Extreme Stochastic Variational Inference: Distributed Inference for Large Scale Mixture Models

Jiong Zhang* Parameswaran Raman* Shihao Ji
zhangjiong724@utexas.edu params@ucsc.edu sji@cs.gsu.edu
UT Austin UC Santa Cruz Georgia State University

Hsiang-Fu Yu †
rofu.yu@gmail.com
Amazon

S.V.N. Vishwanathan
vishy@amazon.com
Amazon

Inderjit S. Dhillon †
derj@cs.utexas.edu
UT Austin & Amazon

Abstract

Mixture of exponential family models are among the most fundamental and widely used statistical models. Stochastic variational inference (SVI), the state-of-the-art algorithm for parameter estimation in such models is inherently serial. Moreover, it requires the parameters to fit in the memory of a single processor; this poses serious limitations on scalability when the number of parameters is in billions. In this paper, we present extreme stochastic variational inference (ESVI), a distributed, asynchronous and lock-free algorithm to perform variational inference for mixture models on massive real world datasets. ESVI overcomes the limitations of SVI by requiring that each processor only access a subset of the data and a subset of the parameters, thus providing data and model parallelism simultaneously. Our empirical study demonstrates that ESVI not only outperforms VI and SVI in wallclock-time, but also achieves a better quality solution. To further speed up computation and save memory when fitting large number of topics, we propose a variant ESVI-TOPK which maintains only the top-k important topics. Empirically, we found that using top 25% topics suffices to achieve the same accuracy as storing all the topics.

† This work was done while the authors were at UT Austin. * denotes equal contribution.

1 Introduction

Mixture of exponential family models generalize a wide collection of popular latent variable models such as Latent Dirichlet Allocation (LDA), Gaussian Mixture Models (GMM), and Mixed Membership Stochastic Block Models (MM-SBM). In recent years, variational inference (VI) has emerged as a powerful technique for parameter estimation in these Bayesian models [14], [3]. One attractive property of VI is that it reduces parameter estimation to the task of optimizing an objective function, often with a well defined “structure”. This opens up the possibility of bringing to bear mature tools from optimization to tackle massive problems. Traditionally, VI in mixture models involves alternating between updating global variables and local variables. Both these operations involve accessing all the data points. Large datasets are usually stored on disk, and the cost of accessing every datapoint to perform updates is prohibitively high.

The first approach to tackle this, is to divide the data across multiple machines and use a distributed framework such as map-reduce to aggregate the computations [11]. The second approach, is to exploit the underlying structure of the optimization problem to reduce the number of iterations (and therefore the corresponding data access) [7]. The key observation here is that the optimization problem corresponding to the local variables is separable, that is, it can be written as a sum of functions, where each function only depends on one data point. Therefore, one can use stochastic optimization to update the local variables. Moreover, in Stochastic Variational Inference (SVI) [7], even before one pass through the dataset, the global variables are updated multiple times, and therefore the model parameters converge rapidly towards their final values. The argument is similar in spirit to how stochastic op-
timization outperforms batch algorithms for maximum 
aposteriori (MAP) estimation [4].

With the advent of the big-data era, we now routinely 
deal with industry-scale problems involving billions of 
documents and tokens. Such massive datasets pose 
another challenge, which, unfortunately neither VI nor 
SVI are able to address; namely, the set of parameters 
is so large that all the parameters do not fit on a single 
processor1. For instance, if we have $D$ dimensional 
data and $K$ mixture components, then the parameter 
size is $O(DK)$. If $D$ is of the order of millions and $K$ 
is in the 100’s or 1000’s, modest numbers by today’s 
standards, the parameter size is a few 100s of GB (see 
our experiments in Section 5).

In this paper, we propose a new framework, Extreme 
Stochastic Variational Inference (ESVI) to address 
these challenges. The main contributions of this paper are:

1. We develop a novel approach to achieve simultaneous 
data and model parallelism in mixture models 
by exploiting the following key idea: instead of updating all the $K$ coordinates of a local variable and then updating all $K$ global variables, we propose updating a small subset of the local variables and the corresponding global variables (see Theorem 1 for proof of correctness). The global variables are exchanged across the processors, and this ensures mixing (see Section 4.2 for more details). This seemingly simple idea has some powerful consequences. It allows multiple processors to simultaneously perform parameter updates independently.

2. Using a classic owner-computes paradigm, we make ESVI asynchronous and lock-free, and thus avoid expensive bulk synchronization between processors. We present an extensive empirical study to evaluate the performance of ESVI by applying it to GMM and LDA models on several large real-world datasets. We find that ESVI outperforms VI and SVI both in terms of time as well as the quality of solutions obtained. For practitioners, ESVI offers the advantage of not having to tune a learning rate, since it makes closed form parameter updates unlike SVI.

3. We develop a variant ESVI-LDA-TOPK to speed up computation and save memory when fitting large number of topics. Empirically, we found that using the most frequent 25% of the topics was enough to obtain the same level of accuracy as storing all the topics.

To the best of our knowledge there is no existing algorithm for VI that sports these desirable properties.

The rest of the paper is structured as follows: We 
present an exhaustive study of related work in Section 2. We briefly review VI and SVI in Section 3. We present our new algorithm ESVI in Section 4, and discuss its advantages. Empirical evaluation is presented in Section 5 and Section 6 concludes the paper.

2 Related Work

Recent research on VI has focused on extending it to non-conjugate models [15] and developing variants that can scale to large datasets such as SVI [7]. Other than the fact that SVI is inherently serial, it also suffers from another drawback: storage of the entire $D \times K$ matrix $\theta$ on a single machine. On the other hand, our method, ESVI, exhibits model parallelism; each processor only needs to store $1/P$ fraction of $\theta$. Black-box variational inference (BBVI) [12] generalizes SVI beyond conditionally conjugate models. The paper proposes a more generic framework by observing that the expectation in the ELBO can be exploited directly to perform stochastic optimization. We view this line of work as complementary to our research. It would be interesting to verify if an ESVI like scheme can also be applied to BBVI.

There has been a flurry of work in the past few years in developing data-parallel distributed methods for Approximate Bayesian Inference. One such popular work includes a classic Map-Reduce style inference algorithm [11], where the data is divided across several worker nodes and each of them perform VI updates in parallel until a final synchronization step during which the parameters from the slaves are combined to produce the final result. This method suffers from the well-known curse of the last reducer, that is, a single slow machine can dramatically slow down the performance. ESVI does not suffer from this problem, because our asynchronous and lock-free updates avoid bulk synchronization altogether.

[5] presents an algorithm that applies VI to the streaming setting by performing asynchronous Bayesian updates to the posterior as batches of data arrive continuously, which is similar in spirit to Hogwild [13]. Their approach uses a parameter server to enable asynchronous local updates. Unlike ESVI, their work cannot guarantee that - (a) each worker works on the latest parameters, (b) the global parameters are all parallely updated. In [1] the authors present Incremental Variational Inference which is also a distributed variational inference algorithm, however it is also only data-parallel. Besides, it requires tuning of a step-size and sequential access of global parameters. ESVI avoids these
drawbacks.

A number of data-parallel approaches exist in the Exact Bayesian Inference literature as well. [6] is a distributed MCMC based approach where workers perform MCMC updates locally and these are aggregated by maintaining a posterior server. [17] proposed a distributed asynchronous algorithm for parameter estimation in LDA [2]. However, the algorithm is specialized to collapsed Gibbs sampling for LDA, and it is unclear how to extend it to other, more general, mixture models. ESVI in contrast is a purely VI based method and provides model-parallelism in addition to data-parallelism.

Somewhat close to our ESVI-TOPK approach is Memorized Online Variational Inference for DP Mixture Models [8]. This paper describes the application of Expectation Truncation to mixture models. In their L-sparse method, unused dimensions are set to zero and used dimensions are shifted. In ESVI-TOPK, unused dimensions are averaged (1-sum of used dimensions).

Another related line of work is Sparse EM [10]. There are some high-level similarities to ESVI in that both the methods update a subset of latent variables at any given time while keeping others frozen. However there are some crucial differences: (a) Sparse EM is not a parallel algorithm while ESVI is, (b) Sparse EM needs to iterate between sparse EM update and full EM update (to select active dimensions occasionally) while each ESVI worker’s job queue will continuously distribute Z’s dimensions to ensure a good mixing, (c) Sparse EM selects active dimensions based on values of Z, while ESVI is designed to ensure the active dimensions of each worker is an unbiased sample of all dimensions.

Since the coordinate-ascent algorithm in VI can be formulated as a message passing scheme applied to general graphical models, we believe ESVI is also related to Variational Message Passing [16]. This connection could be made more concrete if we assume a Mixture Model setup in both cases. Both the d-VMP algorithm (Algorithm 2 in [9]) and ESVI de-couple the global parameters to make the updates scalable, however they differ in some fundamental aspects. d-VMP defines a disjoint partitioning of the global parameters based on their markov-blankets. In contrast, ESVI completely decentralizes the global parameter updates by requiring that the local variables (or assignment vector zi) need to only satisfy local summation constraints (as discussed in Theorem 1 in Section 4). As a side-note, the local updates in d-VMP algorithm do not seem to be de-coupled across the mixture components, whereas this holds true in the case of ESVI.

3 Parameter Estimation for Mixture of Exponential Families

3.1 Generative Model

The following data generation scheme underlies a mixture of exponential family model (Table 1 defines the notations):

Prior:

\[ p(\pi|\alpha) = \text{Dirichlet}(\alpha) \]  

\[ p(\theta_k|n_k, \nu_k) = \exp \left( (n_k \cdot \nu_k, \theta_k) - n_k \cdot g(\theta_k) - h(n_k, \nu_k) \right) \]

where, \( n_k \) and \( \nu_k \) are the parameters of the conjugate prior \( p(\theta_k|n_k, \nu_k) \).

Likelihood:

\[ p(z_i|\pi) = \text{Multinomial}(\pi) \]

\[ p(x_i|z_i, \theta) = \exp \left( \langle \phi(x_i, z_i), \theta_z \rangle - g(\theta_z) \right) \]

where, \( \phi \) denotes the sufficient statistics. Observe that \( p(\theta_k|n_k, \nu_k) \) is conjugate to \( p(x_i|z_i = k, \theta_k) \), while \( p(\pi|\alpha) \) is conjugate to \( p(z_i|\pi) \).

Joint Distribution:

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

3.2 Variational Inference and Stochastic Variational Inference

The goal of inference is to estimate \( p(\pi, z, \theta|x, \alpha, n, \nu) \). This however, involves an intractable marginalization. Therefore, variational inference [3] approximates this distribution with a fully factorized distribution \(^2\) :

\[ q(\pi, z, \theta|\tilde{\pi}, \tilde{z}, \tilde{\theta}) = q(\pi|\tilde{\pi}) \cdot \prod_{i=1}^{N} q(z_i|\tilde{z}_i) \cdot \prod_{k=1}^{K} q(\theta_k|\tilde{\theta}_k). \]

Note that \( \tilde{z}_i \in \Delta_K \) and \( z_{i,k} = q(z_i = k|\tilde{z}_i) \). Moreover, each of the factors in the variational distribution is assumed to belong to the same exponential family as their full conditional counterparts in (5). The variational parameters are estimated by maximizing the

\(^2A \sim \) over a symbol is used to denote that it is a parameter of the variational distribution. See Table 1.
VI performs coordinate ascent updates on $\mathcal{L}$ by optimizing each set of variables, one at a time.

**Update for $\tilde{\pi}_k$:**

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

(8)

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

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

(9)

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

(10)

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

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

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

(11)

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

(12)

It has to be noted that the summation term $\psi \left( \sum_{k'=1}^{K} \tilde{\pi}_{k'} \right)$ cancels out during the $\tilde{z}_{i,k}$ update in (12).

The VI algorithm [14, 3] iteratively updates all the local variables $\tilde{z}_i$ before updating the global variables $\tilde{\pi}_k$ and $\tilde{\theta}_k$ (see Algorithm 1). In contrast, SVI [7], updates $\tilde{z}_i$ corresponding to one data point $x_i$, followed by updating the global parameters $\tilde{\pi}$ and $\tilde{\theta}$ (see Algorithm 2).

Algorithm 1 VI

```plaintext
for i = 1, ..., N do
    Update $\tilde{z}_i$ using (12)
end for

for k = 1, ..., K do
    Update $\tilde{\pi}_k$ using (8)
    Update $\tilde{\theta}_k$ using (9) and (10)
end for
```

Algorithm 2 SVI

- Generate step size sequence $\eta_t \in (0, 1)$
- Pick an $i \in \{1, ..., N\}$ uniformly at random
  - Update $\tilde{z}_i$ using (12)
  - for $k = 1, ..., K$ do
    - Update $\tilde{\pi}_k$ ← $(1 - \eta_t)\tilde{\pi}_k + \eta_t (\alpha + N \cdot \tilde{z}_i, k)$
    - $\tilde{\theta}_k$ ← $\{n_k + N \cdot \tilde{z}_i, k, \nu_k + N \cdot \tilde{z}_{i,k} \cdot \phi(x_i, k)\}$
    - Update $\tilde{\theta}_k$ ← $(1 - \eta_t)\tilde{\theta}_k + \eta_t \tilde{\theta}_k$
  - end for

4 Extreme Stochastic Variational Inference (ESVI)

Algorithm 3 ESVI

- Sample $i \in \{1, ..., N\}$
- Select $K \in \{1, ..., K\}$
  - Update $\tilde{z}_{i,k}$ for all $k \in K$ (see below)
  - Update $\tilde{\pi}_k$ for all $k \in K$ using (8)
  - Update $\tilde{\theta}_k$ for all $k \in K$ using (9), (10)

Algorithm 3 illustrates our proposed updates in ESVI. Before we discuss why this update is advantageous for parallelization, let us first study why updating a subset of coordinates of $\tilde{z}_i$ is justified.
Theorem to a subset of coordinates.

Moreover, to ensure that \( \tilde{z}_{i,k} \) is a valid distribution, one needs to enforce the following constraints:

\[
\sum_{k=1}^{K} \tilde{z}_{i,k} = 1, \quad 0 \leq \tilde{z}_{i,k} \leq 1.
\]

Theorem 1 suggests the following strategy: start with a feasible \( \tilde{z} \), pick, say, a pair of coordinates \( \tilde{z}_{i,k} \) and \( \tilde{z}_{i,k'} \) and let \( \tilde{z}_{i,k} + \tilde{z}_{i,k'} = C \). Solve (16), which has the closed form solution (17). Clearly, if \( \tilde{z} \) satisfied constraints (15) before the update, it will continue to satisfy the constraints even after the update. On the other hand, the conditional ELBO (14) increases as a result of the update. Therefore, ESVI is a valid coordinate ascent algorithm for improving the ELBO (7).

4.1 Access Patterns

In this section we compare the access patterns of variables in the three algorithms to gain a better understanding of their abilities to be parallelized efficiently. In VI, the updates for \( \pi \) and \( \theta \) requires access to all \( z_i \), while update to \( \tilde{z}_i \) requires access to \( \pi \) and all \( \theta_k \). On the other hand, in case of SVI, the access pattern is somewhat different. The updates for \( \pi \) and \( \theta \) require access to only the \( \tilde{z}_i \) that was updated, however the update to \( \tilde{z}_i \) still requires access to \( \pi \) and all the \( \theta_k \). This is a crucial bottleneck to model parallelism. Refer to Figure 9 and Figure 10 in the Appendix for a visual illustration.

In contrast, the following access pattern of ESVI allows multiple processors to access and update mutually exclusive subsets of coordinates \( K \) independently (See Figure 1 for an illustration):

- The update for \( \tilde{z} \) (8) requires access to the coordinates \( \tilde{z}_{i,k} \) for \( k \in K \).
- The update for \( \theta \) (9) and (10) requires access to \( \tilde{z}_{i,k} \) for \( k \in K \).
- The update to \( \tilde{z}_{i,k} \) for \( k \in K \) requires access to \( \pi_k \) and \( \theta_k \) for \( k \in K \).

4.2 Parallelization

In this sub-section, we describe the parallel asynchronous algorithm of ESVI. Let \( P \) denote the number of processors, and let \( I_p \subset \{1, \ldots, N\} \) denote indices of the data points owned by processor \( p \). \( \tilde{z}_i \) for \( i \in I_p \) are local variables assigned to processor \( p \). The global variables are split across the processors. Let \( K_p \subset \{1, \ldots, K\} \) denote the indices of the rows of \( \theta \) currently residing in processor \( p \). Then processor \( p \) can update any \( \tilde{z}_{i,k} \) for \( i \in I_p \) and \( k \in K_p \). Finally, we need to address the issue of how to communicate \( \theta_k \) across processors. For this, we follow the asynchronous communication scheme outlined by [19] and [18]. Figure 2 is an illustration of how this works pictorially. We partition the data and the corresponding \( \tilde{z}_{1:N} \) variables across the processors. Each processors maintains its own queue. Once partitioned, the \( \tilde{z} \) variables never
Algorithm 4  Parallel-ESVI Algorithm

\begin{algorithm}
\begin{algorithmic}
\State $P$: total number of workers,  $T$: maximum computing time
\State $I_p$: data points owned by worker $p$
\State $K_p$: global parameters owned by worker $p$ (concurrent queue)
\State Initialize global parameters $\bar{\theta}^0$, $\bar{\pi}^0$
\For {worker $p = 1 \ldots P$ asynchronously}
\While {Stop criteria not satisfied}
\State Pick a subset $k_s \subset K_p$
\For {All data point $i \in I_p$}
\State Compute $\tilde{z}_{ik}$ using (17)
\State $\tilde{\pi}_k += \tilde{z}_{ik} - \bar{z}_{ik}$
\State $\tilde{n}_k += \bar{z}_{ik} - \tilde{z}_{ik}$
\State $\tilde{v}_k += (\tilde{z}_{ik} - \bar{z}_{ik}) \times \phi(x_i, k)$
\State $\bar{z}_{ik} -= \tilde{z}_{ik}$
\State Pick a random worker $p'$ and send $\tilde{\pi}_k$ and $\tilde{\theta}_k$, push $k$ to $K_{p'}$
\EndFor
\EndFor
\EndWhile
\EndFor
\end{algorithmic}
\end{algorithm}

move. On the other hand, the $\bar{\theta}$ variables move nomadically\footnote{Nomadic movement [19] refers to the distributed setup, where the ownership of parameters rapidly keeps changing after every update. Figure 2 illustrates this.} between processors. Each processor performs ESVI updates using the current subset of $\bar{\theta}$ variables that it currently holds. Then the variables are passed on to the queue of another randomly chosen processor as shown in the second sub-figure in Figure 2. It is this nomadic movement of the $\bar{\theta}$ variables that ensures proper mixing and convergence. The complete algorithm for parallel-ESVI is outlined in Algorithm 4.

4.3 Comparison and Complexity

ESVI updates are stochastic w.r.t. the coordinates, however the update in each coordinate is exact using (17). In contrast, SVI stochastically samples the data and performs inexact or noisy updates and does not guarantee each step to be an ascent step. Moreover, given a $N \times D$ dataset and fixing $K$ clusters, by simple calculation we can see that to update all $\bar{z}_{ik}$ once, VI requires $O(DK)$ updates on $\bar{\theta}$, while SVI needs $O(NDK)$ and parallel ESVI needs $O(PDK)$.

5 Experiments

In our experiments\footnote{Our code and scripts will be made publicly available.}, we compare our proposed ESVI-GMM and ESVI-LDA methods against VI and SVI.

To handle large number of topics in LDA, we also implemented a more efficient version ESVI-LDA-TOPK. We use real-world datasets of varying scale as described in Table 2. We used a large-scale parallel computing platform with node configuration of 20 Intel Xeon E5 - 2680 CPUs and 256 GB memory. We implemented ESVI-LDA in C++ using MPICH, OpenMP and Intel TBB. For Distributed-VI and SVI implementations, we modified the authors' original code in C\textsuperscript{7}. More details on the parameter settings and update equations are available in Appendix C, D, E.

5.1 ESVI-GMM

We first compare ESVI-GMM with SVI and VI in the Single Machine Single thread setting. We use a TOY dataset which consists of $N = 29,983$ data points, $D = 128$ dimensions and the AP-DATA dataset which consists of $N = 2,246$ data points, $D = 10, 473$ dimensions. In both cases, we set the number components $K = 256$. We plot the performance of the methods (ELBO) as a function of time. ESVI-GMM outperforms SVI and VI by quite some margin. For Multi Machine case, we use the NIPS and NY Times datasets and only compare against VI (SVI does not apply; it needs to update all its $K$ global parameters which is infeasible when $K$ is large). Although typically these datasets do not demand running on multiple machines, out intention here is to demonstrate scalability to very large number of components ($K = 1024$) and dimensions, which is typically the case in large scale text datasets with millions of word count features. Traditionally, GMM inference methods have not been able to handle such a scale. Figure 3 clearly indicates that ESVI-GMM is able to outperform VI by a good margin.

5.2 ESVI-LDA

5.2.1 Single Machine Single thread

We compare serial versions of the methods on Enron and NY Times datasets which are medium sized and fit on a single-machine. On both datasets, we run with single machine and single thread. For Enron, we set \#
We stretch the limits of ESVI-LDA method and compare it against distributed VI on large datasets: PubMed and UMBC-3B. UMBC-3B is a massive dataset with 3 billion tokens and a vocabulary of 3 million. We use 32 nodes and 16 cores, and fit $K = 128$ topics. As the results in Figure 5 demonstrate, ESVI-LDA achieves a better solution than distributed VI in all cases. On the largest dataset UMBC-3B, ESVI-LDA is also much faster than VI. In PubMed, VI has a slight initial advantage, however eventually ESVI progresses much faster towards a better ELBO. ESVI-LDA-TOPK is particularly better than the other two methods on both the datasets, especially on PubMed.

5.3 Handling large # of topics in ESVI-LDA

In VI for LDA, the linear dependence of the model size on $K$ prevents scaling to large $K$ due to memory limitations. Our ESVI-LDA-TOPK approach addresses this: instead of storing all $K$ components of the assignment parameter, we only store the most im-

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

![Image](b) Worker 1 finishes processing $\{2, 4\} \in K_1$, it sends them over to a random worker. Here, $\hat{\theta}_2$ is sent from worker 1 to 4 and $\hat{\theta}_4$ from 1 to 3.

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

![Image](d) During the execution of the algorithm, the ownership of the global parameters $\hat{\theta}_k$ changes.

Figure 2: Illustration of the communication pattern in ESVI (asynchronous) algorithm. Parameters of same color are in memory of the same worker. Horizontal and Vertical lines indicate the two directions of partitioning data and parameters. Data $x$ is partitioned horizontally along $N$ and vertically along $D$. Local parameter $\tilde{z}$ is partitioned horizontally along $N$ and vertically along $K$. Global parameters $- \tilde{\pi}$ is partitioned vertically along $K$, and $\hat{\theta}$ is partitioned horizontally along $K$ and vertically along $D$. $\tilde{\pi}$ and $\hat{\theta}$ are nomadically exchanged.

![Image](a) TOY dataset: $P=1 \times 1$, $K=256$

![Image](b) AP-Data dataset: $P=1 \times 1$, $K=256$

![Image](c) NIPS dataset: $P=1 \times 1$, $K=1024$

![Image](d) NY Times dataset: $P=1 \times 1$, $K=1024$

Figure 3: Comparison of ESVI-GMM, SVI and VI. $P = N \times n$ denotes $N$ machines each with $n$ threads.
While the approximation of ELBO must get more accurate as $C \to K$, there might exist a choice of $C \ll K$, which gives a good enough approximation. This will give us a significant boost in speed. On Enron dataset, we varied $C$ as $1, 8, 32, 64, 128$ with true $K = 128$ as our baseline. On NY Times dataset, we varied $C$ as $1, 4, 16, 32, 64$ with the true $K = 64$ as baseline. As we expected, setting the cutoff to a value too low leads to very slow convergence. However, it is interesting to note that at a cut off value of roughly $\frac{K}{2}$ (32 on Enron and 16 on NY Times), we get a good result on par with the baseline. On the larger datasets - PubMed and UMBC-3B, setting $C = 16$ was enough to achieve a similar ELBO as the baseline. (See Figure 7).

5.3.2 Fix $C$ (cutoff for $K$), Scale to large $K$

By tuning $C$, ESVI-LDA-TOPK can be run on large number of topics such as $K = 256$ and $K = 512$ on the largest dataset UMBC-3B (See Figure 8).

6 Conclusion

In this paper, we propose varying $K$ by fixing $C$ Extreme Stochastic Variational Inference (ESVI), a distributed, asynchronous and lock-free algorithm to perform inference for mixture of exponential families. ESVI exhibits simultaneous model and data parallelism, allowing us to handle real-world datasets with large number of documents as well as learn sufficiently large number of parameters. For practitioners, we show how to use ESVI to fit GMM and LDA models on large scale real-world datasets consisting of millions of terms and billions of documents. In our empirical study, ESVI outperforms VI and SVI, and in most cases achieves a better quality solution. ESVI framework is very general and can be extended to several other latent variable models such as Stochastic Block Models.

Acknowledgments: We thank the Texas Advanced Computing Center for infrastructure and support. This research was supported by NSF grants CCF-1564000, IIS-1546452 and IIS-1546459.
References


A Proof of Theorem 1

Proof We prove that \( \tilde{z}_i^* \) is a stationary point by checking the KKT conditions for (16). Let \( h(\tilde{z}_i) = (\sum_{k \in K} \tilde{z}_{i,k}) - C \) and \( g_k(\tilde{z}_i) = -z_{i,k} \). It is clear that \( \tilde{z}_i^* \) satisfies the primal feasibility. Now consider KKT multipliers:

\[
\lambda = \log \frac{C}{\sum_{k' \in K} \exp(u_{i,k'})}, \quad \text{and} \quad \mu_k = 0.
\]

We have

\[
\nabla_k \mathcal{L}_K(\tilde{z}_i^*) = u_{i,k} - \log(\tilde{z}_i^*_{k}) - 1 \\
= u_{i,k} - \left( u_{i,k} + \log \frac{C}{\sum_{k' \in K} \exp(u_{i,k'})} \right) \\
= \log \frac{C}{\sum_{k' \in K} \exp(u_{i,k'})} \\
\lambda \nabla_k h(\tilde{z}_i^*) = \log \frac{C}{\sum_{k' \in K} \exp(u_{i,k'})} \\
\mu_k \nabla_k g_k(\tilde{z}_i^*) = 0.
\]

Then it is easy to verify that \( \nabla_k \mathcal{L}_K(\tilde{z}_i^*) = \lambda = \lambda \nabla_k h(\tilde{z}_i^*) \). Thus, \( \tilde{z}_i^* \) satisfies the stationarity condition:

\[
\nabla \mathcal{L}_K(\tilde{z}_i^*) = \lambda \nabla h(\tilde{z}_i^*) + \sum_{k=1}^{K} \mu_k \nabla g_k(\tilde{z}_i^*).
\]

Due to choice of \( \mu_k = 0 \), complementary slackness and dual feasibility are also satisfied. Thus, \( \tilde{z}_i^* \) is the optimal solution to (16).

B Access Patterns

In this section, we outline the access patterns of VI and SVI in Figure 9 and Figure 10 respectively.

![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.](image)

Figure 9: 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.

**Bottleneck to Model Parallelism**: The local variable \( \tilde{z}_i \) needs to be normalized in order to be maintained on the k-dimensional simplex \( \Delta_k \) after the update (12).

![Access pattern of variables during Stochastic Variational Inference (SVI) updates. Green indicates that the variable is being updated.](image)

Figure 10: 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.

This is the primary bottleneck to model parallelism in both VI and SVI, since this requires access to all \( K \) components. In ESVI, we propose a novel way to overcome this barrier, leading to completely independent local and global variable updates.

C ESVI-LDA

In this section, we show how to apply ESVI to Latent Dirichlet Allocation (LDA). Recall the standard LDA model by Blei et al.[2]. Each topic \( \beta_k, k \in [K] \) is a distribution over the vocabulary with size \( V \) and each document is a combination of \( K \) topics. The generative process is:

- Draw topic weights \( \beta_k \sim \text{Dirichlet}(\eta), k = 1 \ldots K \)
- For every document \( d_i \in \{d_1, d_2 \ldots d_D\} \):
  - Draw \( \theta_i \sim \text{Dirichlet}(\alpha) \)
  - For each word \( n \in [N] \):
    * Draw topic assignment \( z_{in} \sim \text{Multi}(\theta_i) \)
    * Draw word \( w_{in} \sim \text{Multi}(\beta_{z_{in}}) \)

where \( \alpha \in \mathbb{R}^K \) and \( \eta \in \mathbb{R}^V \) are symmetric Dirichlet priors. The inference task for LDA is to characterize the posterior distribution \( p(\beta, \theta, z | w) \). While the posterior is intractable to compute, many methods have been developed to approximate the posterior. Here we use the idea in previous sections to develop extreme stochastic variational inference for LDA.

We denote the assignment of word \( n \) in document \( d_i \) as \( z_{in} \) where \( z_i \in \mathbb{R}^K \). Also \( w_{in} \) denotes the \( n \)-th word in \( i \)-th document. Thus in LDA, the local hidden variables for a word is the word assignment vector \( z_{in} \) and local hidden variable for a document is \( z_i \) and the topic mixture \( \theta_i \). The global hidden variable are the topics \( \beta_k \). Given these, we can formulate the complete
We denote multinomial parameter for who present a scheme for keeping a slowly changing variable is now owned by this machine. After updating with all to store global parameters loads its own chunk of the data, and also has local updated independently. In our setting, each machine that, for each word token, these parameters can be parameters.

In order to update $\phi^k_{in}$ we need only to access $\gamma^k_i$, $\lambda^w_{in}$ and $\pi_k$. And similarly, once $\phi^k_{in}$ is modified, only parameters $\gamma^k_i$, $\lambda^w_{in}$ and $\pi_k$ need to be updated. Following that, for each word token, these parameters can be updated independently. In our setting, each machine loads its own chunk of the data, and also has local model parameters $\gamma$ and $\phi$. Each machine maintains a local job queue that stores global parameters $\lambda$ that is now owned by this machine. After updating with each $\lambda^k_{1..K}$, the machine sends it to another machine while pushing $v$ into the job queue of that machine. This leads to a fully asynchronous and non-locking distributed algorithm.

## D ESVI-GMM

Since the distribution of computation in ESVI-GMM method also works in a similar manner as ESVI-LDA Algorithm 5 (only the local and global updates need to be replaced), in this section, we only present the update rules for the local and global variational parameters for Gaussian Mixture Models (GMM).

### D.1 VI updates for GMM

The generative process for this model assumes that data $x = (x_1, \ldots, x_N)$ is generated by a mixture of $K$ gaussian distributions whose mean and precision are given by $\mu = \{\mu_k\}$ and $\Lambda = \{\Lambda_k\}$. $\pi \in \Delta_k$ denotes the mixing coefficient, where $\Delta_k$ is defined to be the $K$-dimensional simplex. These are the \textit{global variables}. As usual, $z = (z_1, \ldots, z_N)$, $z_i \in \Delta_k$ denotes the latent...
variable to keep track of the component assignments to the data points. These are the local variables.

The conditional distributions for the data $x$ and $z$ (likelihood) can be written as:

$$
p(x|z, \mu, \Lambda) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_i | \mu_k, \Lambda_k^{-1})^{z_{ik}}
$$

$$
p(z|\pi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}}
$$

We now introduce the following conjugate priors to simplify the bayesian inference.

$$
p(\pi) = \text{Dirichlet}(\pi | \alpha_0)
p(\mu, \Lambda) = p(\mu | \Lambda) \cdot p(\Lambda)
= \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, \nu_0)
$$

where, $m_0, \alpha_0, \beta_0, \nu_0, W_0$ are hyper-parameters that can be initialized to some suitable value.

Given this setup, we can express the joint distribution of all our random variables as:

$$
p(x, z, \pi, \mu, \Lambda) = p(x|z, \mu, \Lambda) \cdot p(\mu, \Lambda) \cdot p(z|\pi) \cdot p(\pi)
$$

Clearly, the corresponding posterior distribution $p(z, \pi, \mu, \Lambda|x)$ involves computing expensive high-dimensional integrals and therefore a simpler variational distribution $q$ is used as an approximation:

$$
q(z, \pi, \mu, \Lambda) = q(z) \cdot q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Lambda_k)
= q(z) \cdot q(\pi | \alpha) \cdot \prod_{k=1}^{K} q(\mu_k | m_k, (\beta_k \Lambda_k)^{-1})
\cdot q(\Lambda_k | W_k, \nu_k)
$$

Optimizing the ELBO, leads to the following local and global variable updates.

**Update rules for local variables:**

$$
\rho_{i,k} = \exp \left( \mathbb{E} \left[ \log \pi_k \right] + \frac{1}{2} \mathbb{E} \left[ \log |\Lambda_k| \right] \right)
$$

$$
- \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} \left[ (x_i - \mu_k)^\top \Lambda_k (x_i - \mu_k) \right]
$$

where, the terms $t_1, t_2$ and $t_3$ are given by:

$$
t_1 = \psi(\alpha_k) - \psi \left( \sum_{k=1}^{K} \alpha_k \right)
$$

$$
t_2 = \sum_{j=1}^{D} \psi \left( \frac{\nu_k + 1 - j}{2} \right) + D \log 2 + \log |W_k|
$$

$$
t_3 = D\beta_k^{-1} + \nu_k (x_i - \mu_k)^\top W_k (x_i - \mu_k)
$$

Using these, the local updates can be written as:

$$
\bar{z}_{i,k} = \frac{\rho_{i,k}}{\sum_{k'=1}^{K} \rho_{i,k'}}
$$

**Update rules for global variables:**

For these, first we define some intermediate quantities that are used in the global updates.

$$
N_k = \sum_{i=1}^{N} z_{i,k}
$$

$$
\bar{x}_k = \frac{1}{N_k} \sum_{i=1}^{N} z_{i,k} \cdot x_i
$$

$$
S_k = \frac{1}{N_k} \sum_{i=1}^{N} z_{i,k} (x_i - \bar{x}_k) (x_i - \bar{x}_k)^\top
$$

Using these, the global updates can be written as:

$$
q(\Lambda_k) \sim \mathcal{W}(\Lambda_k | W_k, \nu_k)
$$

$$
q(\mu_k | \Lambda_k) \sim \mathcal{N}(\mu_k | (\beta_k \Lambda_k)^{-1})
$$

$$
q(\pi | \alpha) \sim \text{Dirichlet}(\pi | \alpha)
$$

where the above parameters are given by:

$$
\beta_k = \beta_0 + N_k
$$

$$
m_k = \frac{1}{\bar{\beta}_k} (\beta_0 \cdot m_0 + N_k \cdot \bar{x}_k)
$$

$$
W_k^{-1} = W_0^{-1} + N_k \cdot S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)^\top (\bar{x}_k - m_0) - m_0^\top m_0
$$

$$
\nu_k = \nu_0 + N_k
$$

**D.2 Scaling to large dimensions**

When the dimensions $D$ are large, GMM becomes computationally heavy since it involves the storage and inversion of a large $O(D \times D)$ matrix. To overcome this problem, we make the assumption of diagonal covariance matrix $\Sigma_k = \text{diag} (\sigma_1^2, \ldots, \sigma_D^2)$ for each component $k$, which intuitively means that the dimensions are independent within each mixture component. This lets us run ESVI-GMM on larger datasets.
E Parameter Settings used in the empirical study

We used the following hyper-parameter settings:

- In Gaussian Mixture Models (GMM) experiments, we used $\alpha_0 = 5$, $\beta_0 = 1$, $m_0 = 0$. $\nu_0$ and $W_0$ were turned per dataset based on the best performance. We set $\{\nu_0, W_0\}$ as $\{300000, 0.1\}$ for TOY, $\{300000, 0.1\}$ for AP-DATA, $\{500000, 0.5\}$ for NIPS and $\{500000, 0.5\}$ for NYTIMES datasets.

- For the SVI methods, we used a batch size of 100 (we tried various batch sizes and picked the value we found to provide the best results). The step-size in SVI was decayed following the recommendation in [7], namely, $\eta_t = \frac{\eta_0}{1 + t}$, where $\eta_0$ was carefully tuned and set to 0.1. Here, $t$ denotes the iteration index.