Streaming Variational Bayes

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

Overview

Overview

- Big Data inference generally non-Bayesian

Overview

- Big Data inference generally non-Bayesian
- Why Bayes? Complex models, coherent treatment of uncertainty, etc.

Overview

- Big Data inference generally non-Bayesian
- Why Bayes? Complex models, coherent treatment of uncertainty, etc.
- We deliver: **SDA-Bayes**, a framework for **S**treaming, **D**istributed, **A**synchronous Bayesian inference

Overview

- Big Data inference generally non-Bayesian
- Why Bayes? Complex models, coherent treatment of uncertainty, etc.
- We deliver: **SDA-Bayes**, a framework for **S**treaming, **D**istributed, **A**synchronous Bayesian inference
  - Experiments on streaming topic discovery (Wikipedia: 3.6M docs, Nature: 350K docs)

Background

Background

- **Posterior**: belief about unobserved parameters θ after observing data x

Background

- **Posterior**: belief about unobserved parameters θ after observing data x

- **Variational Bayes** (VB): approximate posterior by solving optimization problem (min KL divergence)

Background

- **Posterior**: belief about unobserved parameters θ after observing data x
- **Variational Bayes** (VB): approximate posterior by solving optimization problem (min KL divergence)
- **Batch VB**: solves VB using coordinate descent

Background

- **Posterior**: belief about unobserved parameters θ after observing data x
- **Variational Bayes** (VB): approximate posterior by solving optimization problem (min KL divergence)
- **Batch VB**: solves VB using coordinate descent
- **Stochastic Variational Inference** (SVI): solves VB using stochastic gradient descent

Background

- **Posterior**: belief about unobserved parameters θ after observing data x
- **Variational Bayes** (VB): approximate posterior by solving optimization problem (min KL divergence)
- **Batch VB**: solves VB using coordinate descent
- **Stochastic Variational Inference** (SVI): solves VB using stochastic gradient descent

NOT STREAMING

SDA-Bayes: **S**treaming

SDA-Bayes: **S**treaming

- Posterior update is iterative:

SDA-Bayes: **S**treaming

- Posterior update is iterative:

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

SDA-Bayes: **S**treaming

- Posterior update is iterative:

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

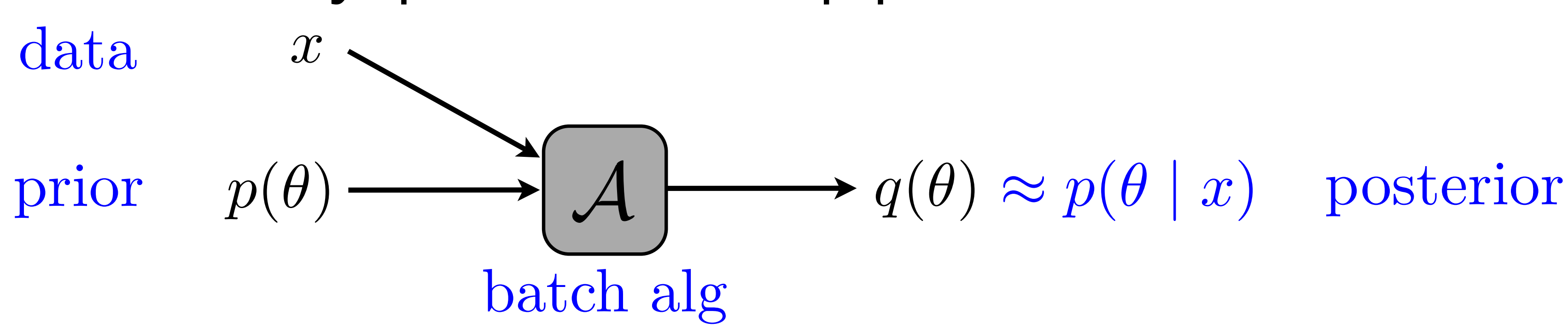
- Choose any posterior approximation A:

SDA-Bayes: **S**treaming

- Posterior update is iterative:

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

- Choose any posterior approximation A:

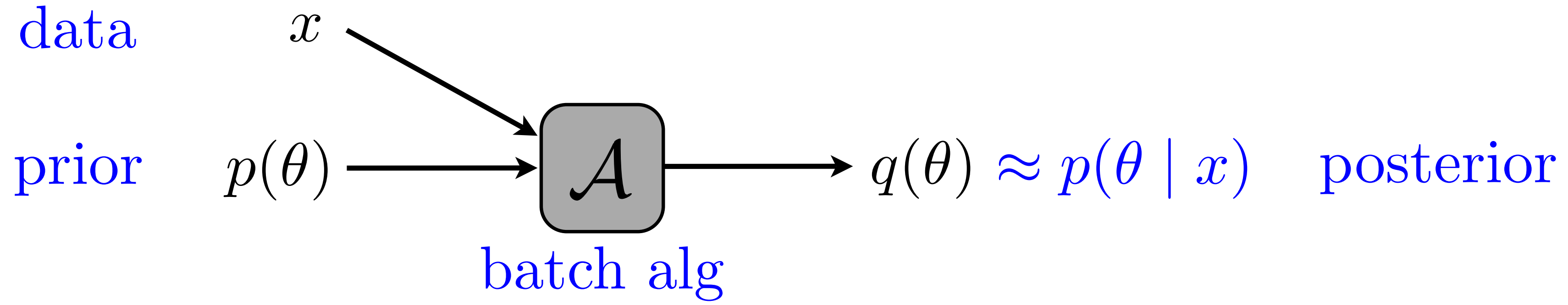


SDA-Bayes: Streaming

- Posterior update is iterative:

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

- Choose any posterior approximation $\mathcal{A}$:



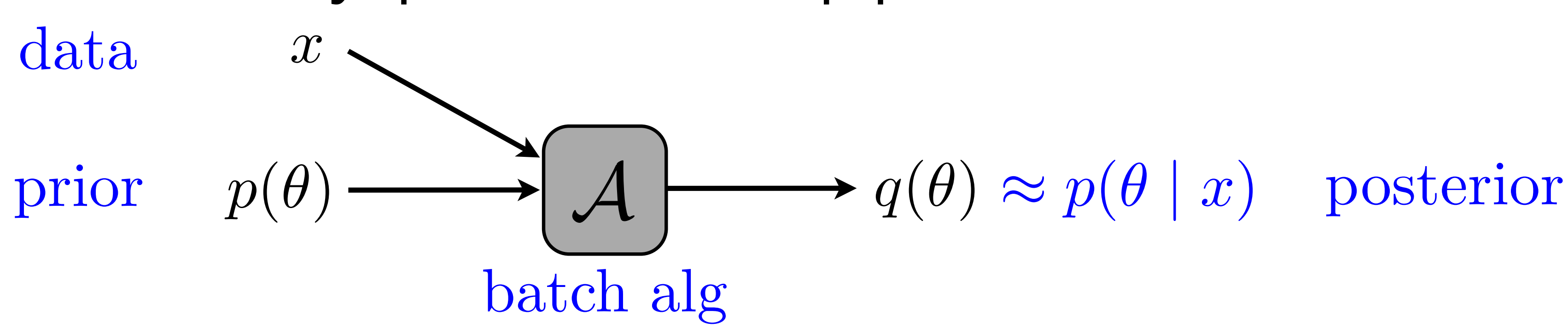
- Iterate approximation if matches prior form:

SDA-Bayes: **S**treaming

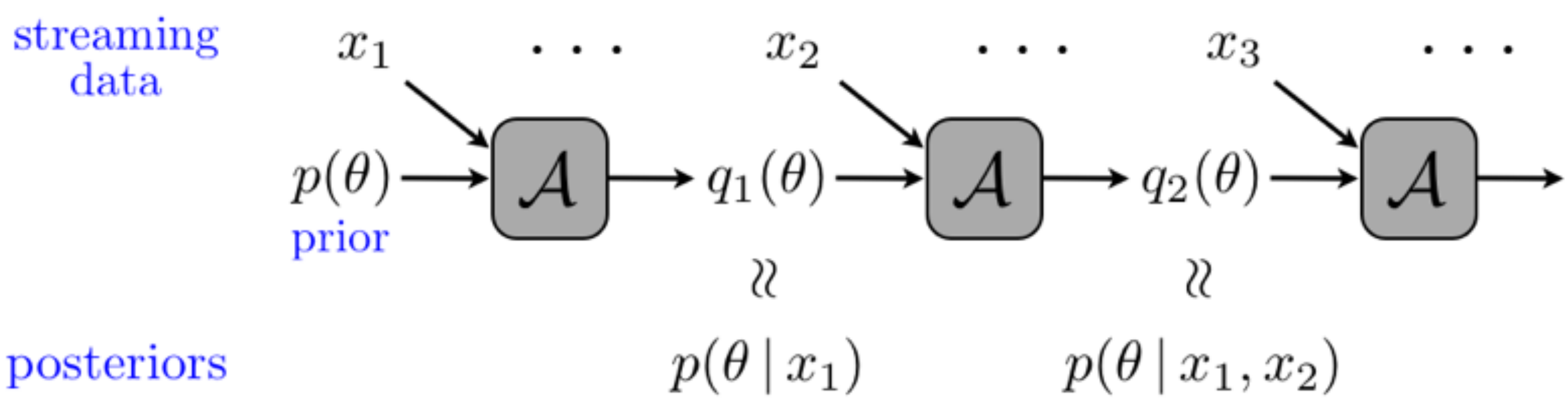
- Posterior update is iterative:

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

- Choose any posterior approximation A:



- Iterate approximation if matches prior form:



S**D**A-Bayes: **D**istributed

S**D**A-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

S**D**A-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

$$p(\theta \mid x_1, \ldots, x_N)$$

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

S**D**A-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

$$p(\theta \mid x_1, \ldots, x_N)$$

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

S**D**A-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

$$p(\theta \mid x_1, \ldots, x_N)$$

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

- Could substitute approximation found by A instead

S**D**A-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

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

- Could substitute approximation found by A instead

- Update is just addition if prior and approximate posterior are in same exponential family:

SDA-Bayes: **D**istributed

- Can calculate posteriors in parallel and combine with Bayes' Rule:

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

- Could substitute approximation found by A instead

- Update is just addition if prior and approximate posterior are in same exponential family:

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

SD**A**-Bayes: **A**synchronous

- Each worker iterates:

  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: LDA

Case Study: LDA

- **Topic**: theme potentially shared by multiple documents

Case Study: LDA

- **Topic**: theme potentially shared by multiple documents
- **Latent Dirichlet Allocation** (LDA): a topic model

Case Study: LDA

- **Topic**: theme potentially shared by multiple documents
- **Latent Dirichlet Allocation** (LDA): a topic model
- (Unsupervised) inference problem: discover the topics and identify which topics occur in which documents

Experimental Setup

Experimental Setup

- **SDA-Bayes** with batch VB for A **vs. SVI** (not designed for streaming)

Experimental Setup

- **SDA-Bayes** with batch VB for A **vs. SVI** (not designed for streaming)
- Training: 3.6M Wikipedia, 350K Nature
- Testing: 10K Wikipedia, 1K Nature

Experimental Setup

- **SDA-Bayes** with batch VB for A **vs. SVI** (not designed for streaming)
- Training: 3.6M Wikipedia, 350K Nature
- Testing: 10K Wikipedia, 1K Nature
- Performance measure: **log predictive probability** on held-out words in held-out testing documents; higher is better

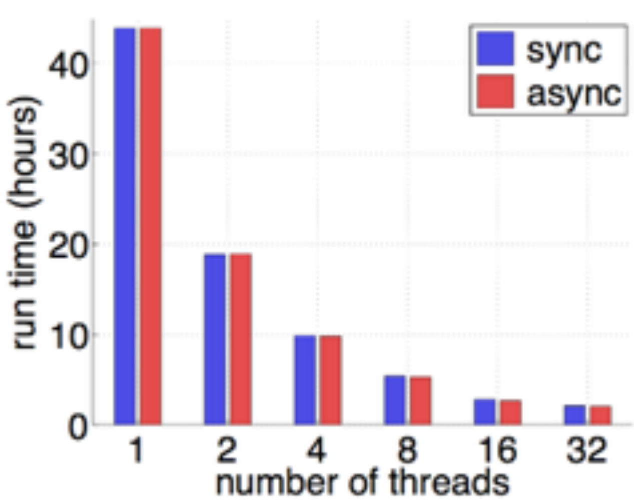
Results

- **SDA-Bayes** (streaming) as good as **SVI** (not streaming); 32 threads and 1 thread shown

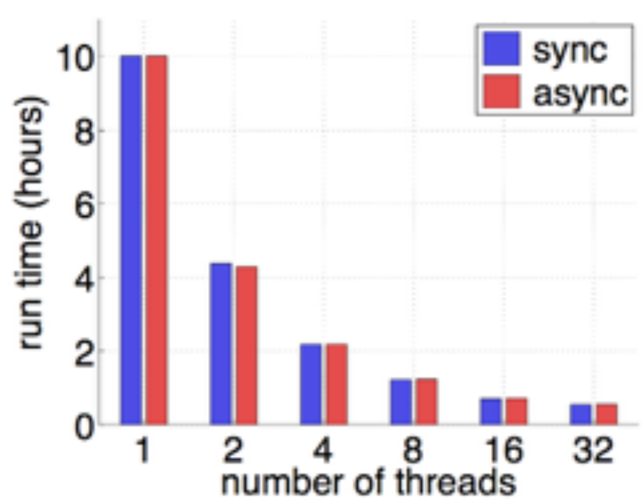
	Wikipedia			Nature		
	32-SDA	1-SDA	SVI	32-SDA	1-SDA	SVI
Log pred prob	$-\mathbf{7.31}$	-7.43	-7.32	-7.11	-7.19	$-\mathbf{7.08}$
Time (hours)	**2.09**	43.93	7.87	**0.55**	10.02	1.22

Results

- More threads in **SDA** improves runtime and performance



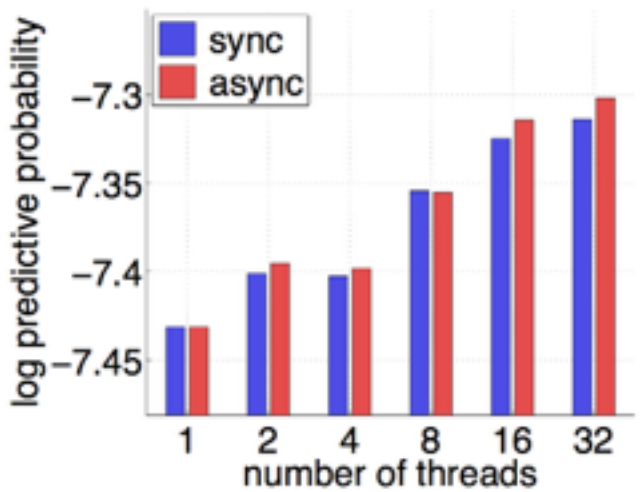
(c) Wikipedia



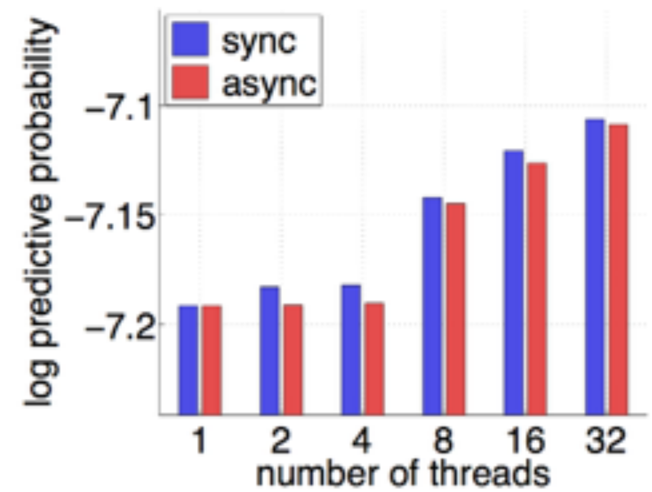
(d) Nature

Results

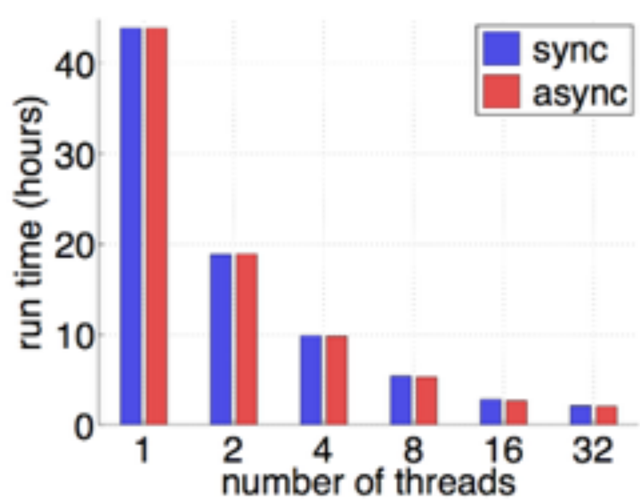
- More threads in **SDA** improves runtime and performance



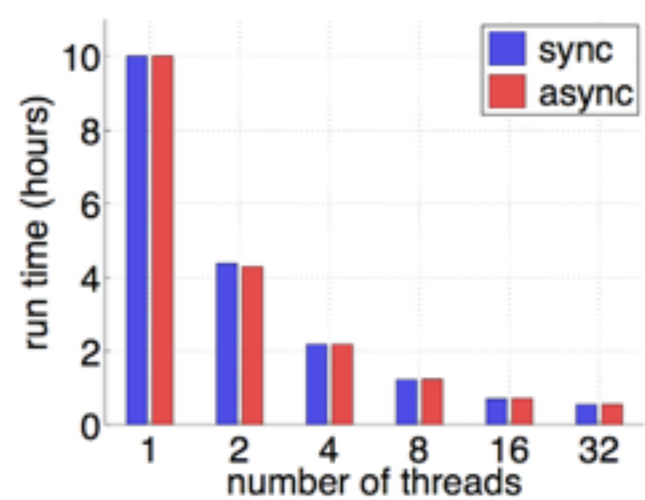
(a) Wikipedia



(b) Nature



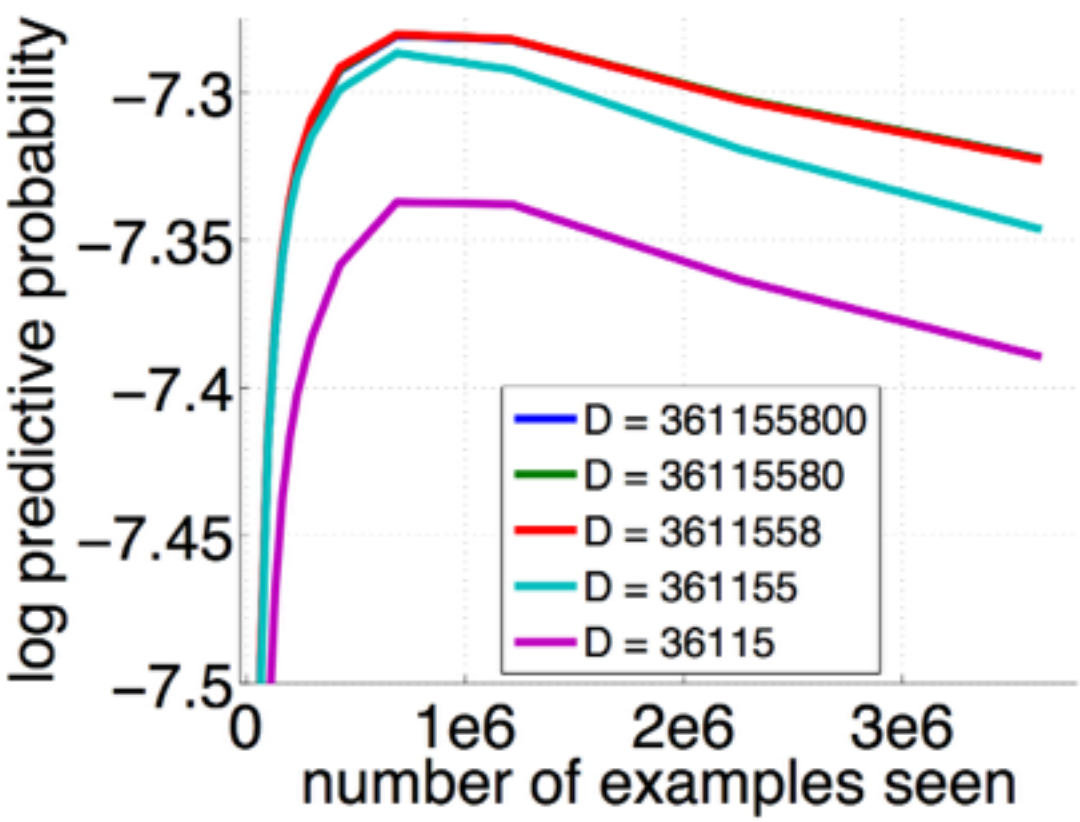
(c) Wikipedia



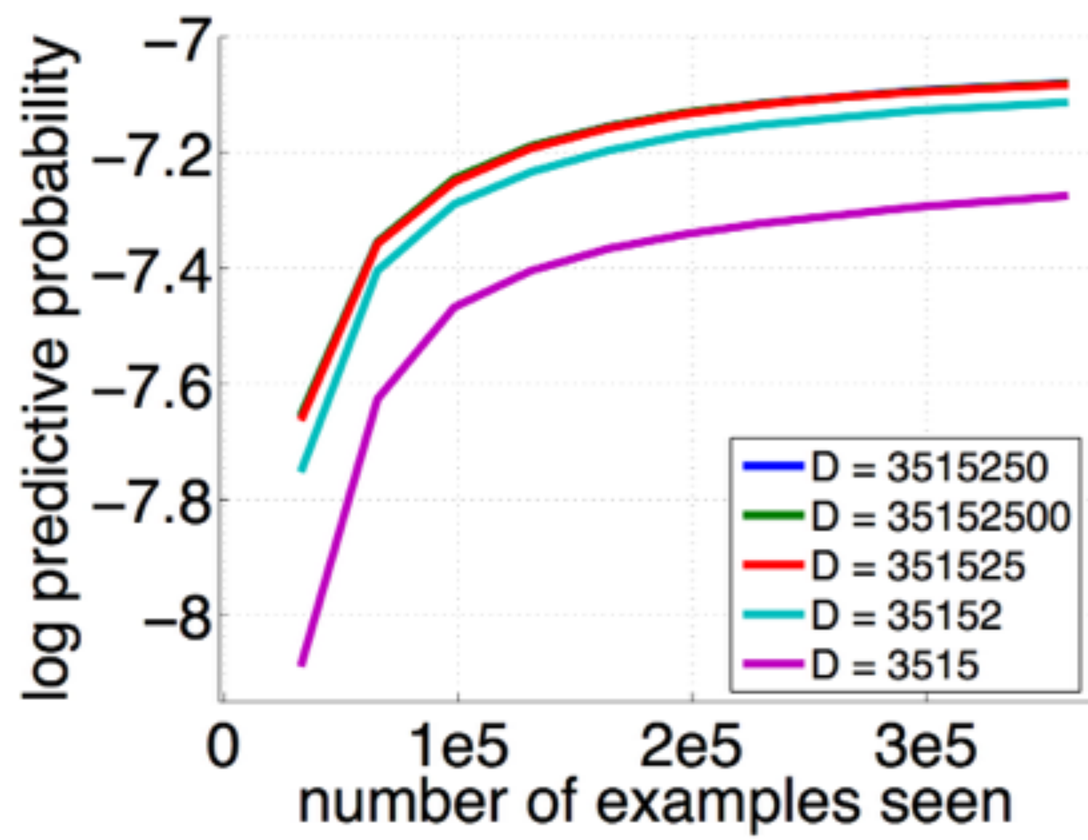
(d) Nature

Results

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



(a) Wikipedia

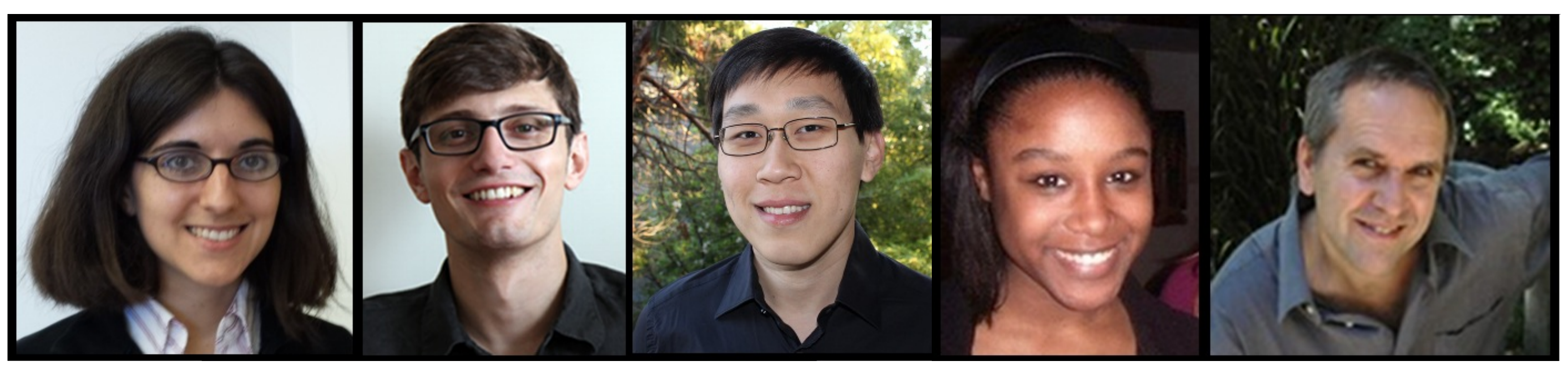


(b) Nature

Further information

- Streaming, distributed Bayesian learning without performance loss

- Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. Streaming variational Bayes. *NIPS* 2013



- Code and slides at www.tamarabroderick.com