# Foundations of Graphical Models: Homework 2

Due: Tuesday 2020-11-24

The total page limit is three pages, though you may use extra pages for figures, and your code can be any length. Please use the LaTeX template on the website. You should submit both the writeup and code to Courseworks.

## Problem 1

Implement either a mixed-membership model or a matrix factorization model. You can choose which inference method you use and should briefly discussed what informed your choice. You can for example implement variational inference for latent Dirichlet allocation or MAP estimation for matrix factorization. As in homework 1, you are expected to implement the inference algorithm yourself.

Apply your code to data and discuss what you learned. We encourage you to apply your code to data sets that aren't widely used in machine learning. We have also included data sets on the course website.

You can plot and discuss whatever you'd like. As for homework 1, a good guide line is to work through one or two iterations of Box's loop. Among the plots, we would like to see a figure depicting the convergence of your inference procedure (for example, if you are performing variational inference you can plot the ELBO as a function of iteration; if you're using MCMC, you can examine the log joint or samples for latent variables of interest). We would also like to see a plot that checks the model on held-out data.

What challenges did you encounter while building and fitting your model?

## Problem 2

Write an "aspirational abstract" for your final project. Note you are not committed to deliver everything you mention on the abstract. Rather, preparing the abstract is a chance to think concretely and envision a successful final project.

## Online Data Sets

**MovieLens** One of the online data sets contains movie ratings submitted by users on MovieLens, a movie recommendation website. The data set contains 100,000 ratings applied to 9,000 movies by 600 users. There's a `README.txt` file in the folder that contains more information about the data set. (Alternatively, read an online version.)

**AP** The other provided data set contains the text of 2,246 articles from the Associated Press. After unzipping, you'll find three files in the folder `ap`: `ap.dat`, `ap.txt`, and `vocab.txt`.

`ap.txt` is an XML file that contains the full text of every article.

You'll probably want to work with `ap.dat`, which contains the counts of each word for each article. Every line is a different article in a bag-of-words format. The first number in each line is the total number of words in that article. Following this number, the rest of the line contains word counts in the format `word_index:count`. For example, the line "`5 0:1 5:2 140:1 2031:1`" indicates a 5-word article that has 1 occurrence of word 0, 2 occurrences of word 5, 1 occurrences of word 140, and 1 occurrence of word 2031.

Finally, the file `vocab.txt` contains a list of each word, zero-indexed to match the indices in `ap.dat`.