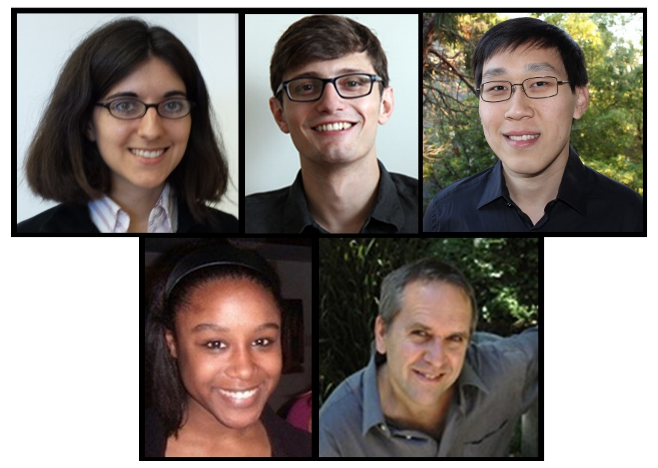
Streaming Variational Bayes

Tamara Broderick **Nick Boyd** **Andre Wibisono** **Ashia C. Wilson** **Michael I. Jordan**



Overview

- Large, streaming data sets are increasingly the norm
- Inference for Big Data has generally been non-Bayesian
- Advantages of Bayes: complex models, coherent treatment of uncertainty, etc.

We deliver:

- **SDA-Bayes**, a framework for **S**treaming, **D**istributed, **A**synchronous Bayesian inference
- Experiments demonstrating streaming topic discovery with comparable predictive performance to non-streaming algorithms
  - Corpuses used are Wikipedia (3.6M documents) and the scientific journal Nature (350K documents)

Background

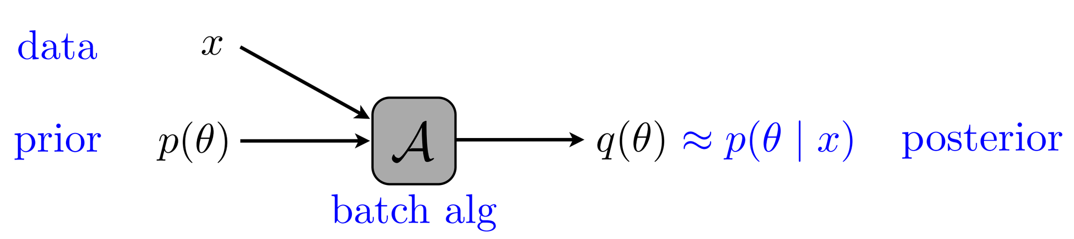
- **Posterior:** adjusted belief about unknowns θ after observing data x
- **Variational Bayes (VB):** finds approximate posterior by solving an optimization problem (minimize Kullback-Leibler divergence)
- **Batch VB:** solves a VB optimization problem using coordinate descent
  - Requires passing over the data multiple times
- **Stochastic Variational Inference (SVI)**: solves a VB optimization problem using stochastic gradient descent
  - Requires specifying the data size in advance (so not streaming)
  - Generally much better predictive performance after a single data pass than batch VB

SDA-Bayes: Streaming

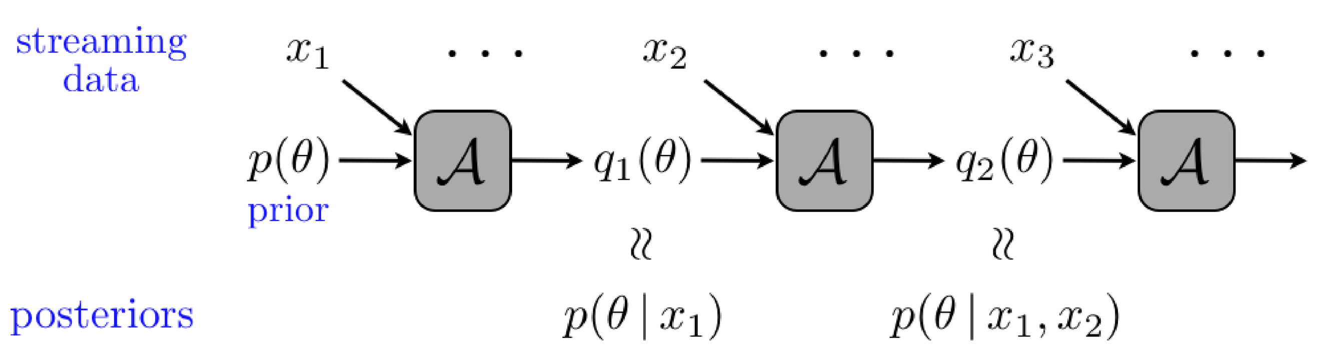
- Can iteratively update posterior after new data using Bayes' theorem

$$p(\theta \mid x_{\text{old}}, x_{\text{new}}) \propto p(\theta \mid x_{\text{old}}) \cdot p(x_{\text{new}} \mid \theta)$$

- Choose any batch approximation $\mathcal{A}$ to the posterior



- Can iterate as long as approximation has same form as prior



SDA-Bayes: Distributed

- Posteriors calculated in parallel can be combined using Bayes' rule:

$$p(\theta \mid x_1, \ldots, x_N) \propto \left[\prod_{n=1}^{N} p(x_n \mid \theta)\right] p(\theta) \propto \left[\prod_{n=1}^{N} p(\theta \mid x_n)\, p(\theta)^{-1}\right] p(\theta)$$

- Can combine approximated posteriors in similar fashion
- If the prior and approximate posterior are in the same exponential family, the update is simply vector addition
  - Sufficient statistic $T(\theta)$, prior parameter ξ_0, nth approximate posterior parameter ξ_n

$$p(\theta \mid x_1, \ldots, x_N) \approx q(\theta) \propto \exp\left\{\left[\xi_0 + \sum_{n=1}^{N}(\xi_n - \xi_0)\right] \cdot T(\theta)\right\},$$

SDA-Bayes: Asynchronous

- Each worker iterates the following steps.
  1. Collect a new data point x.
  2. Copy the master posterior parameter locally: $\xi^{(\text{local})} \leftarrow \xi^{(\text{post})}$
  3. Compute the local approximate posterior parameter ξ using $\mathcal{A}$ with $\xi^{(\text{local})}$ as the prior parameter
  4. Return $\Delta\xi := \xi - \xi^{(\text{local})}$
- Each time the master receives $\Delta\xi$ from a worker, it updates synchronously: $\xi^{(\text{post})} \leftarrow \xi^{(\text{post})} + \Delta\xi$

Case Study: Latent Dirichlet Allocation (LDA)

- LDA: a model for the content of documents
- **Topic**: a theme potentially shared by multiple documents
- (Unsupervised) inference problem: discover the topics and identify which topics occur in which documents

Experimental Setup

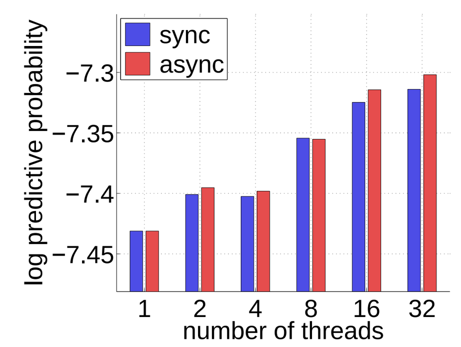
- We compare SDA-Bayes with a batch VB primitive for $\mathcal{A}$ ("Streaming Variational Bayes") to SVI
- All algorithms learn topics using an LDA model with 100 topics
- **Data**: 3.6M Wikipedia and 350K Nature documents for training; 10K Wikipedia and 1K Nature documents for testing.
- Documents seen in **minibatches** (small groups) rather than one by one
- **Log predictive probability**: on held-out words in held-out testing documents
  - We use an approximation of this as our performance measure in experiments; higher is better

Results

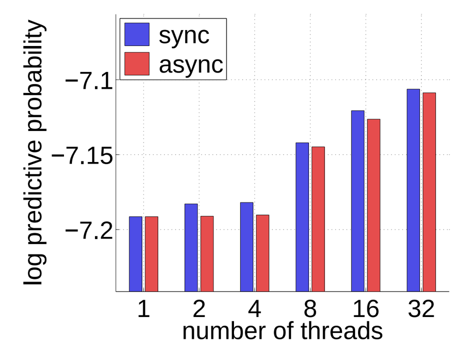
- SDA performs at least as well as SVI, an algorithm not designed for the streaming setting (32 threads and 1 thread depicted in table)

	Wikipedia			Nature		
	32-SDA	1-SDA	SVI	32-SDA	1-SDA	SVI
Log pred prob	**−7.31**	−7.43	−7.32	−7.11	−7.19	**−7.08**
Time (hours)	**2.09**	43.93	7.87	**0.55**	10.02	1.22

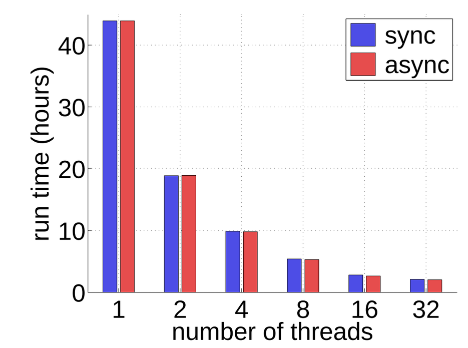
- Using more threads in SDA improves performance and runtime



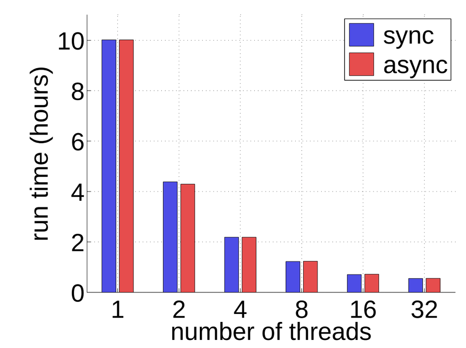
(a) Wikipedia



(b) Nature

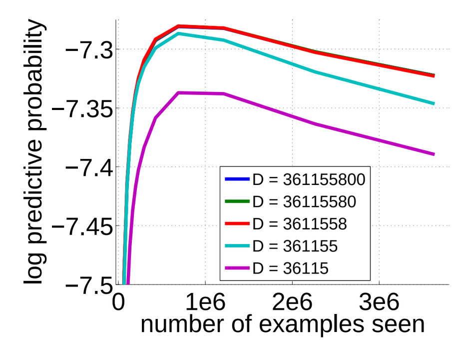


(c) Wikipedia

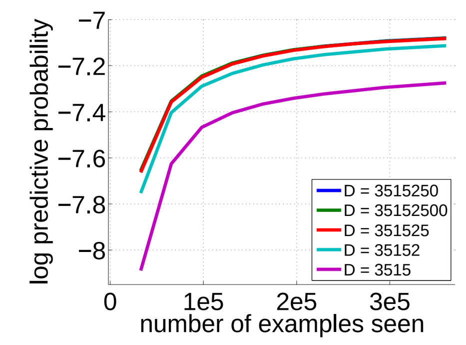


(d) Nature

- SVI is sensitive to the pre-specified number of documents D



(a) Wikipedia



(b) Nature

References

[1] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Neural Information Processing Systems*, 2013.

[2] M. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.