Austin Anderson
CSCI 5622
Homework 7 Analysis

In the lda.py script, new_phi and m_step were altered according to the assignment's directions.

Alterations to new_phi

In new_phi, a new value for the variational parameter phi is calculated. This is accomplished by looking at the appropriate column of the Beta matrix for the given word index. The exponential digamma terms are calculated for the sum of over the gamma vectors and for the vector of gammas. The exponential term is calculated as a vector and is element-wise multiplied by the extracted Beta column. Then the normalization term is calculated as the sum of the unscaled phi result divided by the count term to result in the proper scaling result. The phi vector is returned. All computations are done using numpy calls to improve computation speed.

Alterations to m_step

In the m_step function, the new Beta matrix is calculated. In the e_step function call, topic_counts is computed based on the final phi computed in the inner loop of the document iteration loop. This set of counts is passed to the m_step function as an input. Then we know that the Beta columns can be approximated by taking the counts of the phi matrix represented in topic_counts and for each topic, dividing by the total counts. The comments in the code gives us that each topic is a row, so we simply take the sum over the rows then scale each row by one over that relevant sum. The row sums are taken using a numpy call on the topic_counts matrix. The scaling is done in a loop. A check is done to prevent divides by zero. The resulting matrix is returned.

Comments on Code Performance

The code runs the full training dataset in 3.8 minutes on a laptop with 4 cores and 8 GB of RAM. Performance was maintained by using numpy calls for all computations and using matrix and vector operations in place of loops. The only place in the code where a loop is run is to scale the new Beta matrix rows. This could probably be converted to a matrix or vector operation, but as the whole thing runs well below 10 minutes with default settings, it wasn't pursued. Memoization of the digamma function was also explored, but as one of the digamma calls is done as a vector operation, the basic memoization scheme failed during development debugging. Again, as the code was performing fine, the memoization effort was abandoned.

Results

The resulting topics.txt list was generated using the default code settings and running the lda script against the train.dat and voc.dat datasets. The resulting topics did not match the order of the provided example results on github, but they generally matched (ie the topic about money showed up as topic 1 in my results instead of topic 6 for the example results). Generally, though, the first 10 topics made sense from a human perspective, so the code is believed to be functioning properly.