF POEW TTEE  ON HOERPESTIRDETACTY B

random:
EJPIAAFMTPOENGEFQEARNCE  X RBTE VCWYYTZJHTIQIB WCLSGJEMUMQQCRXHQG NUSCKQNGHAVAWKDSLTLSVU GLVNWDHZBNCMAW
```