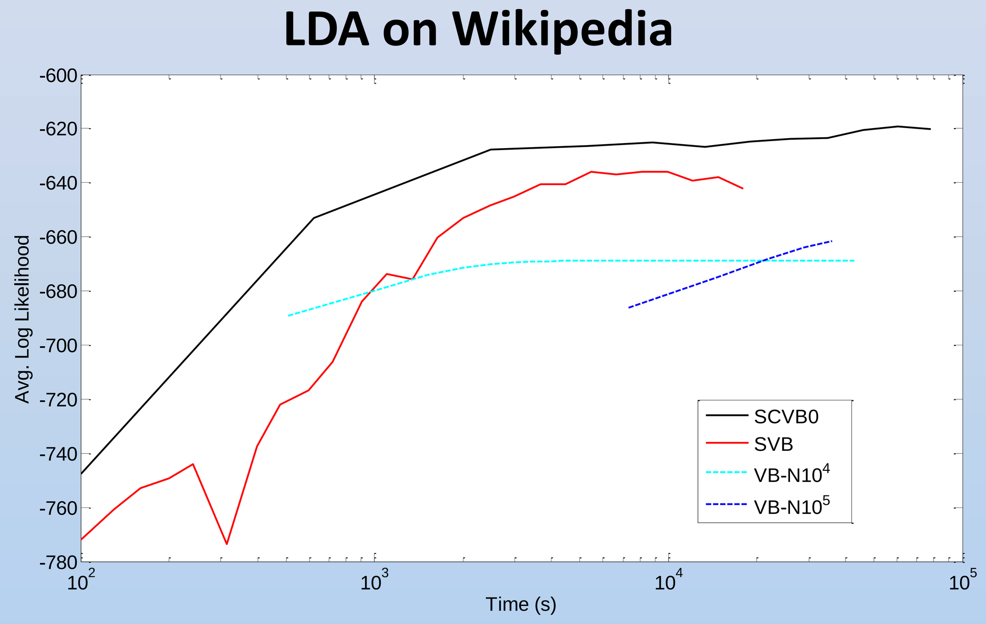
# Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation

James Foulds[1], Levi Boyles[1], Christopher DuBois[2], Padhraic Smyth[1], Max Welling[3]

[1]University of California Irvine, Computer Science [2]University of California Irvine, Statistics [3]University of Amsterdam, Computer Science

## Motivation

**LDA on Wikipedia**




### Stochastic Optimization for ML

- **Batch algorithms**
  - •While (not converged)
    - •Process the entire dataset
    - •Update parameters
- **Stochastic algorithms**
  - •While (not converged)
    - •Process a subset of the dataset
    - •Estimate quantities needed for an update, and extrapolate
    - •Update parameters

### Collapsed Variational Bayes for LDA

- Maintain variational distributions for the topic of each token
- Mean field assumption
- CVB0 (Asuncion et al., 2009)

$$\gamma_{ijk} :\propto \frac{N^{\Phi\neg ij}_{w_{ij}k} + \eta_{w_{ij}}}{N^{Z\neg ij}_k + \sum_w \eta_w}(N^{\Theta\neg ij}_{jk} + \alpha_k)$$

$$N^Z_k \triangleq \sum_{ij} \gamma_{ijk}$$

$$N^\Theta_{jk} \triangleq \sum_{i} \gamma_{ijk}$$

$$N^\Phi_{wk} \triangleq \sum_{ij: w_{ij}=w} \gamma_{ijk}$$

**Advantages of the collapsed representation**

- **Simpler**, faster and fewer update equations
- **Better mixing** for Gibbs sampling
- **Better variational bound** for VB

## Stochastic CVB0

### How to estimate CVB0 statistics?

- Pick a random word (i,j) from the corpus

$$N^Z_k \triangleq \sum_{ij} \gamma_{ijk}$$

$$E[N^Z_k] = \text{Total words in the corpus} \times \gamma_{ijk}$$

- In an online algorithm, we cannot store the variational parameters, but we can *update them*

$$\gamma_{ijk} :\propto \frac{N^\Phi_{w_{ij}k} + \eta_{w_{ij}}}{N^Z_k + \sum_w \eta_w}(N^\Theta_{jk} + \alpha_k)$$

- Keep an online average of the CVB0 statistics

$$\mathbf{N}^\Theta_j := (1-\rho^\Theta_t)\mathbf{N}^\Theta_j + \rho^\Theta_t \hat{\mathbf{N}}^\Theta_j$$
$$\mathbf{N}^\Phi := (1-\rho^\Phi_t)\mathbf{N}^\Phi + \rho^\Phi_t \hat{\mathbf{N}}^\Phi$$
$$\mathbf{N}^Z := (1-\rho^\Phi_t)\mathbf{N}^Z + \rho^\Phi_t \hat{\mathbf{N}}^Z$$

### The Full Algorithm

- Optional burn-in passes per document
- Minibatches
- Operating on sparse counts

$$\mathbf{N}^\Theta_j := (1-\rho^\Theta_t)\mathbf{N}^\Theta_j + \rho^\Theta_t C_j \gamma_{ij}$$
$$\mathbf{N}^\Phi := (1-\rho^\Phi_t)\mathbf{N}^\Phi + \rho^\Phi_t \hat{\mathbf{N}}^\Phi$$
$$\mathbf{N}^Z := (1-\rho^\Phi_t)\mathbf{N}^Z + \rho^\Phi_t \hat{\mathbf{N}}^Z$$

where $\hat{\mathbf{N}}^\Phi = \frac{C}{|M|}\sum_{ij\in M}\mathbf{Y}^{(ij)}$ and $\hat{\mathbf{N}}^Z = \frac{C}{|M|}\sum_{ij\in M}\gamma_{ij}$

$$\mathbf{N}^\Theta_j := (1-\rho^\Theta_t)^{m_{aj}}\mathbf{N}^\Theta_j + C_j\gamma_{aj}(1-(1-\rho^\Theta_t)^{m_{aj}})$$

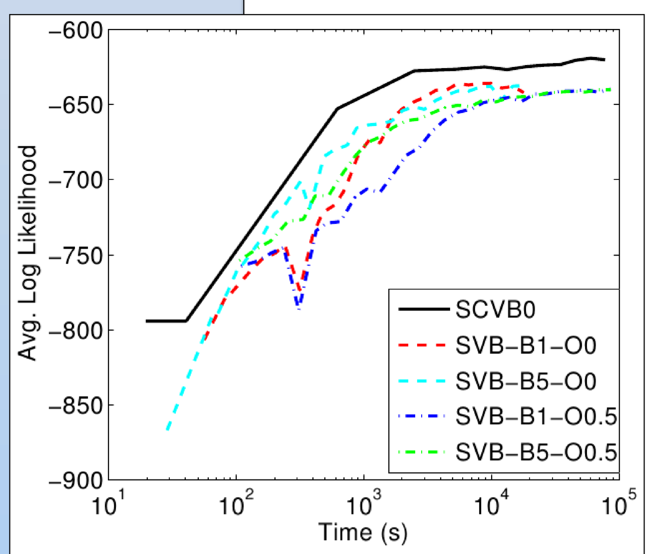
- Randomly initialize $\mathbf{N}^\Phi$, $\mathbf{N}^\Theta$; $\mathbf{N}^Z := \sum_w \mathbf{N}^\Phi_w$
- For each minibatch $M$
  - $\hat{\mathbf{N}}^\Phi := \mathbf{0}$; $\hat{\mathbf{N}}^Z := \mathbf{0}$
  - For each document $j$ in $M$
    - For zero or more "burn-in" passes
      - For each token $i$
        - Update $\gamma_{ij}$
        - Update $\mathbf{N}^\Theta_j$
    - For each token $i$
      - Update $\gamma_{ij}$
      - Update $\mathbf{N}^\Theta_j$
      - $\hat{\mathbf{N}}^\Phi_{w_{ij}} := \hat{\mathbf{N}}^\Phi_{w_{ij}} + \frac{C}{|M|}\gamma_{ij}$
      - $\hat{\mathbf{N}}^Z := \hat{\mathbf{N}}^Z + \frac{C}{|M|}\gamma_{ij}$
  - Update $\mathbf{N}^\Phi$
  - Update $\mathbf{N}^Z$

SCVB0 is a Robbins Monro **stochastic approximation** algorithm for finding the fixed points of (a variant of) CVB0

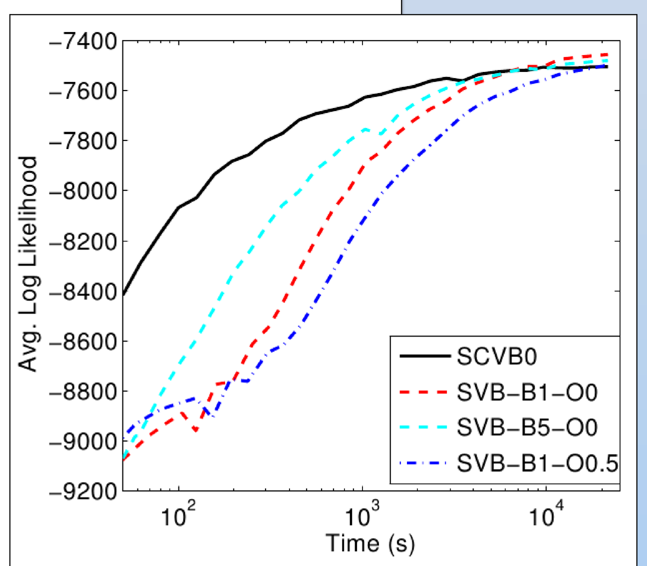
Theorem: with an appropriate sequence of step sizes, **SCVB0 converges to a stationary point of the MAP**, with adjusted hyper-parameters

## Experimental Results

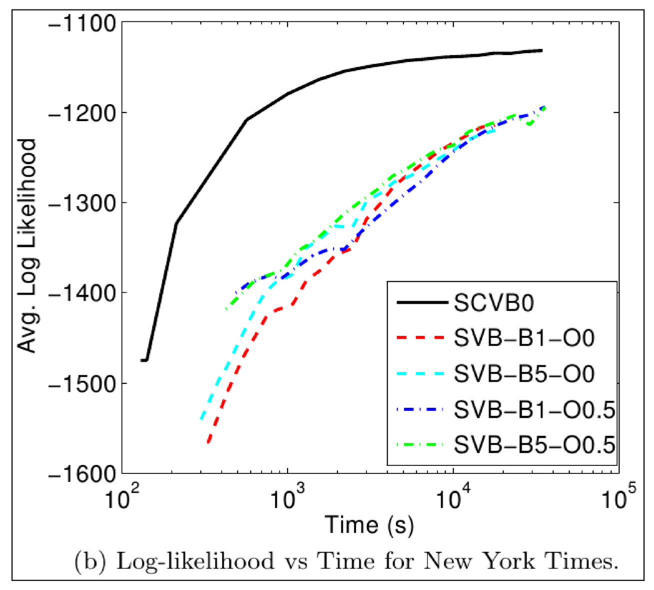
### Large-Scale Experiments




(c) Log-likelihood vs Time for Wikipedia.




(a) Log-likelihood vs Time for PubMed Central.



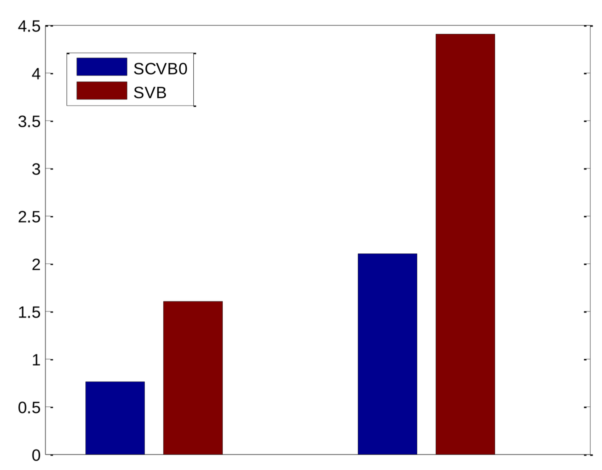

(b) Log-likelihood vs Time for New York Times.

### Small-Scale Experiments

- **Real-time** or near real-time results are important for EDA applications
- Topics on-demand

> Here are 20 collections of related words. Some words may not seem to "belong" with the other words. Count the total number of words in each collection that don't "belong."

Mean # of errors




NIPS (5 Seconds) New York Times (60 Seconds)

- We introduced stochastic CVB0 for LDA, which is **fast, easy** to implement, and **accurate**
- Experimental results show SCVB0 is useful for both **large-scale** and **small-scale** analysis
- Future work: Exploit **sparsity**, **parallelization**, **non-parametric** extensions