Online Learning for Latent Dirichlet Allocation

Matt Hoffman (Princeton Computer Science Dept., Columbia Statistics Dept.)
David Blei (Princeton Computer Science Dept.)
Francis Bach (INRIA, Ecole Normale Supérieure)

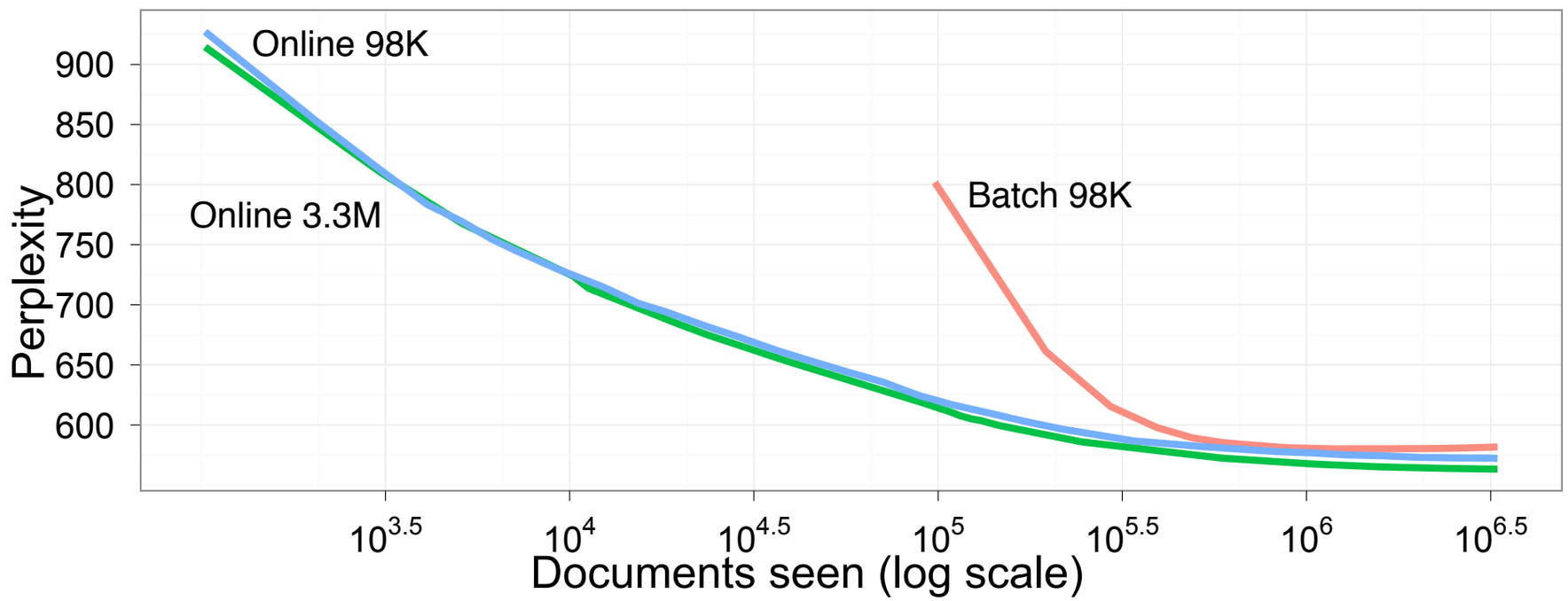
Latent Dirichlet Allocation

- Unsupervised Bayesian hierarchical model of document collections
- Discovers a set of latent “topic” distributions over words shared across corpus
- Our goal: Efficiently discover topics from very large datasets

Variational Bayes

- Deterministic alternative to Markov Chain Monte Carlo (MCMC) for fitting Bayesian hierarchical models
- Basic idea: minimize KL divergence between convenient parametric distribution and true posterior
- Typically converges in significantly fewer iterations than MCMC
- Still scales linearly with number of documents, so slow for very large datasets

Wikipedia Experiment



We can fit VB objective with stochastic natural gradient instead of batch updates

Online algorithm has nearly converged before batch has finished one iteration on 3% of Wikipedia