Extreme Stochastic Variational Inference (ESVI)

Parameswaran Raman

November 13, 2017

Joint work with:
Jiong Zhang (UT Austin), Hsiang-Fu Yu (UT Austin), Shihao Ji (Intel), S.V.N Vishwanathan (UCSC), Inderjit Dhillon (UT Austin)

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Recap - Exponential Family

Broad umbrella of distributions that can be expressed in the form,

$$p(x;\theta) = p_0(x)\exp(\langle\phi(x),\theta\rangle - g(\theta))$$

- $p_0(x)$ - base measure
- $\phi(x)$ - sufficient statistics
- θ - natural parameter
- $g(\theta) = \log\int_x \exp(\langle\phi(x),\theta\rangle)dx$ - log-partition function

Examples: Gaussian, Multinomial, Exponential, Dirichlet, Poisson, Gamma, ...

Quick Recap - Exponential Family

Key Properties:

- $g(\theta)$ is convex
- Derivatives of $g(\theta)$ generate moments of $\phi(x)$
  - $\partial_\theta g(\theta) = \mathbb{E}_{p(x;\theta)}[\phi(x)]$
  - $\partial^2_\theta g(\theta) = \mathrm{Var}_{p(x;\theta)}[\phi(x)]$
- Every exponential family distribution has a conjugate prior

$$p(\theta|n,\nu) = \exp(\langle n \cdot \nu, \theta \rangle - n \cdot g(\theta) - h(n,\nu))$$

where,

- new sufficient statistics are $(\theta, -g(\theta))$
- new natural parameter is $(n \cdot \nu, n)$
- new log-partition function is $h(n,\nu)$

Bayesian Inference

Given data $x = \{x_1, x_2, \ldots, x_n\}$,

$$\underbrace{p(\theta|x)}_{\text{posterior}} = \frac{\overbrace{p(x|\theta)}^{\text{likelihood}} \cdot \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int p(x,\theta)d\theta}_{\text{marginal likelihood (model evidence)}}}$$

Modeling Assumptions:

- $p(x|\theta) \sim \exp(\langle \phi(x), \theta \rangle) - g(\theta))$
- x are iid

Most inference problems will be one of:

- **Marginalization** $p(x) = \int p(x,\theta)d\theta$
- **Expectation** $\mathbb{E}[f(x|z)] = \int f(x)p(x|z)dz$
- **Prediction** $p(y|x) = \int p(y|\theta,x)p(\theta|x)d\theta$

Computational Challenges

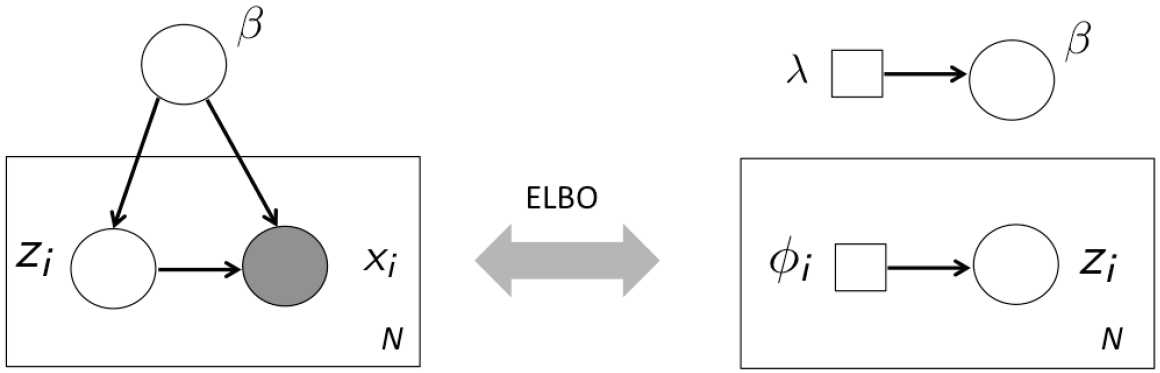
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x, \theta) d\theta}$$

- Computing the log-partition function

Solution: Approximate Inference techniques!

Approaches: Type I (sampling based), Type II (variational approximation based)

Recap - Variational Inference



- We are interested in computing posterior for the following model

$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- Introduce a **fully factored variational distribution** over the latent variables

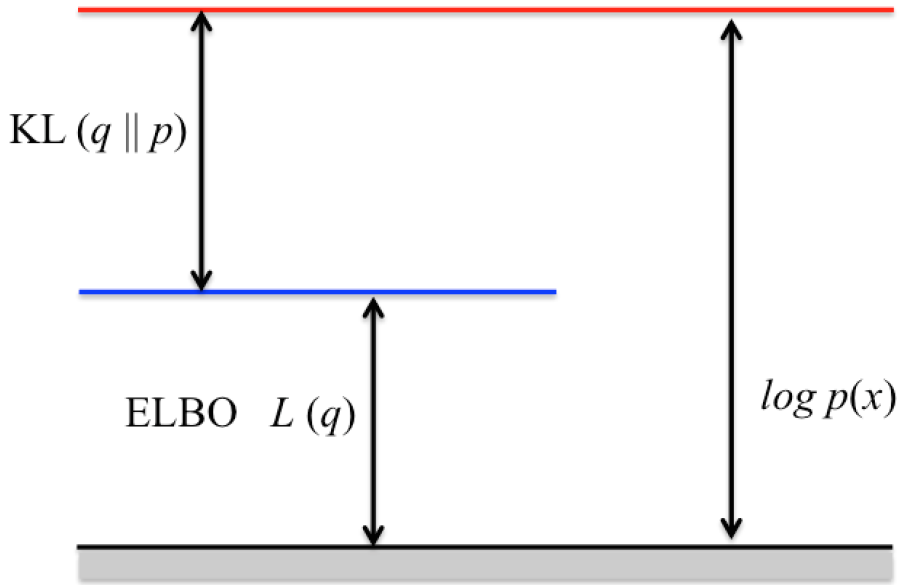
$$q(\beta, z) = q(\beta|\lambda) \prod_{i=1}^{N} q(z_i|\phi_i)$$

Evidence Lower Bound (ELBO)

- Optimize the **evidence lower bound** (ELBO) with respect to q

$$\log p(x) \geq \mathbb{E}_q \Big[\log p(\beta, Z, x) \Big] - \mathbb{E}_q \Big[\log q(\beta, Z) \Big]$$

- Up to a constant, this is the negative KL divergence between q and the posterior



Optimization of ELBO

- Objective function: We optimize the ELBO wrt variational parameters

$$\mathcal{L}(\lambda, \phi_{1:N}) = \mathbb{E}_q[\log p(\beta, x, z)] - \mathbb{E}_q[\log q(\beta, z)]$$

- Same as finding the $q(\beta, z)$ that is closest in KL Divergence to $p(\beta, z|x)$

Algorithm 1 Coordinate Ascent

// **Local Step** (*Var E-Step*):

for $i = 1, \ldots, N$ **do**

  Update ϕ_i

// **Global Step** (*Var M-Step*):

Update λ

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Mixture Models

- k-means
- Gaussian Mixture Models (GMM)
- Latent Dirichlet Allocation (LDA)
- Stochastic Mixed Membership Models

Mixture Models

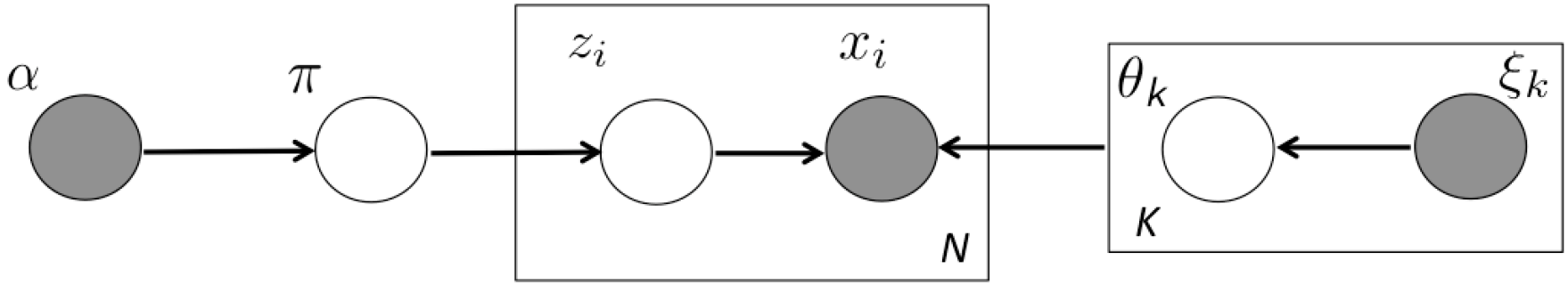


Figure: Plate Diagram of Mixture Models

Mixture Models

Generative process:

$$p(\pi|\alpha) = \text{Dirichlet}(\alpha) \tag{1}$$

For $k = 1, \ldots, K$

$$p(\theta_k|n_k, \nu_k) = \exp\left(\langle n_k \cdot \nu_k, \theta_k \rangle - n_k \cdot g(\theta_k) - h(n_k, \nu_k)\right) \tag{2}$$

for $i = 1, \ldots, N$

$$p(z_i|\pi) = \text{Multinomial}(\pi) \tag{3}$$

$$p(x_i|z_i, \theta) = \exp\left(\langle \phi(x_i, z_i), \theta_{z_i} \rangle - g(\theta_{z_i})\right) \tag{4}$$

Mixture Models

The joint distribution of the data and latent variables can be written as:

$$p\left(x, \pi, z, \theta | \alpha, n, \nu\right) = p\left(\pi | \alpha\right) \cdot \prod_{k=1}^{K} p\left(\theta_k | n_k, \nu_k\right) \cdot \prod_{i=1}^{N} p\left(z_i | \pi\right) \cdot p\left(x_i | z_i, \theta\right) \tag{5}$$

Variational inference approximates this distribution with a fully, factorized distribution of the following form:

$$q\left(\pi, z, \theta | \tilde{\pi}, \tilde{z}, \tilde{\theta}\right) = q\left(\pi | \tilde{\pi}\right) \cdot \prod_{i=1}^{N} q\left(z_i | \tilde{z}_i\right) \cdot \prod_{k=1}^{K} q\left(\theta_k | \tilde{\theta}_k\right). \tag{6}$$

Mixture Models

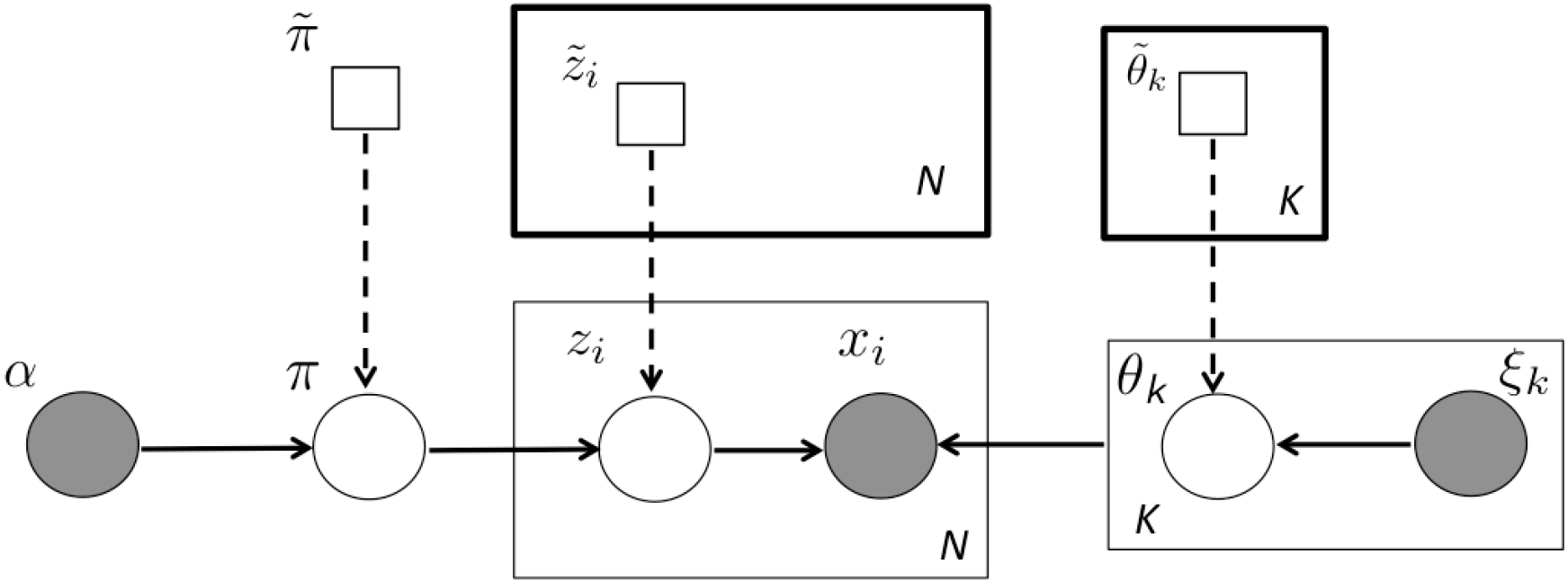


Figure: Plate Diagram illustrating true and variational models

Mixture Models

Update for $\tilde{\pi}$

$$\tilde{\pi}_k = \alpha + \sum_{i=1}^{N} \tilde{z}_{i,k} \tag{7}$$

Update for $\tilde{\theta}_k$ The components of $\tilde{\theta}_k$ namely $\tilde{n}_k$ and $\tilde{\nu}_k$ are updated as follows:

$$\tilde{n}_k = n_k + N_k \tag{8}$$

$$\tilde{\nu}_k = n_k \cdot \nu_k + N_k \cdot \bar{x}_k \tag{9}$$

where $N_k := \sum_{i=1}^{N} \tilde{z}_{i,k}$ and $\bar{x}_k := \frac{1}{N_k} \sum_{i=1}^{N} \tilde{z}_{i,k} \cdot \phi(x_i, k)$.

Mixture Models

Update for $\tilde{z}_i$ Let u_i be a K dimensional vector whose k-th component is given by

$$u_{i,k} = \psi\left(\tilde{\pi}_k\right) - \psi\left(\sum_{k'=1}^{K} \tilde{\pi}_{k'}\right) + \left\langle \phi\left(x_i, k\right), \mathbb{E}_{q\left(\theta_k|\tilde{\theta}_k\right)}\left[\theta_k\right]\right\rangle - \mathbb{E}_{q\left(\theta_k|\tilde{\theta}_k\right)}\left[g\left(\theta_k\right)\right] \tag{10}$$

$$\tilde{z}_{i,k} = \frac{\exp\left(u_{i,k}\right)}{\sum_{k'=1}^{K} \exp\left(u_{i,k'}\right)} \tag{11}$$

Mixture Models

Inference using VI

Algorithm 2 VI

// E-Step

for $i = 1, \ldots, N$ **do**

Update $\tilde{z}_i$ using (11)

// M-Step

for $k = 1, \ldots K$ **do**

Update $\tilde{\pi}_k$ using (7)

Update $\tilde{\theta}_k$ using (8) and (9)

Appendix

VI Access Pattern:

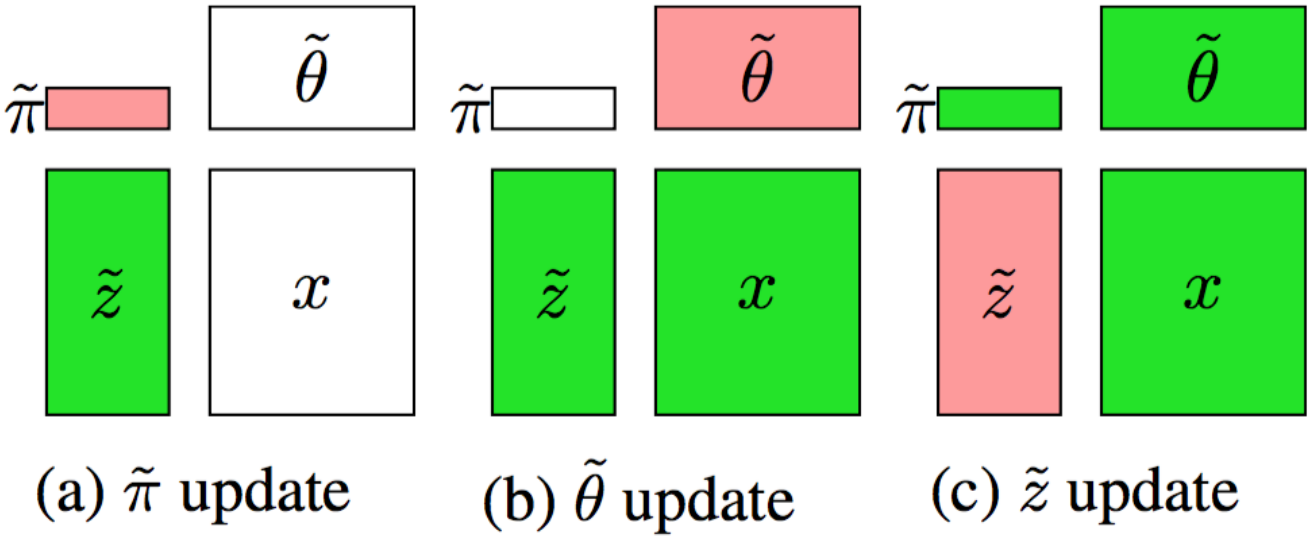


Figure: Access pattern of variables during Variational Inference (VI) updates. Green indicates that the variable or data point is being read, while red indicates that the variable is being updated.

Mixture Models

Inference using SVI

Algorithm 3 SVI

Generate step size sequence $\eta_t \in (0, 1)$

Pick an $i \in \{1, \ldots, N\}$ uniformly at random

Update $\tilde{z}_i$ using (11)

for $k = 1, \ldots K$ **do**

Update $\tilde{\pi}_k$ using (7)

Update $\tilde{\theta}_k$ using (8), (9)

Appendix

SVI Access Pattern:

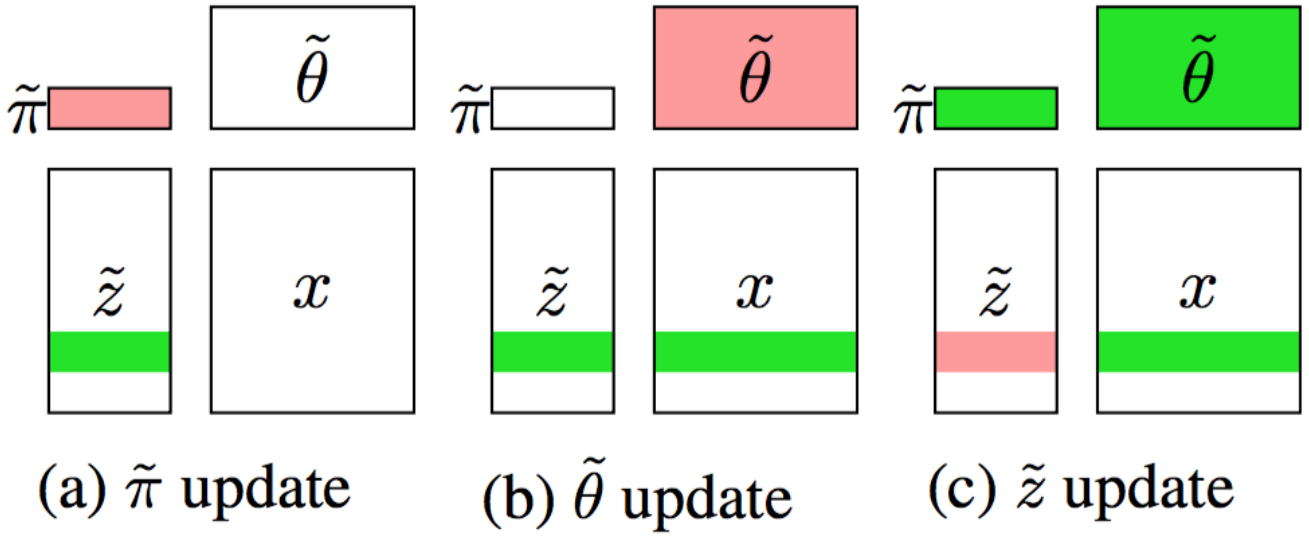


Figure: Access pattern of variables during Stochastic Variational Inference (SVI) updates. Green indicates that the variable or data point is being read, while red indicates that the variable is being updated.

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

ESVI - Key Idea in a nutshell

Instead of *updating all the K coordinates of a local variable* and then updating all K global variables, ESVI [1] proposes the following:

- We only update a *small subset of the local variables* and the *corresponding* global variables
- The global variables nomadically move through the network, and *this ensures mixing*

ESVI

Inference using ESVI

Algorithm 4 ESVI

Sample $i \in \{1, \ldots, N\}$

Select $\mathcal{K} \subset \{1, \ldots, K\}$

Update $\tilde{z}_{i,k}$ for all $k \in \mathcal{K}$ (see below)

Update $\tilde{\pi}_k$ for all $k \in \mathcal{K}$ using (7)

Update $\tilde{\theta}_k$ for all $k \in \mathcal{K}$ using (8), (9)

ESVI - Access Pattern of Data and Parameters

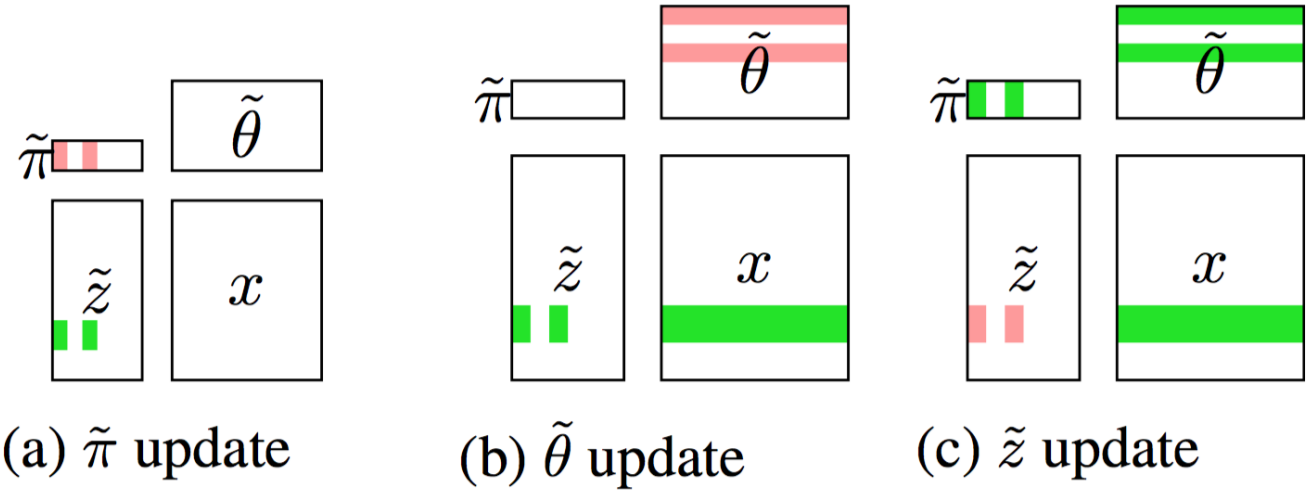


Figure: Access pattern during ESVI updates. Green indicates the variable or data point being read, while Red indicates it being updated.

ESVI - Why picking a subset of coordinates works?

Lemma

For $2 \leq K' \leq K$, *let* $\mathcal{K} \subset \{1, \ldots, K\}$ *be s.t.* $|\mathcal{K}| = K'$. *For any* $C > 0$, *the problem*

$$\max_{z_i \in \mathbb{R}^{K'}} \mathcal{L}_{\mathcal{K}} = \sum_{k \in \mathcal{K}} \tilde{z}_{i,k} \cdot u_{i,k} - \tilde{z}_{i,k} \cdot \log \tilde{z}_{i,k} \tag{12}$$

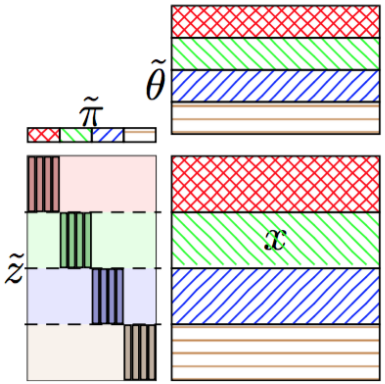
$$s.t. \quad \sum_{k \in \mathcal{K}} \tilde{z}_{i,k} = C \quad and \quad 0 \leq \tilde{z}_{i,k},$$

has the closed form solution:

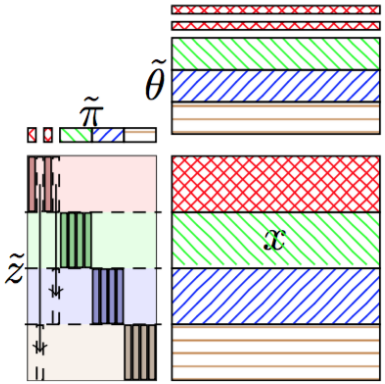
$$\tilde{z}^*_{i,k} = C \frac{\exp\left(u_{i,k}\right)}{\sum_{k' \in \mathcal{K}} \exp\left(u_{i,k'}\right)}, \ \textit{for } k \in \mathcal{K}. \tag{13}$$

Proof: Available in the paper.

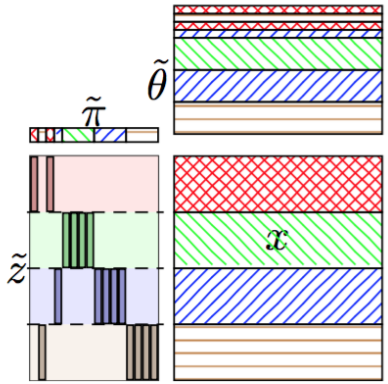
ESVI - Parallelization



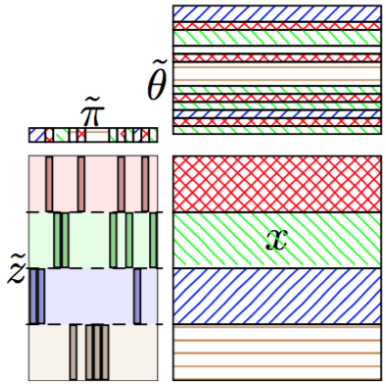
(a) Initial assignment of $\tilde{\theta}$ and x. We plot diagonal initialization while in real case random initialization is used.



(b) Worker 1 finishes processing $\{2, 4\} \in \mathcal{K}_1$, it sends them over to a random worker. Here, $\tilde{\theta}_2$ is sent from worker 1 to 4 and $\tilde{\theta}_4$ from 1 to 3.



(c) Upon receipt, the column is processed by the new worker. Here, worker 4 can now operate on $\tilde{\theta}_2$ and 3 on $\tilde{\theta}_4$



(d) During the execution of the algorithm, the ownership of the global parameters $\tilde{\theta}_k$ changes.

ESVI - Comparison and Complexity

- ESVI updates are stochastic w.r.t. the coordinates, however the update in each coordinate is exact using (11)
- SVI stochastically samples the data and performs inexact or noisy updates and does not guarantee each step to be an ascent step.

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Outline

- Background
- Mixture Models and Inference
- ESVI
- Experiments
- Appendix
- References

Experiments

Datasets:

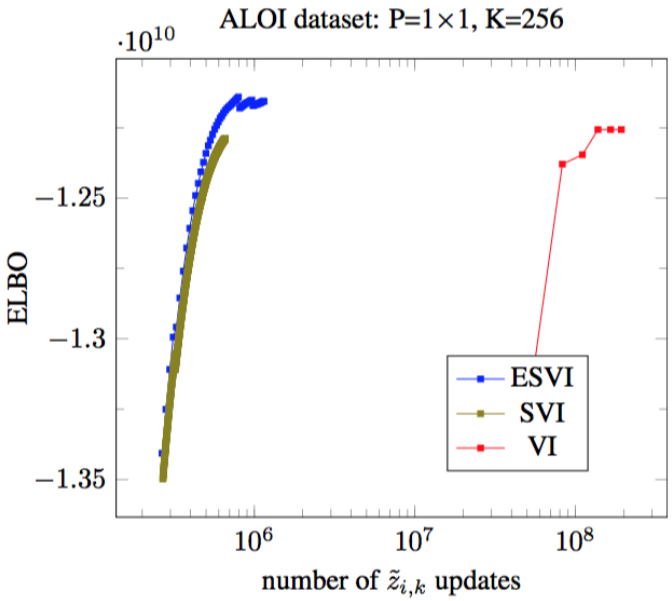
	# documents	# vocabulary	#words
NIPS	1,312	12,149	1,658,309
Enron	37,861	28,102	6,238,796
Ny Times	298,000	102,660	98,793,316
PubMed	8,200,000	141,043	737,869,083
UMBC-3B	40,599,164	3,431,260	3,013,004,127

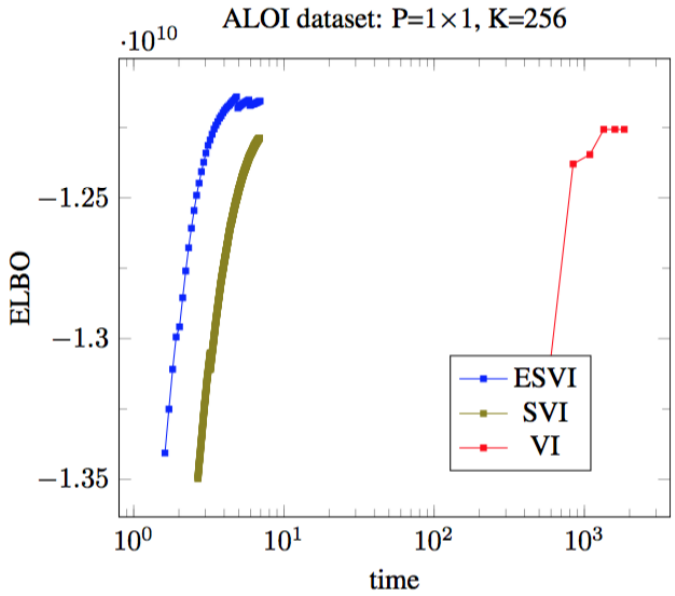
Table: Data Characteristics

We apply ESVI to GMM and LDA in our experiments.

Experiments

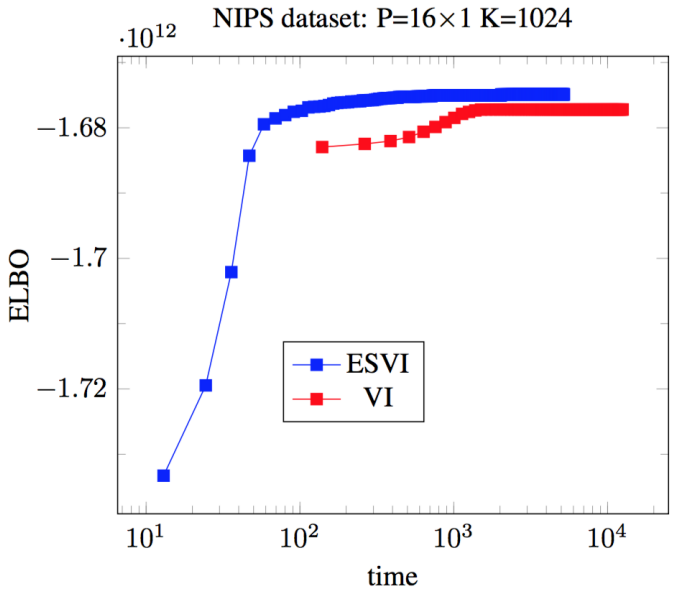
ESVI-GMM (Single-Machine)

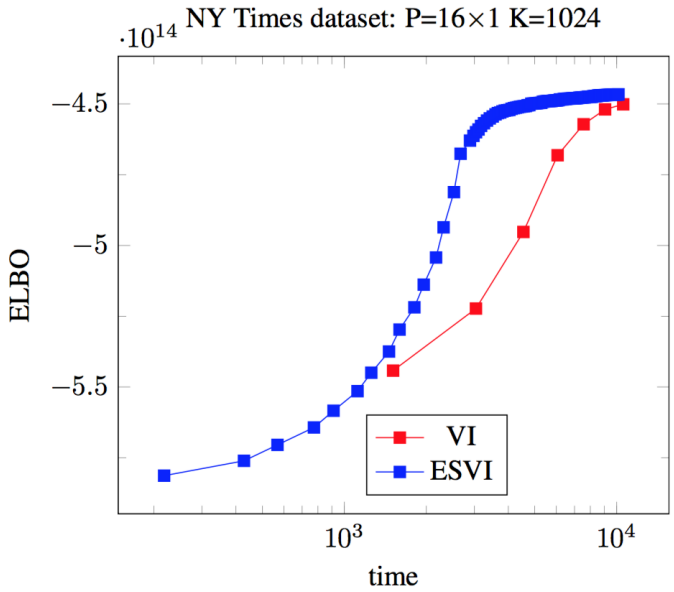




Experiments

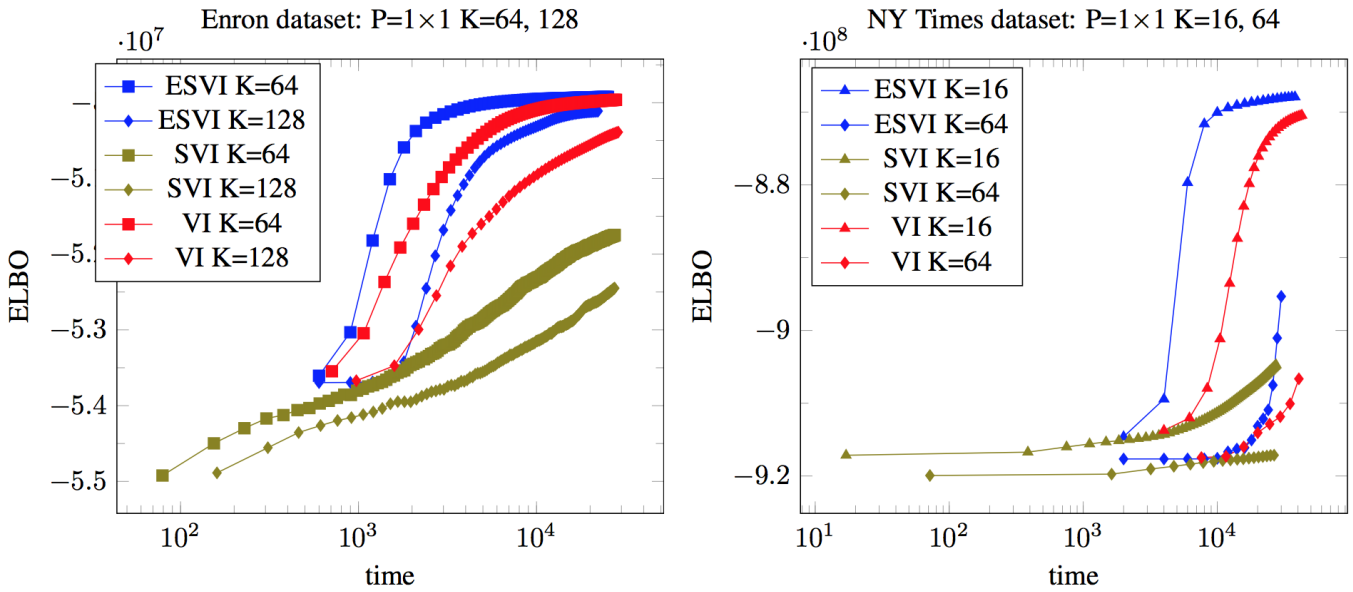
ESVI-GMM (Multi-Machine)





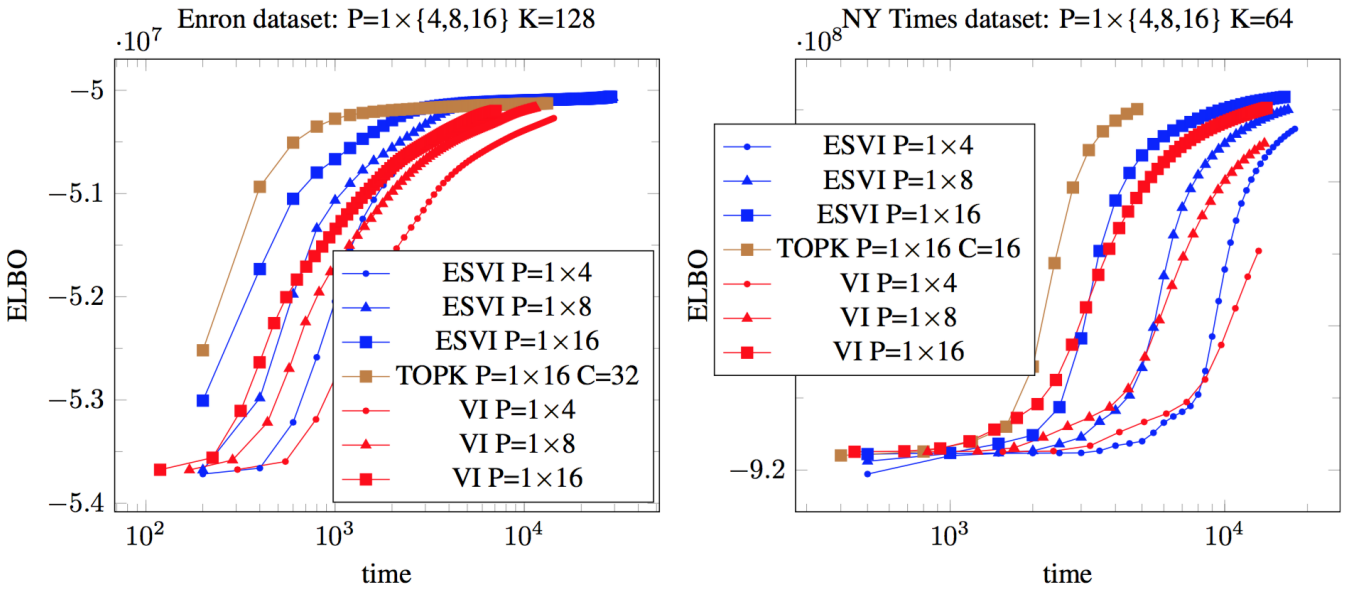
Experiments

ESVI-LDA (Single-Machine) - Varying K, Fixed P



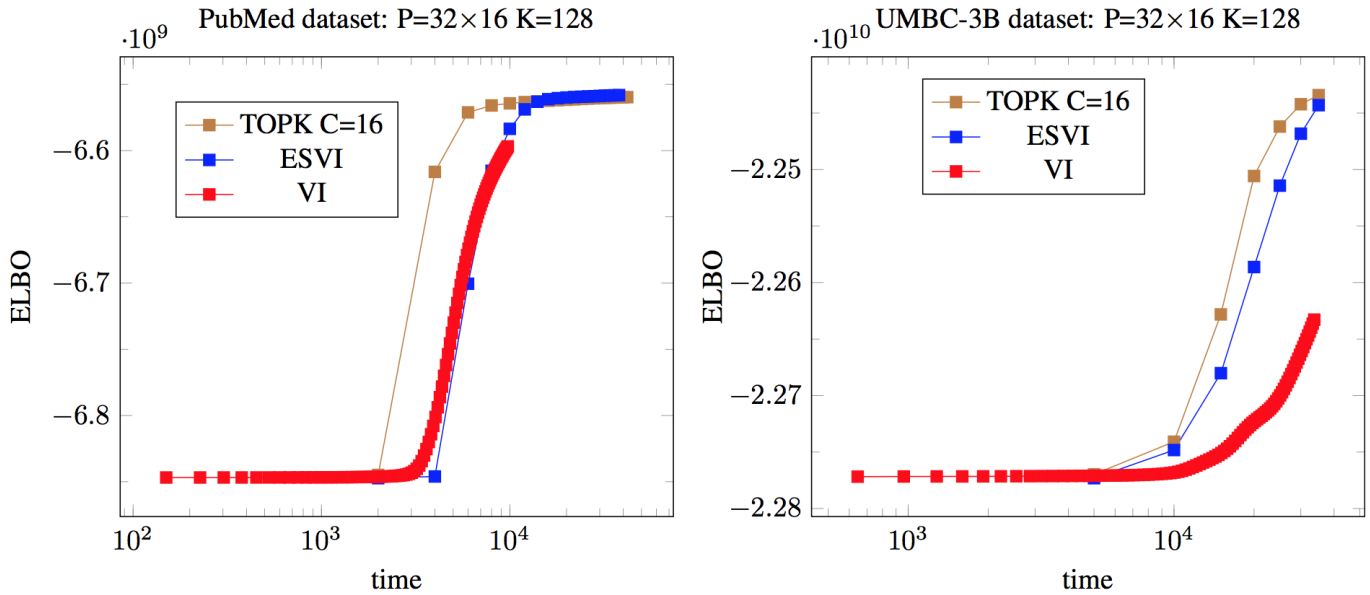
Experiments

ESVI-LDA (Single-Machine) - Varying P, Fixed K



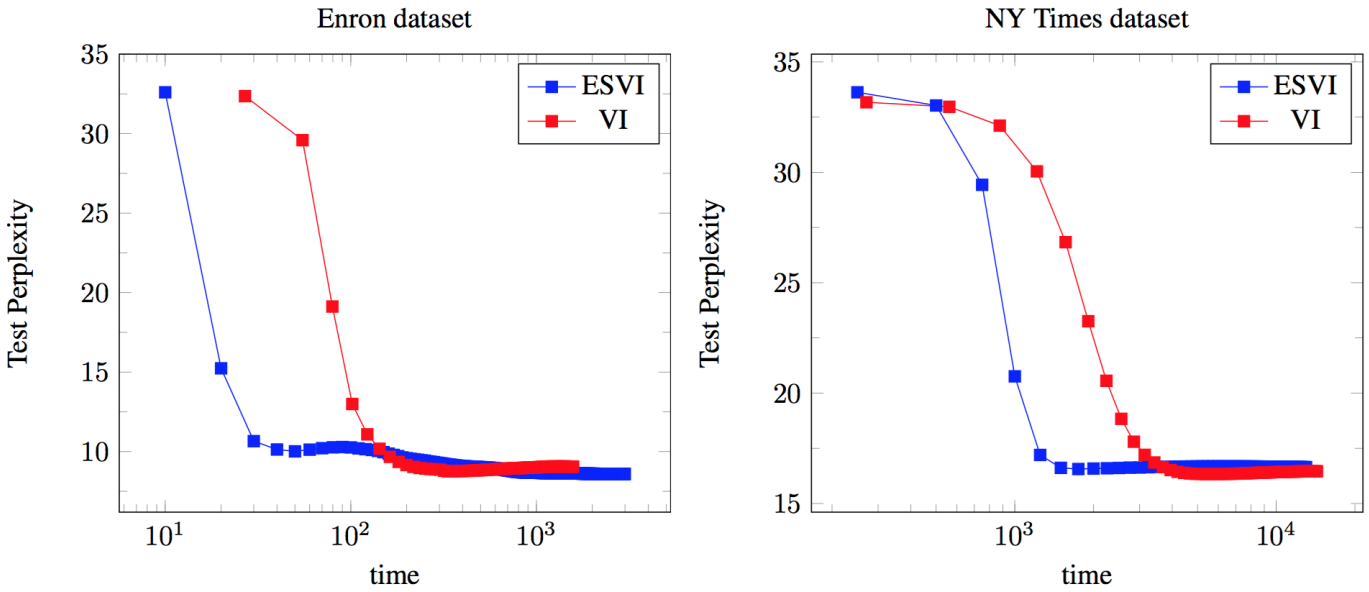
Experiments

ESVI-LDA (Multi-Machine) - PubMed, UMBC



Experiments

ESVI-LDA Predictive Performance



Conclusion

Data \ *Parameters*	Fit	Do not Fit
Fit	VI, SVI, **ESVI**	**ESVI**
Do not Fit	VI, **ESVI**	**ESVI**

Figure: Applicability of the three algorithms to common scenarios in distributed machine learning

- Distributed (both multi-core and multi-machine)
- Asynchronous and Lock-free

References

[1] Parameswaran Raman, Jiong Zhang, Hsiang-Fu Yu, Shihao Ji, and SVN Vishwanathan. Extreme stochastic variational inference: Distributed and asynchronous. *arXiv preprint arXiv:1605.09499*, 2016.

Appendix

ESVI-LDA - Effect of varying TOPK cut-off (C)