STAT 391
Homework 2
Out April 5, 2018
Due April 12, 2018

©Marina Meilă
mmp@cs.washington.edu

[Problem 1 – Estimation of small probabilities – Not Graded]

This is a follow-up to Problem 4 from Homework 1. Estimate a model for "Lincoln-English" using the following texts (available in the files `lincoln_text.txt`).

```
"What constitutes the bulwark of our own liberty and independence?  It is not our frowning
battlements, our bristling sea coasts, the guns of our war steamers, or the strength of our
gallant and disciplined army.  These are not our reliance against a resumption of tyranny
in our fair land.  All of them may be turned against our liberties, without making us stronger
or weaker for the struggle.  Our reliance is in the love of liberty which God has planted
in our bosoms.  Our defense is in the preservation of the spirit which prizes liberty as
the heritage of all men, in all lands, every where.  Destroy this spirit, and you have planted
the seeds of despotism around your own doors."
```

–From the September 13, 1858 Speech at Edwardsville, IL

```
``It is not merely for to-day, but for all time to come that we should perpetuate for our
children's children this great and free government, which we have enjoyed all our lives.
I beg you to remember this, not merely for my sake, but for yours.  I happen temporarily
to occupy this big White House.  I am a living witness that any one of your children may
look to come here as my father's child has.  It is in order that each of you may have through
this free government which we have enjoyed, an open field and a fair chance for your industry,
enterprise and intelligence; that you may all have equal privileges in the race of life,
with all its desirable human aspirations.  It is for this the struggle should be maintained,
that we may not lose our birthright--not only for one, but for two or three years.  The nation
is worth fighting for, to secure such an inestimable jewel.''
```

–August 22, 1864 Speech to the 166th Ohio Regiment

a Estimate $\hat{\theta}_a, \hat{\theta}_b, \ldots \hat{\theta}_z$ the parameters of the Lincoln-English language model from the above texts. This is the same question as Homeowork 1 Problem 4,a. What is *different* for the previous version of this question, is that for letters that never appear, you will set $\theta^{ML} = 0$.

b. Let S be the sample space $\{\texttt{a,b,c,...z}\}$, with $m = |S| = 26$. Determine the sets $S_0, S_1, \ldots S_n$, where $S_k = \{j \in S, n_j = k\}$.

c. Let $r_k = |S_k|$ and r be the number of unique letters observed in the Lincoln-English corpus above. Verify that $r = \sum_{k=1}^{n} r_k$, $m = \sum_{k=0}^{n} r_k$, and $n = \sum_{k=1}^{n} k r_k$.

Problem 2 – Estimate letter probabilities from text

Submit the code used for this problem through the Assignments web page.
This problem requires you to use Maximum Likelihood estimation and the smoothing methods from Lecture 1 to estimate the probabilities of the letters in the English alphabet.

We assume that sentences in a language are generated by sampling letters independently from the alphabet { A, B, C, ... Z }. Spaces and punctuation are ignored. For instance, the probability of the sentence `` ``Who's on first?'' `` is

$$\theta_W \theta_H \theta_O^2 \theta_S^2 \theta_N \theta_F \theta_I \theta_R \theta_T$$

because the sentence contains (W, H, O, S, O, N, ... T) in this order. You will estimate the parameters $\theta_{A:Z}$ of this simple model from the text below (also available in `hw2-mlk-letter-estimation.txt`).

> `To save man from the morass of propaganda, in my opinion, is one of the chief aims of education. Education must enable one to sift and weigh evidence, to discern the true from the false, the real from the unreal, and the facts from the fiction. [...] The function of education, therefore, is to teach one to think intensively and to think critically. But education which stops with efficiency may prove the greatest menace to society.`
>
> Martin Luther King, Jr., *The Purpose of Education*

First, preprocess this text: Turn all letters to lower (or upper) case, eliminate spaces and punctuation. Then proceed with the questions of the homework.

a. Get the sufficient statistics: Count the number of times each letter appears in the sentence. These are the counts n_a, n_b, $\ldots$ n_z. Print out the counts $n_{a:j}$ only.

What is the fingerprint r_k, $k = 0, \ldots$ of this data set?

For the following estimation questions, choose one letter for each type (i.e., for $k = 0, 1, \ldots$ choose a letter i for which $n_i = k$), and display the estimate only for those selected letters.

b. Compute the ML estimates $\theta^{ML}_{A:Z}$ of the letter probabilities.

c. Compute now the Laplace $\theta^{Lap}_{A:Z}$ or Bayes (with $n_0 = 1$) estimates $\theta^{B}_{A:Z}$ of the same probabilities

[**d.** – **Optional, extra credit**] Compute the Witten-Bell estimates $\theta^{WB}_{A:Z}$ of the same probabilities.

[**e.** – **Optional, extra credit**] Compute the smoothed[1] Good-Turing estimates $\theta^{GT}_{A:Z}$ of the same probabilities.

f. Compute the Ney-Essen estimates $\theta^{NE}_{A:Z}$ of the same probabilities, taking $\delta = 1$.

g. Now use the estimates you obtained to compute the (log-)probability of the text in either one of `hw2-test-letter-estimation.txt` or `hw2-test-letter-estimation-large.txt`. Also compute the log-probability of the *training data* `hw2-mlk-letter-estimation.txt`).

Print out the results obtained by each method (including ML). Which method gives the highest likelihood of the new data? Which method gives the highest likelihood of the training data?

Problem 3 – ML estimation

Sam rolls a die n times, and observes a data set $\mathcal{D}$ with counts $n_1, \ldots n_6$. He is told that the die is not a fair one: the odd faces have the same probabilitly of coming up, denoted by θ_o, the even faces also have the same probabilitly of coming up, denoted by θ_e, but $\theta_o \neq \theta_e$, i.e. the distribution P defined by the die is given by $\theta_1 = \theta_3 = \theta_5 = \theta_o$ and $\theta_2 = \theta_4 = \theta_6 = \theta_e$.

a. Write the expression of the probability $P(3, 2, 1, 1, 6)$.

b. Write the expression of $l(\theta_o, \theta_e)$ the log-likelihood of data set $\mathcal{D}$ as a function of θ_o, θ_e and the counts $n_{1:6}$.

c. Transform $l(\theta_o, \theta_e)$ into a function of one variable, $l(\theta_e)$.

d. Now find the ML estimate of θ_e by equating the derivative of $l(\theta_e)$ with 0.

[**e–Extra credit**] Explain why the result above is intuitive/not surprising/natural.

Problem 4 – The ML estimate as a random variable

Submit the code used for this problem through the Assignments web page.

Consider the coin toss experiment ($m = 2$) with $\theta_1 = 0.3141$. The coin is tossed $n = 100$ times, obtaining independent outcomes from which we estimate the parameters $\theta^{ML}_1, \theta^{ML}_0 = 1 - \theta^{ML}_1$ by the max likelihood method.

1. What is the set of possible values S_{θ_1} for θ^{ML}_1 ? Does the true θ_1 belong to S_{θ_1}?

[1] These use the approximation of $E[r_{k+1}/E[r_k]$.

2. Write the expression of the probability of each outcome in S_{θ_1}, i.e the probability that $\theta_1^{ML} = j/n$ for $j = 0, 1, \ldots n$.

3. Make a plot of the probability distribution of θ_1^{ML}. Preferably, this should be a "stem and flower" plot (the `stem` function in Matlab) like in figure 4.2 in the book. To avoid numerical overflow/underflow in the computation of the probabilities, consider using logarithms for the intermediate computations. The final results should not be in logarithm form, however. Take figure 4.2 as an example of how your plot should look like.

4. Let $\epsilon = 0.02$. Answer using the probability distribution computed previously (numerical answer only is OK):

$$\begin{aligned} \delta_{abs} &= P[|\theta_1^{ML} - \theta_1| > \epsilon] \ = ? \\ \delta_{rel} &= P[\frac{|\theta_1^{ML} - \theta_1|}{\theta_1} > \epsilon] \ = ? \end{aligned}$$

5. For $\epsilon = 0,\ 0.005,\ 0.01,\ 0.015,\ \ldots,\ 1$ plot the graph of $\delta(\epsilon) = P[|\theta_1^{ML} - \theta_1| > \epsilon]$ vs. ϵ. (Note that there is a linear time recursive algorithm to compute $\delta(\epsilon)$.) Is the function $\delta(\epsilon)$ monotonically increasing, decreasing[2] or neither?

6. Simulate tossing the coin with $\theta_1 = 0.3141$ $n = 100$ times and compute θ_1^{ML}. What is the value you θ_1^{ML} have obtained, and what are the absolute and relative errors $|\theta_1^{ML} - \theta_1|$, $\frac{|\theta_1^{ML} - \theta_1|}{\theta_1}$?

7. Let θ_1' have the value θ_1^{ML} of the previous question. Repeat questions 3– 6 using "the guess" θ_1' instead of "the truth" θ.

[Problem 5 – Rare outcomes and data set size – Not Graded]

Here we will be concerned with a biased coin for which outcome 1 has a very low probability, i.e $0 < \theta_1 < \theta_0 << 1$. Assume our experiment consists of n independent tosses of this coin.

1. What is the probability $p_0 = P(n_1 = 0)$ that the outcome sequence contains no 1's ? Write the answer as a function of θ_1 and n.

2. What is the probability $p_1 = P(n_1 = 1)$ that the outcome sequence contains a single 1, ? Write the answer as a function of θ_1 and n.

3. Note that for $n = 1$ (and in general for small values of n) $p_0 >> p_1$. We are interested in the value(s) of n for which p_0 and p_1 are of comparable size. This gives us the answer to the question: how many times do I need to toss the coin in order to have a chance of observing the rare outcome once (p_1) that is about the same to that of not observing it (p_0)?

To obtain the answer, solve the equation

$$p_0 \ = \ p_1$$

for n a function of θ_1. [You will not necessarily obtain an integer].

4. Compute the above n for $\theta_1 = 10^{-3}$, 10^{-4}, 10^{-5}.

[2] Some math textbook call non-decreasing (respectively non-increasing) a function which can only increase (decrease) or stay constant. I'm using the simpler term increasing for such a function.