Reflection - Stochastic Variational Inference - Hoffman et. al 2013
Nicholas Hoernle **February 12, 2018**

Hoffman et. al tackle the scalability of Bayesian posterior analysis by using a stochastic method for updating the global variational parameters in variational inference (in LDA these global parameters correspond to the topic Dirichlet parameters while the local parameters are the per-document topic proportions and the per-word topic assignments). Due to the large sizes of modern data (they present some nice examples of analysis on large text corpora) this is an important objective and they certainly achieve (1) an impressive speed-up over variational inference and (2) a higher evaluation log predictive probability by using the stochastic optimisation methods. Their methods center on stochastically updating the the global parameters in the coordinate ascent method presented by Blei 2003, by evaluating the local parameter updates on a randomly drawn subset of data (instead of the entire dataset) and completing a weighted update of parameters after every step. As the global parameters are updated at every step, their convergence is faster than the alternative of having to iterate through the entire dataset before an update is executed.

The introduction of natural gradients is important in this context for two reasons. Firstly, the natural gradient corrects for the ineffective dissimilarity metric provided by Euclidean distance between parameter vectors of probability distributions. Secondly, the natural gradients provide a convenient computational efficiency by not requiring the pre-multiplication of the Fisher information matrix that would be required in the standard gradient calculation.

In response to the data streaming problem (a), it depends on the nature of the documents. If a strong assumption of exchangeability among documents in time holds then the variational objective is constant (i.e. if the topics within the documents are not evolving over time, or if there is no possible bias in the sampling over time, then the joint probability $p(\beta, \theta, z, X)$ is constant and thus the variational inference approximation should solve the same optimisation problem, even as more data arrives. For both cases, of an evolving distribution of topics over time and a static distribution, stochastic variational inference will allow a streaming variational approximation. In the first case, the 'forgetting' nature of the update will approximate the most recent batch data with a higher weight in the model (thus converging to the more recent distributions that are present). For the later case, there is no difference between this and the case presented in the paper.