

Towards Big Topic Modeling

Jian-Feng Yan, *Member, IEEE*, Jia Zeng, *Senior Member, IEEE*, Zhi-Qiang Liu, and Yang Gao

Abstract—To solve the big topic modeling problem, we need to reduce both time and space complexities of batch latent Dirichlet allocation (LDA) algorithms. Although parallel LDA algorithms on the multi-processor architecture have low time and space complexities, their communication costs among processors often scale linearly with the vocabulary size and the number of topics, leading to a serious scalability problem. To reduce the communication complexity among processors for a better scalability, we propose a novel communication-efficient parallel topic modeling architecture based on power law, which consumes orders of magnitude less communication time when the number of topics is large. We combine the proposed communication-efficient parallel architecture with the online belief propagation (OBP) algorithm referred to as POBP for big topic modeling tasks. Extensive empirical results confirm that POBP has the following advantages to solve the big topic modeling problem: 1) high accuracy, 2) communication-efficient, 3) fast speed, and 4) constant memory usage when compared with recent state-of-the-art parallel LDA algorithms on the multi-processor architecture.

Index Terms—Big topic modeling, latent Dirichlet allocation, communication complexity, multi-processor architecture, online belief propagation, power law.



1 INTRODUCTION

Probabilistic topic modeling [1], [2] provides a powerful method for data analysis in machine learning and applied statistics. In this paper, we study one of the most successful topic modeling algorithms, latent Dirichlet allocation (LDA) [3], which has been widely used in many fields such as text mining, computer vision and computational biology. Big topic modeling algorithms have attracted intensive research interests because big data have become increasingly common in recent years such as billions of tweets, images and videos on the web.

However, it is still a big challenge to reduce both time and space complexities of traditional batch LDA algorithms such as variational Bayes (VB) [3], collapsed Gibbs sampling (GS) [4], and belief propagation (BP) [5] for big topic modeling tasks. For example, if we use the batch BP [5] to extract 10,000 topics from the PUBMED data set containing 8.2 million documents [6], the memory to store all documents and LDA parameters takes around 36 TB, and the time consumption for 200 iterations is around 3 months on a single processor. Therefore, both time and space costs are unaffordable in many real-world applications. Recent big topic modeling solutions fall into three categories: 1) fast batch LDA algorithms, 2) online LDA algorithms, and 3) parallel LDA algorithms.

Fast batch LDA algorithms observe the fact that the probability mass of the topic distribution is concentrated only on a small set of the topics when the number of topics is very large. This sparseness property facilitates fast Gibbs sampling (FGS) [6] and sparse Gibbs sampling (SGS) [7] algorithms. The basic idea is to sample a topic by checking the topics with high concentrated probability mass first. Generally, FGS and SGS run around 8 ~ 20 times faster than traditional GS [4] when the number of topics is very large. Active belief propagation (ABP) [8] is a sub-linear BP algorithm [5] for topic modeling. At each iteration, it scans only a subset of topics and documents for a fast convergence speed. In practice, ABP is around 10 ~ 20 times faster than SGS or FGS to reach convergence with a higher topic modeling accuracy. Despite of the fast speed on large data sets, anchor word recovery-based topic modeling algorithms [9] scale nonlinearly with the vocabulary size and the number of topics. Although a significant speedup has been achieved, these fast batch LDA algorithms still require a large memory space to store both data and LDA parameters.

Unlike fast batch solutions, online LDA algorithms require only a constant memory space by treating both data and LDA parameters as streams composed of several small mini-batches. After sequentially loading each mini-batch into memory for computation until convergence, we free each mini-batch from memory after one look. In practice, we need to confirm that online algorithms can converge to the local optimum point of LDA's objective function. Within the stochastic optimization framework [10], online variational Bayes (OVb) [11] and online belief propagation (OBP) [12] have been proved to fulfill this goal. Generally, online algorithms are faster than their batch

• Jian-Feng Yan, Jia Zeng and Yang Gao are with the School of Computer Science and Technology, Soochow University, Suzhou 215006, China. Jia Zeng is the corresponding author. E-mail: j.zeng@ieee.org.

• Zhi-Qiang Liu is with the School of Creative Media, City University of Hong Kong, Tat Chee Ave. 83, Kowloon Tong, Hong Kong, P.R. China.

counterparts by a factor of 2 to 5 due to fast local gradient descents. However, online algorithms rarely use the powerful parallel architectures to further scale their performances because of high communication costs or serious race conditions [13], [14].

Parallel LDA algorithms use the widely available parallel architecture to speed up topic modeling process. Currently, there are two types of parallel architectures: multi-processor [15] and multi-core [13], where the difference lies in the way to use memory. In the multi-processor architecture (MPA), all processes have separate memory spaces and communicate to synchronize LDA parameters at the end of each iteration. In the multi-core architecture (MCA), all threads share the same memory space so that race condition is serious. There are three important questions remaining to be addressed in recent parallel LDA algorithms:

- 1) Accuracy: Can parallel LDA algorithms produce the same results as those of batch counterparts on a single processor?
- 2) Communication cost: How to reduce the communication cost in MPA?
- 3) Race condition: How to alleviate the race condition in MCA?

Almost all parallel GS (PGS) algorithms [6], [13], [14], [15], [16], [17], [18] can yield only an approximate result with that of batch GS [4], while the parallel VB (PVB) [19] is able to produce exactly the same result with that of batch VB [3]. To alleviate race conditions on the GPU MCA, a streaming approach is proposed to partition data into several non-conflict data streams in memory [13]. But this partition process may introduce the loading imbalance problem for a low parallel efficiency. As far as MPA is concerned, the reduction of communication cost still remains an unsolved problem since the communication cost is often too big to be masked by computation time in web-scale applications. The experimental results confirm that the communication cost may exceed the computation cost to become the primitive cost of big topic modeling [16], [17]. Therefore, in this paper we focus on reducing the communication complexity in MPA for big topic modeling tasks. It is not difficult to combine MPA and MCA for a better parallel architecture to solve big topic modeling problems.

To achieve the communication-efficient goal, we propose a novel MPA based on the power law [20], which has a few orders of magnitude less communication cost when compared with the current state-of-the-art parallel LDA algorithms [6], [14], [15], [19], [21]. Besides, we combine this parallel architecture with the current state-of-the-art online LDA algorithm OBP [12] referred to as POBP for big topic modeling tasks with the following advantages:

- 1) Convergence to the local optimum of the LDA's objective function;
- 2) Communication-efficient;

TABLE 1
Notations.

$1 \leq d \leq D$	Document index
$1 \leq w \leq W$	Word index in vocabulary
$1 \leq k \leq K$	Topic index
$1 \leq m \leq M$	Mini-batch index
$1 \leq n \leq N$	Processor index
$\mathbf{x}_{W \times D} = \{x_{w,d}\}$	Document-word matrix
$\mathbf{z}_{W \times D} = \{z_{w,d}^k\}$	Topic labels for words
$\boldsymbol{\theta}_{K \times D}$	Document-topic distribution
$\boldsymbol{\phi}_{K \times W}$	Topic-word distribution
α, β	Dirichlet hyperparameters

- 3) Fast speed;
- 4) Constant memory usage.

In experiments, our POBP runs 5 ~ 100 times faster, uses constant memory space, consumes around 5% ~ 20% communication time, but achieves 20% ~ 65% higher topic modeling accuracy than current state-of-the-art parallel LDA algorithms. Therefore, we anticipate that the proposed communication-efficient MPA scheme can be generalized to other parallel machine learning algorithms.

The remainder of this paper is organized as follows. Section 2 reviews 1) online belief propagation (OBP) algorithm [12] and 2) the current MPA scheme [15] for big topic modeling. Section 3 presents our solution POBP and introduces how to use power law to significantly reduce the communication complexity in MPA. Section 4 compares the proposed POBP with several state-of-the-art parallel LDA algorithms. Finally, Section 5 makes conclusions and envisions further work.

2 RELATED WORK

We briefly review OBP [12] and MPA [15] for big topic modeling. We show that a simple combination of OBP and MPA will cause unaffordable communication costs for a bad scalability performance. Table 1 summarizes important notations in this paper.

LDA allocates a set of thematic topic labels, $\mathbf{z} = \{z_{w,d}^k\}$, to explain non-zero elements in the document-word co-occurrence matrix $\mathbf{x}_{W \times D} = \{x_{w,d}\}$, where $1 \leq w \leq W$ denotes the word index in the vocabulary, $1 \leq d \leq D$ denotes the document index in the corpus, and $1 \leq k \leq K$ denotes the topic index. Usually, the number of topics K is provided by users. The nonzero element $x_{w,d} \neq 0$ denotes the number of word counts at the index $\{w, d\}$. For each word token $x_{w,d,i} = \{0, 1\}$, $x_{w,d} = \sum_i x_{w,d,i}$, there is a topic label $z_{w,d,i}^k = \{0, 1\}$, $\sum_{k=1}^K z_{w,d,i}^k = 1$, $1 \leq i \leq x_{w,d}$. We define the soft topic label for the word index $\{w, d\}$ by $z_{w,d}^k = \sum_{i=1}^{x_{w,d}} z_{w,d,i}^k x_{w,d,i} / x_{w,d}$, which is an average topic labeling configuration over all word tokens at index $\{w, d\}$. The objective of LDA is to maximize the joint probability $p(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta)$, where $\boldsymbol{\theta}_{K \times D}$ and $\boldsymbol{\phi}_{K \times W}$ are two non-negative matrices of multinomial parameters for document-topic and topic-word distributions, satisfying $\sum_k \theta_d(k) = 1$ and $\sum_w \phi_w(k) =$

1. Both multinomial matrices are generated by two Dirichlet distributions with hyperparameters α and β . For simplicity, we consider the smoothed LDA with fixed symmetric hyperparameters [4].

2.1 OBP

Online belief propagation (OBP) [12] combines active belief propagation (ABP) [22] with stochastic gradient descent framework [10]. It partitions the document-word matrix $\mathbf{x}_{W \times D}$ into mini-batches $x_{w,d}^m, 1 \leq d \leq D_m, 1 \leq m \leq M$. After loading the m th mini-batch into memory, OBP infers the posterior probability called message $\sum_k \mu_{w,d}^m(k) = 1, \mu_{w,d}^m(k) = p(z_{w,d,i}^{k,m} = 1 | x_{w,d,i}^m = 1, \theta, \phi; \alpha, \beta)$,

$$\mu_{w,d}^m(k) \propto \frac{[\hat{\theta}_{-w,d}^m(k) + \alpha] \times [\hat{\phi}_{w,-d}^m(k) + \beta]}{\hat{\phi}_{-(w,d)}^m(k) + W\beta}, \quad (1)$$

where $\hat{\theta}$ and $\hat{\phi}$ are the *sufficient statistics* for the online LDA model,

$$\hat{\theta}_{-w,d}^m(k) = \sum_{-w} x_{w,d}^m \mu_{w,d}^m(k), \quad (2)$$

$$\hat{\phi}_{w,-d}^m(k) = \hat{\phi}_w^{m-1}(k) + \sum_{-d} x_{w,d}^m \mu_{w,d}^m(k), \quad (3)$$

where $-w$ and $-d$ denote all word indices except w and all document indices except d , and $-(w,d)$ denotes all indices except $\{w,d\}$. The multinomial parameters of document-topic and topic-word distributions θ and ϕ can be obtained by normalizing sufficient statistics $\hat{\theta}$ and $\hat{\phi}$. Each mini-batch is swept for several iterations T_m until the convergence condition is reached. Then, OBP frees from memory the m th mini-batch, the local $\mu_{w,d}^m(k)$ and $\hat{\theta}_{-w,d}^m(k)$. The global topic-word distribution $\hat{\phi}_w^m(k)$ in memory will be re-used by the next mini-batch. When the size of $\hat{\phi}_w^m(k)$ is very large, we may also store the entire matrix in hard disk and load the partial matrix in memory for computation [12].

OBP is an ideal choice for big stream topic modeling on the single-processor platform because of several advantages. First, OBP guarantees convergence to the stationary point of LDA's likelihood function within the online expectation-maximization (EM) framework [23], [24], [25]. Second, OBP is memory-efficient by using disk as the storage extension. Its space complexity in memory is proportional to the mini-batch size D_m and the number of topics K . Finally, OBP is built upon time-efficient ABP algorithm [22], whose time complexity is insensitive to the number of topics K and the number of documents in each mini-batch D_m . However, the communication complexity is intractable if we directly parallelize OBP in MPA for big topic modeling tasks, which will be explained in detail in the next subsection.

2.2 MPA

The MPA scheme has been widely used in many parallel batch LDA algorithms [6], [15], [16], [17], [19], [21]. Here, we extend this scheme to parallelize online LDA algorithms. The MPA [15] distributes each mini-batch $\mathbf{x}_{W \times D_m}$ documents over $1 \leq n \leq N$ processors. The processor n gets approximately $D_{m,n} = D_m/N$ documents. The local $\hat{\theta}_{K \times D_{m,n}}^{m,n}$ can be also distributed into N processors, but the global $\hat{\phi}_{K \times W}^{m,n}$ have to be shared by N processors since each distributed mini-batch $\mathbf{x}_{W \times D_{m,n}}$ may still cover the entire vocabulary words. After sweeping each mini-batch $\mathbf{x}_{W \times D_{m,n}}$ at the end of each iteration $1 \leq t \leq T_m$, the N processors have to communicate and synchronize the global matrix $\hat{\phi}_{K \times W}^{m,t}$ from N local matrices $\hat{\phi}_{K \times W}^{m,n,t}$ by

$$\hat{\phi}_w^{m,t}(k) = \hat{\phi}_w^{m,t-1}(k) + \sum_{n=1}^N [\hat{\phi}_w^{m,n,t}(k) - \hat{\phi}_w^{m,t-1}(k)]. \quad (4)$$

Then, the synchronized matrix $\hat{\phi}_w^{m,t}(k)$ is transferred to each processor to replace $\hat{\phi}_w^{m,n,t}(k)$ for the next mini-batch. Thus, the communication complexity is

$$\text{Communication complexity} \propto NMTKW, \quad (5)$$

where N is the number of processors, M the number of mini-batches, K the number of topics, W the vocabulary size, and $T = \sum_{m=1}^M T_m/M$ the average number of iterations to reach convergence for each mini-batch. For example, suppose that we use 1000 processors to learn $K = 2000$ topics with $T = 100$ from the PUBMED data set [6] having $W = 141,043$ and $M = 500$ mini-batches. The total communication cost reaches around 100 PB (10^{15} bytes) according to (5). Meanwhile, the time complexity of OBP reduces linearly with the number of processors N . So, the communication cost will be greater than the computation cost when $N \rightarrow \infty$. In this situation, adding more processors will not reduce the entire topic modeling time, leading to serious scalability issues. The major reason why MPA still works in previous parallel batch LDA algorithms [6], [15], [16], [17], [21] is that the communication cost depends only on the number of batch iterations T' rather than the number of iterations over mini-batches MT , where practically $T' \ll MT$. If $T' = 500$, the parallel batch LDA algorithms require only 1PB communication cost in the above example, which is significantly smaller than that of parallel online LDA algorithms. For some real-world big data streams, the number of mini-batches may reach infinity [26], i.e., $M \rightarrow \infty$. Thus, communication cost of parallel online LDA algorithms may become so huge as to seriously damage parallel efficiency.

Fig. 1 compares the communication costs between parallel batch and online LDA algorithms. Parallel batch LDA algorithms communicate and synchronize $\hat{\phi}_{K \times W}$ at the end of each batch iteration, while

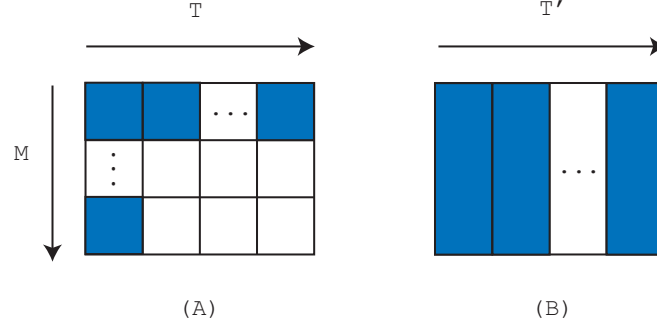


Fig. 1. A comparison of communication costs between parallel (A) online and (B) batch LDA algorithms. Each blue box denotes a communication operation. In (A), the communication rate depends on the number of iterations over all mini-batches MT , while in (B), the communication rate depends on the number of iterations T' .

parallel online algorithms do it at the end of each mini-batch iteration. Generally, the number of batch iterations T' is significantly smaller than the number of mini-batch iterations MT . Thus, the higher communication rate leads to the larger communication cost in parallel online LDA algorithms. Therefore, it is nontrivial to reduce the communication complexity (5) for parallel online LDA algorithms [11], [12], [21], [27], [28] in order to achieve a better scalability performance. Moreover, not all parallel batch LDA algorithms based on MPA have been proved to converge to the local optimum of the LDA's objective function. Typical examples include those GS-based parallel algorithms [6], [15], [16], [17], [21] in MPA framework.

3 POBP

In this paper, we propose a communication-efficient MPA and explain this scheme using power law [20]. Combining with OBP, we propose the parallel OBP (POBP) to solve the big topic modeling problem. We show that POBP has low time, space and communication complexities, and can converge to the local optimum of the LDA's objective function within the online EM framework [23], [24], [25].

3.1 Communication-Efficient MPA

From (5), there are two straight-forward solutions to reduce the communication cost. The first is to reduce the average communication rate T . For example, we may communicate and synchronize the global matrix at every two mini-batch iterations to reduce around half communication cost. This heuristic solution has been widely used in MPA [15] but with two problems: 1) the lower communication rate may cause the lower topic modeling accuracy; and 2) the overall communication rate depends also on the number of mini-batches M , which is often constrained by each processor's memory space. Therefore, we investigate the second solution to communicate and synchronize only the subset of global matrix at each mini-batch

iteration, i.e., reduce the size KW in (5). To our best knowledge, there are very few investigations in related work following this research line. We will further explain why selecting the subset of global matrix dynamically does not influence the topic modeling accuracy very much based on power law.

We propose a two-step strategy to select the subset of global matrix at each iteration in a dynamic manner. First, we select a subset of vocabulary words with size $\lambda_W W$ referred to as the *power words*. For each power word, we select a subset of topics with size $\lambda_K K$ referred to as the *power topics*. In this way, we reduce the communication complexity (5) from KW to $\lambda_K \lambda_W KW$ as follows,

$$\text{Communication complexity} \propto \lambda_K \lambda_W NMTKW, \quad (6)$$

where the ratios $0 < \lambda_K \ll 1$ and $0 < \lambda_W \ll 1$. Obviously, Eq. (6) shows a sublinear complexity of (5). The remaining question is how to select both power words and topics.

Our selection criterion is inspired by the residual belief propagation (RBP) [29], [30]. At each processor n , we define the residual between message vectors (1) at two successive iterations t and $t-1$,

$$r_{w,d}^{m,n,t}(k) = x_{w,d}^{m,n} |\mu_{w,d}^{m,n,t}(k) - \mu_{w,d}^{m,n,t-1}(k)|, \quad (7)$$

$$r_w^{m,n,t}(k) = \sum_d r_{w,d}^{m,n,t}(k). \quad (8)$$

We then communicate and synchronize the residual matrix $r_w^{m,n,t}(k)$ across N processors similar to (4),

$$r_w^{m,\cdot,t}(k) = r_w^{m,\cdot,t-1}(k) + \sum_{n=1}^N [r_w^{m,n,t}(k) - r_w^{m,\cdot,t-1}(k)]. \quad (9)$$

From (9), we further obtain the synchronized residual vector of vocabulary words,

$$r_w^{m,\cdot,t} = \sum_k r_w^{m,\cdot,t}(k). \quad (10)$$

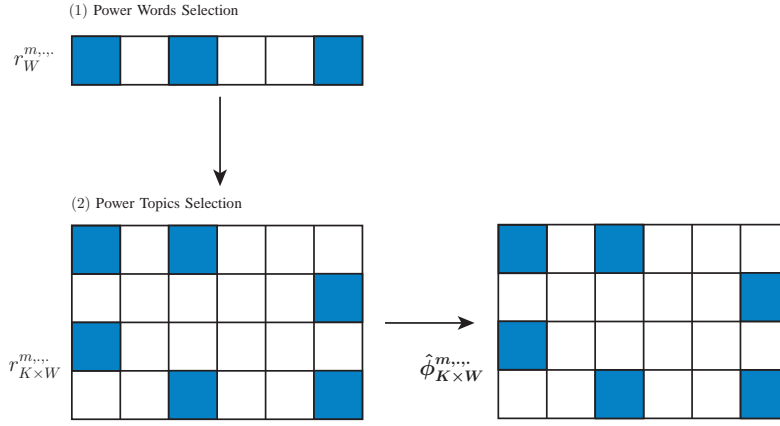


Fig. 2. The two-step power words and topics selection process for a global matrix $\hat{\phi}_{K \times W}$ with $K = 4$ and $W = 6$, where $\lambda_K = \lambda_W = 0.5$. The blue boxes denote the selected power words and topics. In the first step, we select power words by sorting the synchronized residual vector $r_w^{m,t}$. In the second step, for each selected power word we select further power topics by sorting the synchronized residual matrix $r_w^{m,t}(k)$ in K dimensions.

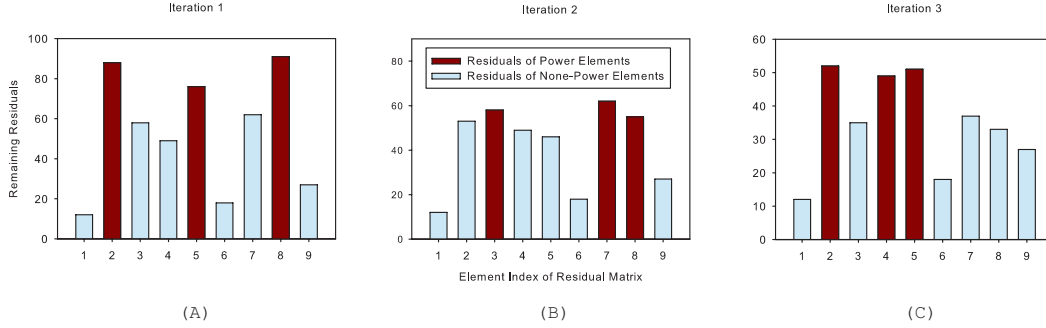


Fig. 3. A dynamic scheduling example of a residual matrix $r_{3 \times 3}$ with 3 words and 3 topics, where the 9 elements in $r_{3 \times 3}$ are shown in one dimension. (A) In the first iteration $t = 1$, the elements $\{2, 5, 8\}$ are chosen as power elements. (B) In the second iteration $t = 2$, the elements $\{3, 7, 8\}$ are chosen as power elements because residuals of $\{2, 5\}$ become relatively smaller. (C) In the third iteration, the elements $\{2, 4, 5\}$ are selected as the power elements because residuals of $\{3, 7, 8\}$ become relatively smaller.

Finally, we sort vector (10) in the descending order, and select the power words with $\lambda_W W$ largest residuals. For each power word, we sort matrix (9) in the K dimension, and select $\lambda_K K$ power topics for each word with largest residuals.

Fig. 2 shows an example of the two-step selection method for the global matrix $\hat{\phi}_{4 \times 6}$. We set the selection ratios as $\lambda_K = \lambda_W = 0.5$. In the first step, we select three power words with largest residuals in the vector $r_w^{m,t}$ denoted by the blue boxes. In the second step, for each selected power word, we select two power topics with largest residuals in the matrix $r_w^{m,t}(k)$ in the K dimension.

This two-step selection process follows the dynamical scheduling scheme. For m th mini-batch at the first iteration $t = 1$, we need to communicate and synchronize the entire matrices $\hat{\phi}_{K \times W}^{m,1}$ and $r_{K \times W}^{m,1}$. When $2 \leq t \leq T_m$, we communicate and synchronize only the partial matrices $\hat{\phi}_{\lambda_K K \times \lambda_W W}^{m,2 \leq t \leq T_m}$ and $r_{\lambda_K K \times \lambda_W W}^{m,2 \leq t \leq T_m}$, while we keep the remaining elements untouched. Residuals (7) of power words and topics are getting

smaller and smaller in the message passing process according to Eq. (1). Therefore, the power words and topics in the previous iteration may be no longer power ones due to their relatively smaller residuals in the next iteration. In this way, all vocabulary words and topics have the chances to be selected as power ones before convergence. When all elements in residual matrix reach zeros, i.e., $r_w^{m,t}(k) \rightarrow 0$, the message passing process reaches the convergence state.

For a better understanding of the dynamic scheduling process, Fig. 3 shows an example $r_{3 \times 3}^{t=1,2,3}$ at different iterations, where the nine elements are shown in one dimension for simplicity. Fig. 3A shows that in the first iteration, the elements $\{2, 5, 8\}$ are selected as the power elements to pass messages such that the residuals for the three elements decrease while other residuals remain unchanged. Fig. 3B shows that elements $\{3, 7, 8\}$ are selected as the power elements in the second iteration because the elements $\{2, 5\}$ get relatively smaller residuals. However, they could be power elements again in next iterations when their

residuals become relatively higher than those of other elements. Fig. 3C shows that the elements $\{2, 4, 5\}$ are chosen as power elements in the third iteration. Therefore, we can guarantee that no information gets lost since all elements have chance to become power elements to pass messages, which ensures the topic modeling accuracy of the algorithm.

3.2 The POBP Algorithm

Although we focus on developing parallel online belief propagation (POBP) algorithm for big topic modeling tasks in this subsection, the proposed communication-efficient MPA can be applied to both parallel batch and online LDA algorithms. Fig. 4 summarizes the proposed POBP algorithm. We distribute each incoming mini-batch $x_{w,d}^{m,n}$ into N processors in parallel (line 2). At the first iteration $t = 1$, we random initialize and normalize messages $\mu_{w,d}^{m,n,0}$ (line 3), which are used to update sufficient statistics $\hat{\theta}_d^{m,n,0}(k)$ and $\hat{\phi}_w^{m,n,0}(k)$ using Eqs. (2) and (3) (lines 4 and 5). Note that we use the stochastic gradient descent [10], [31] to update (3) in line 5, where the initial $\hat{\phi}^{m=0}$ is set as the zero matrix. Then, we update both messages $\mu_{w,d}^{m,n,1}(k)$ and residuals $r_w^{m,n,1}(k)$ using Eqs. (1) and (7) (lines 6 and 7). The messages are in turn used to update sufficient statistics $\hat{\theta}_d^{m,n,1}(k)$ and $\hat{\phi}_w^{m,n,1}(k)$ (line 8). At the end of the first iteration, all processors communicate and synchronize two global matrices $\hat{\phi}_w^{m,\dots,1}(k)$ and $r_w^{m,\dots,1}(k)$, and transfer the global matrices back to each processor (lines 9 and 10). Using two-step selection method, we select the power words and topics from the global residual matrix (lines 12 and 13). We use the partial sort to find the power words and topics with top largest $\lambda_W W$ and $\lambda_K K$. The computation cost of partial sort algorithm is significantly lower than quick sort since we do not need the complete sorting. Also, we use the parallel implementations of partial sort algorithm to further speed up the selection process. The time complexity of partial sort is at most $W \log W$ and $K \log K$, where W is the vocabulary size and K is the number of topics.

In the following iterations $2 \leq t \leq T$, we update only the subsets of messages $\mu_{w,d}^{m,n,t}(k)$ and $r_w^{m,n,t}(k)$ residuals based on the selected power words and topics (lines 17 and 18), and communicate only the subsets of matrices $\hat{\phi}_w^{m,\dots,t}(k)$ and $r_w^{m,\dots,t}(k)$ (lines 23 and 24). In the dynamical scheduling process, we select the power words and topics based on the synchronized residual matrix $r_w^{m,\dots,t}(k)$ (lines 27 and 28). If the average of the residual matrix is blow a threshold (line 26), we terminate all processors and load the next mini-batch $x_{w,d}^{m+1,n}$ after freeing memory except for the global topic-word matrix $\hat{\phi}_w^{m,\dots,t}(k)$. POBP terminates until all M mini-batches have been processed (line 1). When $M \rightarrow \infty$, POBP can be viewed as a life-long or never-ending topic modeling algorithm. The output

is the global sufficient statistics $\hat{\phi}_{K \times W}$, which can be normalized to obtain the topic-word multinomial parameter matrix $\phi_{K \times W}$. If $N = 1$, POBP reduces to the OBP [12] algorithm on a single processor. If $M = 1$, POBP reduces to the parallel batch BP algorithm on N processors [32].

3.2.1 Convergence Analysis

The objective of LDA is to maximize the joint probability $p(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi} | \alpha, \beta)$ [3], [5], [33]. According to the MAP inference [33], [34], [35], this objective can be achieved by the iterative EM algorithm [12], where the E-step has almost the same message update equation with (1), and the M-step resembles Eqs. (2) and (3). OBP uses the stochastic gradient descent method [10], [31] to update the topic-word matrix,

$$\hat{\phi}_w^m(k) = \hat{\phi}_w^{m-1}(k) + \frac{1}{m-1} \Delta \hat{\phi}_w^m(k), \quad (11)$$

where $\Delta \hat{\phi}_w^m(k) = \sum_d x_{w,d}^m \mu_{w,d}^m(k)$ in Eq. (3) is the gradient generated by the current mini-batch. Eq. (11) has a learning rate $1/(m-1)$ because $\hat{\phi}_w^{m-1}(k)$ accumulates sufficient statistics of previous $m-1$ mini-batches, and $\Delta \hat{\phi}_w^m(k)$ accumulates only sufficient statistics of the current mini-batch. The parameter estimation is invariant to the scaling of sufficient statistics (3). Since this learning rate satisfies two conditions,

$$\sum_{m=2}^{\infty} \frac{1}{m-1} = \infty, \quad (12)$$

$$\sum_{m=2}^{\infty} \frac{1}{(m-1)^2} < \infty, \quad (13)$$

the online stochastic approximation [31] shows that sufficient statistics $\hat{\phi}_w^m(k)$ will converge to a stationary point, and the gradient $\Delta \hat{\phi}_w^m(k)$ will converge to zero when $m \rightarrow \infty$. Using (11), OBP can incrementally improve $\hat{\phi}^m$ to maximize the log-likelihood $\ell(\cdot)$ of the joint probability of LDA within the online EM framework [23], [24], [25],

$$\ell(\hat{\phi}^{m+1}) \geq \ell(\hat{\phi}^m). \quad (14)$$

More detailed proof of (14) can be referred to [12]. In this sense, when $m \rightarrow \infty$, OBP can converge to the local optimum of the LDA's log-likelihood function.

Similarly, we show that POBP in Fig. 4 can also achieve this goal on N processors. As far as the m th mini-batch is concerned, the global $\hat{\phi}_w^{m-1}(k)$ of the previous mini-batch remains unchanged for N processors (line 5). Indeed, all processors update just the local gradient $\Delta \hat{\phi}_w^{m,n,t}(k)$ from the current mini-batch $x_{w,d}^{m,n}$, and communicate this local gradient according to (4) as follows,

$$\Delta \hat{\phi}_w^{m,\dots,t}(k) = \Delta \hat{\phi}_w^{m,\dots,t-1}(k) + \sum_{n=1}^N [\Delta \hat{\phi}_w^{m,n,t}(k) - \Delta \hat{\phi}_w^{m,\dots,t-1}(k)], \quad (15)$$

```

input   :  $\mathbf{x}_{W \times D}, K, \lambda_K, \lambda_W, \alpha, \beta$ .
output :  $\hat{\phi}_{K \times W}$ .
1 for  $m \leftarrow 1$  to  $M$  do
2   for each processor in parallel  $n \leftarrow 1$  to  $N$  do
3      $\mu_{w,d}^{m,n,0}(k) \leftarrow$  random initialization and normalization;
4      $\hat{\theta}_d^{m,n,0}(k) \leftarrow \sum_w x_{w,d} \mu_{w,d}^{m,n,0}(k)$ ;
5      $\hat{\phi}_w^{m,n,0}(k) \leftarrow \hat{\phi}_w^{m-1,n}(k) + \sum_d x_{w,d} \mu_{w,d}^{m,n,0}(k)$ ; // stochastic gradient descent
6      $\mu_{w,d}^{m,n,1}(k) \leftarrow \text{normalize}([\hat{\theta}_{-w,d}^{m,n,0}(k) + \alpha][\hat{\phi}_w^{m,n,0}(k) + \beta]/[\hat{\phi}_{-(w,d)}^{m,n,0}(k) + W\beta])$ ; // update messages
7      $r_w^{m,n,1}(k) \leftarrow \sum_d x_{w,d} |\mu_{w,d}^{m,n,1}(k) - \mu_{w,d}^{m,n,0}(k)|$ ; // update residuals
8      $\hat{\theta}_d^{m,n,1}(k), \hat{\phi}_w^{m,n,1}(k) \leftarrow \text{update}(\mu_{w,d}^{m,n,1}(k))$ ; // update sufficient statistics
9      $r_w^{m,\cdot,1}(k) \leftarrow \sum_n r_w^{m,n,1}(k), r_w^{m,\cdot,1} \leftarrow \sum_k r_w^{m,\cdot,1}(k), r_w^{m,n,1}(k) \leftarrow r_w^{m,\cdot,1}$ ;
10     $\hat{\phi}_w^{m,\cdot,1}(k) \leftarrow \sum_n \hat{\phi}_w^{m,n,1}(k), \hat{\phi}_w^{m,\cdot,1}(k) \leftarrow \hat{\phi}_w^{m,\cdot,1}(k)$ ;
11    // communicate  $r_{K \times W}$  and  $\hat{\phi}_{K \times W}$ 
12     $\lambda_W W \leftarrow \text{partial sort}(r_w^{m,\cdot,1}, \text{'descend'})$ ; // select power words
13     $\lambda_K K \leftarrow \text{partial sort}(r_w^{m,\cdot,1}(k), \text{'descend'})$ ; // select power topics
14    for  $t \leftarrow 2$  to  $T$  do
15      for  $w \in \lambda_W W$  do
16        for  $k \in \lambda_K K$  do
17           $\mu_{w,d}^{m,n,t}(k) \leftarrow \text{normalize}([\hat{\theta}_{-w,d}^{m,n,t-1}(k) + \alpha][\hat{\phi}_w^{m,n,t-1}(k) + \beta]/[\hat{\phi}_{-(w,d)}^{m,n,t-1}(k) + W\beta])$ ;
18           $r_w^{m,n,t}(k) \leftarrow \sum_d x_{w,d} |\mu_{w,d}^{m,n,t}(k) - \mu_{w,d}^{m,n,t-1}(k)|$ ;
19          // update the subset of messages and residuals
20           $\hat{\theta}_d^{m,n,t}(k), \hat{\phi}_w^{m,n,t}(k) \leftarrow \text{update}(\mu_{w,d}^{m,n,t}(k))$ ; // update sufficient statistics
21        end
22      end
23       $r_w^{m,\cdot,t}(k) = r_w^{m,\cdot,t-1}(k) + \sum_n [r_w^{m,n,t}(k) - r_w^{m,\cdot,t-1}(k)], r_w^{m,\cdot,t} \leftarrow \sum_k r_w^{m,\cdot,t}(k), r_w^{m,n,t}(k) \leftarrow r_w^{m,\cdot,t}(k)$ ;
24       $\hat{\phi}_w^{m,\cdot,t}(k) = \hat{\phi}_w^{m,\cdot,t-1}(k) + \sum_n [\hat{\phi}_w^{m,n,t}(k) - \hat{\phi}_w^{m,\cdot,t-1}(k)], \hat{\phi}_w^{m,\cdot,t}(k) \leftarrow \hat{\phi}_w^{m,\cdot,t}(k)$ ;
25      // communicate the subsets  $r_{\lambda_K K \times \lambda_W W}$  and  $\hat{\phi}_{\lambda_K K \times \lambda_W W}$ 
26      if  $\sum_w r_w^{m,\cdot,t} / \sum_{w,d} x_{w,d} \leq 0.1$  then break;
27       $\lambda_W W \leftarrow \text{partial sort}(r_w^{m,\cdot,t}, \text{'descend'})$ ; // select power words dynamically
28       $\lambda_K K \leftarrow \text{partial sort}(r_w^{m,\cdot,t}(k), \text{'descend'})$ ; // select power topics dynamically
29    end
30  end
31 end

```

Fig. 4. The POBP algorithm for LDA.

where the synchronized gradient is almost the same with (11). Also, the learning rate is still $1/(m-1)$, which guarantees the convergence of POBP. If we do not communicate at each iteration, Eq. (15) produces the inaccurate local gradient (11), and thus leads to a slow convergence speed. However, from (11) and (15), lowering the communication rate does not change the convergence property of POBP, but reduces its convergence speed. The proposed POBP communicates more frequently than its offline counterparts as shown in Fig. 1, which ensures its superiority over offline algorithms in terms of convergence performance.

3.2.2 Complexity and Scalability

Table 2 compares the complexities of POBP with those of OBP [12] and PGS [15] algorithms. For simplicity, we assume that the number of non-zero element in $\mathbf{x}_{W \times D}$ is ηWD , where η is a very small constant value depending on the data sets because $\mathbf{x}_{W \times D}$ is very sparse. Similarly, we assume that the total number of

word tokens in $\mathbf{x}_{W \times D}$ is $\eta' WD = \sum_{w,d} x_{w,d}$, where η' is also a constant value depending on data sets. Generally, $\eta \ll \eta'$ for most data sets. Suppose that the overall computation cost is A , and the communication cost for each processor is B , and thus the overall cost of N processors can be simplified as

$$\text{Overall cost} = \frac{A}{N} + BN, \quad (16)$$

where

$$N^* = \sqrt{\frac{A}{B}}, \quad (17)$$

minimizes the overall cost (16) to $2\sqrt{AB}$. From (17), we see that it is the ratio between computation and communication costs that determines the scalability, i.e., the best number of processors for the minimum overall cost. Note that in practice the communication cost per processor B is a variable that depends also on the bandwidth limitation between processors. When N increases, B will also increase nonlinearly

TABLE 2
Comparison of complexities.

Algorithms	Computation cost	Memory cost	Communication cost
POBP	$\eta\lambda_K\lambda_W KWD T/N$	$K(\eta WD + D)/MN + 2KW$	$\lambda_K\lambda_W KWMNT'$
OBP [12]	$\eta\lambda_K\lambda_W KWD T$	$K(\eta WD + D)/M + 2KW$	—
PGS [15]	$\eta' KWD T'/N$	$(K \times D + \eta' WD)/N + KW$	$NKWT'$

due to complex communication operations over limited bandwidth. Although Eq. (16) is a simplified estimation of relationship between computation and communication costs, it provides clues for estimation of the optimal number of processors in practice. For simplicity, we use (16) and (17) in the following analysis, where we use the size of communicated and synchronized matrices of each processor in Table 2 to approximate B .

For each mini-batch at the first iteration ($t = 1$), POBP requires to scan the entire mini-batch and communicate two complete matrices $\hat{\phi}_{K \times W}$ and $r_{K \times W}$. In the following iterations, POBP scans only the subset of mini-batch, and communicate the subsets of matrices $\hat{\phi}_{K \times W}$ and $r_{K \times W}$. Since the number of iterations for convergence is often very large (for example, $T \approx 200$), the total computation and communication costs are dominated by the rest iterations ($2 \leq t \leq T$). So, we approximate the overall computation and communication costs (without considering the small partial sorting costs) shown in Table 2. The real-world costs are proportional to these values. According to (16) and (17), the best number of processors is

$$N^* \propto \sqrt{\frac{\eta D}{M}} = \sqrt{\eta D_m}, \quad (18)$$

and the minimal overall cost is

$$\text{POBP's minimum cost} \propto 2\lambda_K\lambda_W KWT\sqrt{\eta DM}. \quad (19)$$

This analysis is consistent with our intuition that the best number of processors in POBP scales linearly with the mini-batch size D_m . When $M = 1$, POBP reduces to the parallel batch BP algorithm with the minimum overall cost when the best number of processors reaches the maximum.

However, the memory cost of each processor becomes very high shown in Table 2 because we have to store the local message matrix $\mu_{K \times \eta WD}$, the document-topic matrix $\hat{\theta}_{K \times D/M}$, the global topic-word matrix $\hat{\phi}_{K \times W}$ and the residual matrix $r_{K \times W}$. Therefore, POBP provides a flexible solution by setting the number of mini-batches M for big topic modeling tasks. When each processor has enough memory space, we can set the smaller number of mini-batches M and use more processors for the fast speed. When there is not enough memory for each processor, we can set larger number of mini-batches M and use less processors for a relatively slow speed. Note that the minimum overall cost of POBP scales with the square root \sqrt{DM} , which is often significantly lower

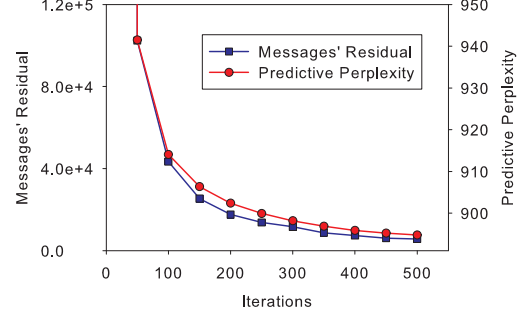


Fig. 5. The residual (blue curve) and predictive perplexity (red curve) as a function of iterations on ENRON. The predictive perplexity goes down with the residual, which indicates the convergence.

than that of the OBP (e.g., OBP scales linearly with the number of documents D) on a single processor shown in Table 2. Moreover, POBP uses less memory of each processor than OBP. In this sense, POBP is more suitable than OBP for big topic modeling tasks in real-world applications.

Table 2 also shows the complexities of PGS algorithm [15], which is one of the widely-used big topic modeling solutions introduced in Section 2. Its computation scales linearly with the number word tokens in $\mathbf{x}_{W \times D}$. According to (16) and (17), the best number of processors is $\sqrt{\eta' D}$ and the minimum overall cost is $2KWT'\sqrt{\eta' D}$. Obviously, POBP often has the lower minimum cost (19) than that of PGS. Moreover, POBP consumes less memory than PGS, so that it is more suitable for big topic modeling tasks. Indeed, if $\lambda_W = 0.1$ and $\lambda_K = 50/K$ (See experiments in subsection 4.1), POBP's minimum cost is insensitive to the number of topics K and the vocabulary size. This is a good property since big data sets often contain a big number of topics and vocabulary words [26]. Although parallel FGS (PFGS) [6] and SGS (PSGS) [21] are also insensitive to the number of topics K , they still consume more memory space than POBP. Also, lowering the computation cost instead of communication cost will make the scalability worse as shown in Eq. (17), i.e., the best number of processors N^* will become smaller.

3.3 Power Law Explanation

Power law, also known as the long-tail principle or the 80/20 rule [36], refers to the fact that a major

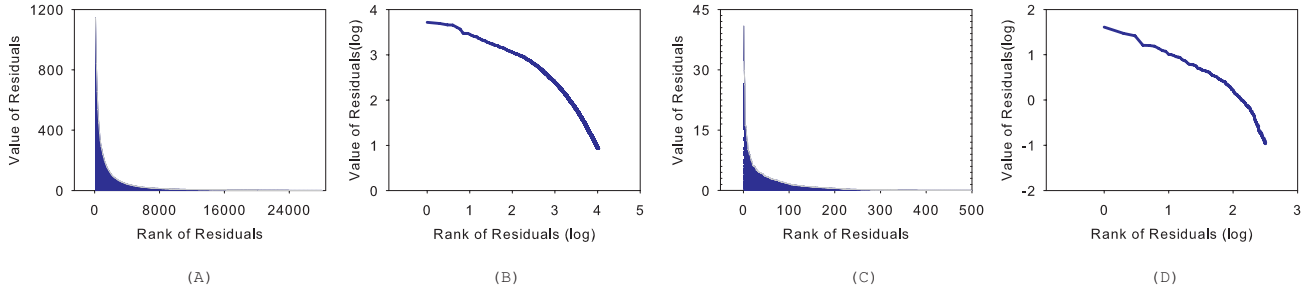


Fig. 6. The message value as a function of the message rank when $K = 500$ at the 10th iteration on ENRON. (A) Linear plot for message rank for vocabulary words. (B) Log-log plot for message rank for vocabulary words. (C) Linear plot for message rank for topics. (D) Log-log plot for message rank for topics.

proportion of effects come from a small fraction of the causes for many events. We show that the selected power words and topics based on residuals (9) and (10) follow power law. In this paper, we take the ENRON data set [6] as an example to show the appearance of power law. We set the number of topics as 500 and select the messages at the 10th iteration.

First, we show that the residual (7) can evaluate the convergence of topic modeling process, where the predictive perplexity (9) has been widely used as the convergence condition of LDA algorithms [3], [5], [33]. Fig. 5 compares the predictive perplexity of LDA and the average residual over all words. We see that the two curves have almost the same trend reflecting the convergence state. This is why we use the average residual as the convergence condition in the POBP algorithm in Fig. 4 (line 26). Intuitively, if residuals become zeros, the message values do not change so that the parameters are fixed at the local optimum. In this sense, we can speed up convergence by minimizing the larger residuals first and then the smaller residuals. This is the first motivation of our two-step selection method in communication-efficient MPA in subsection 3.1.

Second, we show that the distribution of residuals approximately follows power law at each mini-batch iteration. A simple way to identify power-law behavior in either natural or man-made systems is to draw a histogram with both axis plotted on logarithmic scales called log-log plot [36]. If the log-log plot approximates a straight line, we consider that power law applies. We sort the residuals in descending order. We draw the linear plot for r_w in Fig. 2 with the x-axis for residual ranks and y-axis for residual values. Fig. 6A indicates that a small fraction of words take a vary large proportion of residuals. Fig. 6B shows that the corresponding log-log plot approximately follows power law. This phenomenon confirms that only a small subset of vocabulary words contribute almost all residual values. More specifically, the top 10% words account for 79% of the total residual value, while the top 20% words account for almost 90% of

the total residual value. Therefore, it is efficient to minimize residuals of those power words first to speed up the convergence. Fig. 6C shows the linear plot for $r_w(k)$ in Fig. 2, and Fig. 6D shows the corresponding log-log plot. Both confirm that the residual distribution of power topics approximately follows power law. Therefore, we only need to do computation and communication for power words and topics, which will be updated through the dynamical scheduling at each iteration in Fig. 3.

4 EXPERIMENTS

We compare the proposed POBP with parallel FGS (PFGS) [6], parallel SGS (PSGS) [21], Yahoo LDA (YLDA) [14], and parallel variational Bayes (PVB) [19]. All these benchmark algorithms have open source codes. For a fair comparison, we re-write their source codes in C++ language [37]. Also, we use the integer type to store LDA parameters in the GS-based algorithms, while we use the single-precision floating-point format to represent LDA parameters in both PVB and POBP algorithms. Such an implementation difference is caused by the sampling process in the GS-based algorithms [4].

We run the above algorithms on a cluster with up to 1024 processors (1.9GHz CPU, 2GB memory) to perform the experiments. All the processors communicate through a high-speed Infiniband with 20GB per second bandwidth. Following [6], we use the fixed hyper-parameters $\alpha = 2/K$ and $\beta = 0.01$ for all algorithms to guarantee a fair comparison. To reach the convergence state, we run PFGS, PSGS, YLDA and PVB using 500 iterations [15]. For POBP, we set $NNZ \approx 45,000$ in each-mini batch since OBP's performance is insensitive to the mini-batch size [12]. Also, this mini-batch size can be easily fit into 2GB memory of each processor. We evenly distribute D documents to N processors to avoid load imbalance.

We use four publicly available data sets: ENRON, NYTIMES, PUBMED [6] and WIKIPEDIA,¹ where

1. <http://en.wikipedia.org>

TABLE 3
Summarization of four data sets.

Data sets	D	W	N_{token}	NNZ	Size (M)
ENRON	39,861	6,536	6,412,172	2,374,385	28.34
NYTIMES	300,000	7,871	99,542,125	44,379,275	568.88
WIKIPEDIA	4,360,095	5,363	665,375,061	154,934,308	1983.77
PUBMED	8,200,000	6,902	737,869,083	222,399,377	3043.04

ENRON is a relatively smaller data set so that we use it for parameters tuning. The other three data sets are relatively bigger with up to 8 million documents and we use them for web-scale experiments. We follow [11] and remove the words out of a fixed truncated vocabulary to get a shorter vocabulary because some vocabulary words occur rarely and contribute little to topic modeling. While the vocabulary size W has been greatly reduced, most of the word tokens N_{token} and none-zero-elements NNZ are still reserved. For example, though we reduce the vocabulary size of PUBMED from 141,043 to 6,902 with a ratio of 4.89%, we reduce the number of word tokens from about 7 million to 3 million with a ratio over 40%. As a result, we can fit the word-topic distribution $\phi_{K \times W}$ in 2GB memory of each processor when K is large. Table 3 summarizes the statistics of data sets, where D denotes the number of documents, W the vocabulary size, N_{token} the number of word tokens, NNZ the number of non-zero elements, and “Size (M)” size of data sets in MByte.

We use the predictive perplexity (\mathcal{P}) [5], [33] to measure accuracy of different parallel LDA algorithms. To calculate the predictive perplexity, we randomly partition each document into 80% and 20% subsets. Fixing the word-topic distribution $\phi_{K \times W}$, we estimate $\theta_{K \times D}$ on the 80% subset by the training algorithms from the same random initialization after 500 iterations, and then calculate the predictive perplexity on the rest 20% subset,

$$\mathcal{P} = \exp \left\{ - \frac{\sum_{w,d} x_{w,d}^{20\%} \log [\sum_k \theta_d(k) \phi_w(k)]}{\sum_{w,d} x_{w,d}^{20\%}} \right\}, \quad (20)$$

where $x_{w,d}^{20\%}$ denotes word counts in the the 20% subset. The lower predictive perplexity represents a higher accuracy.

4.1 Ratios λ_W and λ_K

POBP introduces two parameters λ_W and λ_K to control the ratio of power words and topics at each iteration. The parameter λ_K determines the ratio of power topics evolved at each iteration. The smaller λ_K will lead to less computation and communication cost. However, this may also result in a lower topic modeling accuracy. In practice, each word may not be allocated to many topics, and thus $\lambda_K K$ is often a fixed value. To study the effect of different $\lambda_K K$, we evaluate a range of $\lambda_K K$ values on the ENRON data set when $K = 500$.

TABLE 4
Perplexity gap between POBP and PFGS.

K	NYTIMES	WIKIPEDIA	PUBMED
500	24.41%	31.64%	48.54%
1000	24.57%	36.07%	60.46%
2000	24.69%	39.51%	66.68%

Fig. 7A shows the predictive perplexity and training time as a function of λ_W by fixing $\lambda_K = 1$, where $\lambda_W = 1$ denotes that all the vocabulary words are scanned at each iteration. We decrease the value of λ_W from 0.4 to 0.025 in an exponential manner. While the training time decreases with the decrease of λ_W , the predictive perplexity also increases indicating a degraded performance. However, when $\lambda_W \geq 0.1$, the increase of perplexity is so small that can be neglected. This result confirms that a subset of power words at each iteration contributes to almost all topic modeling performance. Also, we see that a small value of λ_W may lead to an increase of perplexity. For example, when $\lambda_W = 0.025$, the predictive perplexity increases around 8% to 526.8.

Fig. 7B shows the predictive perplexity and training time as a function of $\lambda_K K$ by fixing $\lambda_W = 1$, where $\lambda_K K = 500$ means that all the topics are scanned at each iteration. We change $\lambda_K K$ from 30 to 70 in a step of 10. The results show that the predictive perplexity increases slightly and the training time decreases steadily with the decrease of $\lambda_K K$. Fig. 7B also confirms that a subset of power topics plays an important role in topic modeling. Finally, we combine different values of λ_W and $\lambda_K K$. Fig. 7C shows that $\{\lambda_W = 0.1, \lambda_K K = 50\}$ can achieve a reasonable speedup while keeping a high accuracy (e.g., the predictive perplexity change is within 15). We also use this setting in subsection 3.2.2 for complexity and scalability analysis.

4.2 Accuracy

Fig. 8 shows the predictive perplexity as a function of training time (in second log-scale) on NYTIMES, PUBMED and WIKIPEDIA using 256 processors when $K = 2000$. We see that POBP converges fastest among all the algorithms, around 10 to 100 times faster than GS-based algorithms and 50 to 400 times faster than PVB. This result is consistent with our convergence analysis in subsection 3.2.1. Also, POBP always reaches the lowest predictive perplexity, indicating its

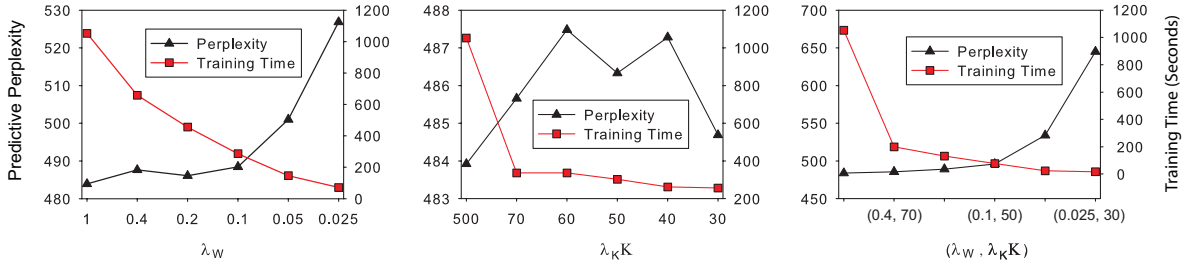


Fig. 7. Predictive perplexity and training time as a function of λ_K and λ_W on ENRON data set. We fix $K = 500$ and use 12 processors. The left axis denotes the predictive perplexity and the right axis denotes the training time in seconds. (A) Fixing $\lambda_K = 1$, we test different $\lambda_W = \{0.025, 0.05, 0.1, 0.2, 0.4, 1\}$. (B) Fixing $\lambda_W = 1$, we test different $\lambda_K K = \{30, 40, 50, 60, 70, 500\}$. (C) We test some combinations of λ_W and $\lambda_K K$. We see that when $\lambda_W = 0.1$ and $\lambda_K K = 50$ POBP can achieve a significant speedup while achieving a good accuracy.

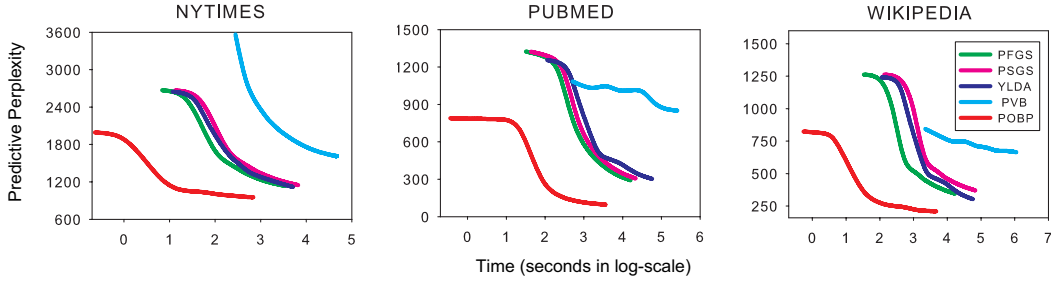


Fig. 8. Predictive perplexity as a function of training time (in second log-scale) on NYTIMES, PUBMED and WIKIPEDIA data sets using 256 processors when $K = 2000$.

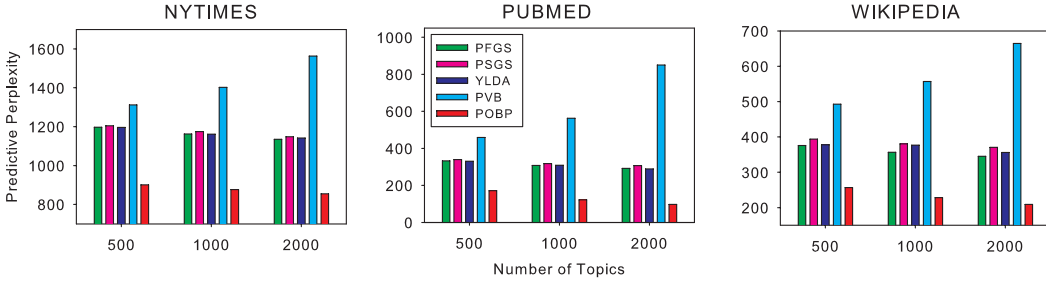


Fig. 9. Comparison of predictive perplexity for all algorithms on NYTIMES, PUBMED and WIKIPEDIA data sets using 256 processors, where the number of topics $K \in \{500, 1000, 2000\}$.

good convergence property. Fig. 9 also shows that POBP yields the lowest predictive perplexity on all data sets given different number of topics on 256 processors. The GS-based algorithms such as PFGS, PSGS and YLDA have slightly higher perplexity while PVB produces the highest perplexity. These results are consistent with observations in previous work [5], [22], [33]. We see that the predictive perplexity of PVB increases with the number of topics partly due to the overfitting phenomenon.

Table 4 compares the perplexity gap between POBP and PFGS calculated by

$$\text{gap} = \frac{\mathcal{P}_{\text{PFGS}} - \mathcal{P}_{\text{POBP}}}{\mathcal{P}_{\text{PFGS}}} \times 100\%, \quad (21)$$

where \mathcal{P} is the predictive perplexity (20). When $K = 500$, the gap is about 24.41% on relatively smaller data set NYTIMES but the gap increases to 31.64% and

48.54% on larger data sets WIKIPEDIA and PUBMED, respectively. Besides, the gap increases for all data sets when K increases from 500 to 2000. Such an excellent predictive performance makes POBP a very competitive topic modeling algorithm on real-world big data streams.

4.3 Communication Time

Fig. 10 shows the communication time (in second log-scale) of all algorithms on NYTIMES, PUBMED and WIKIPEDIA using 256 processors when $K \in \{500, 1000, 2000\}$. We see that POBP consumes around 5% ~ 20% communication time of other algorithms on all data sets. Among all algorithms, PVB has the longest communication time because the topic-word distribution $\hat{\phi}_{K \times W}$ in PVB is of single-precision floating type, leading to an approximately double communication amount than that of GS-based algorithms

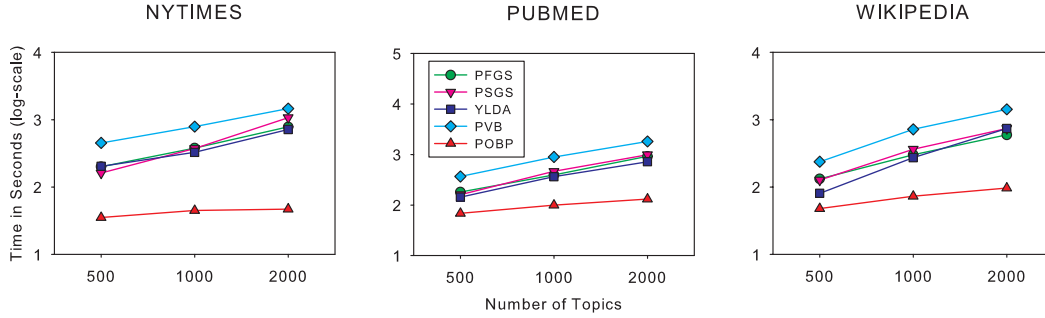


Fig. 10. The communication time (in second log-scale) on NYTIES, PUBMED and WIKIPEDIA using 256 processors when $K \in \{500, 1000, 2000\}$.

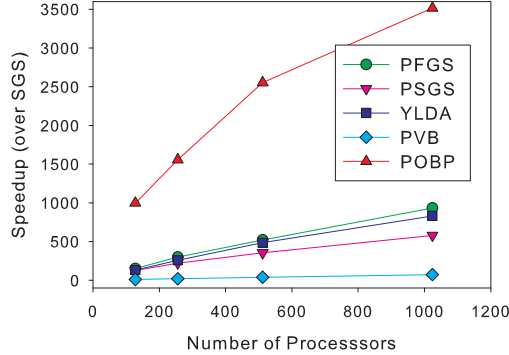


Fig. 12. The speedup performance when $K = 2000$. We choose 1/128 training time of PSGS [21] on 128 processors as the baseline.

using integer type. Although POBP also store $\hat{\phi}_{K \times W}$ in single-precision floating-point format, it selects only a subset of matrix $\hat{\phi}_{K \times W}$ for communication in subsection 3.1. Hence, POBP is more communication-efficient than GS-based algorithms. According to the analysis in subsection 3.2.2, the total communication time of POBP is proportional to the number of mini-batches M . In our experiments, the number of mini-batches on NYTIES, PUBMED and WIKIPEDIA is 6, 19 and 17, respectively. Therefore, POBP has the least total communication time on NYTIES. This result suggests that if the memory is big enough, we should try to minimize the number of mini-batches M in POBP to reach the minimum communication time.

4.4 Speed and Scalability

Fig. 11 shows the training time of all algorithms as a function of the number of topics. We see that POBP is the fastest among all algorithms. PFGS, PSGS and YLDA have a comparable speed, and PVB runs the slowest. On all data sets, POBP is around 5 to 100 times faster than other algorithms. Such a high speed has been largely attributed to three reasons. First, POBP has the least communication time as shown in Fig. 10. Second, POBP runs fast at each iteration

TABLE 5
Memory usage (MB) on PUBMED when $K = 2000$.

N	PFGS	PSGS/YLDA	PVB	POBP
1024	349	279	438	1, 133
512	541	349	560	1, 133
256	924	487	804	1, 133
128	1, 690	765	1, 293	1, 133
64	N/A	1320	N/A	1, 133
32	N/A	N/A	N/A	1, 133

because it selects the subset of words and topics for computation as shown in Fig. 4. Finally, POBP converges very fast as shown in Fig. 8.

We use the speedup performance with the number of processors [15] to evaluate the scalability of parallel algorithms. We choose the 1/128 training time of PSGS on 128 processors as baseline, which approximates the training time of SGS on a single processor without parallelization. Then, the speedup is calculated as the ratio between the baseline and the training time of other parallel algorithms. Fig. 12 shows the speedup performance of all algorithms on PUBMED when $K = 2000$. We show the speedup curve on $N \in \{128, 256, 512, 1024\}$ processors. Although the speedup curve of POBP bends earlier than other algorithms, POBP always has much better speedup performance than other parallel algorithms. This phenomenon confirms that POBP requires only a small number of processors N^* in (18) to achieve the best speedup performance, while other parallel algorithms often need more processors to fulfill it. Moreover, the best performance of POBP is much better than those of other algorithms following the analysis in subsection 3.2.2. In this sense, POBP has a good scalability because it uses the least number of processors to achieve a much better speedup performance than other parallel algorithms.

4.5 Memory Usage

Big topic modeling tasks are often limited by the memory space of each processor. Table 5 shows the memory usage of all algorithms in each processor on the PUBMED data set when $K = 2000$. The memory

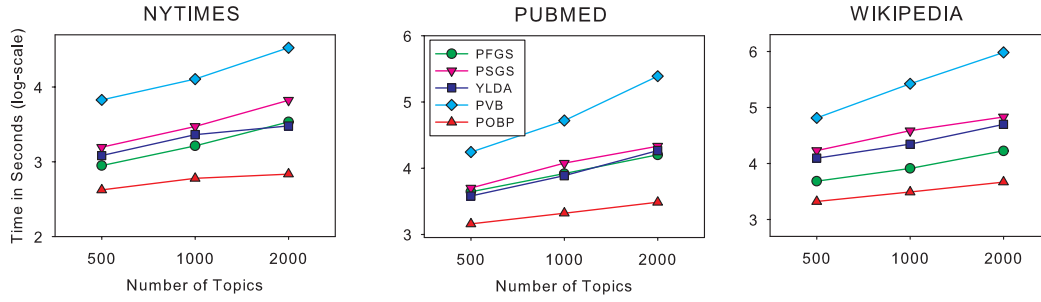


Fig. 11. Training time in second (log-scale) of all algorithms on NYTIMES, PUBMED and WIKIPEDIA when $K \in \{500, 1000, 2000\}$ using 256 processors.

usage of PSGS, YLDA, PFGS and PVB decreases with the number of processors, while POBP consumes a constant memory space. The major reason is that parallel batch LDA algorithms can distribute both data $\mathbf{x}_{W \times D}$ and document-topic matrix $\hat{\theta}_{K \times D}$ into N processors, so that the entire memory usage of each processor will decrease linearly with N . However, when N is small, parallel batch LDA algorithms may not load $1/N$ data and document-topic matrix into memory for computation (e.g., when $N \leq 64$, PFGS and PVB fail to process PUBMED in Table 5). On the other hand, POBP is an online algorithm that loads only a mini-batch of data and document-topic matrix into memory, which is a constant value dependent on the mini-batch size D_m . In practice, users can provide D_m according to each processor's memory quota. Generally, we maximize D_m to reduce M for the minimum communication time (19). To further reduce the memory usage of POBP, we may use hard disk as extended memory to store the word-topic matrix $\hat{\phi}_{K \times W}$ like [12]. Another strategy is to distribute $\hat{\phi}_{K \times W}$ into N processors by adding more communication costs. In this way, we can extract more topics from more vocabulary words without truncation in our experimental settings.

5 CONCLUSIONS

This paper proposes a novel parallel multi-processor architecture (MPA) for big topic modeling tasks. This communication-efficient MPA can be combined with both batch and online LDA algorithms. For example, we combine this MPA with OBP [12] referred to as the POBP algorithm for big data streams in this paper. At each iteration, POBP computes and communicates the subsets of vocabulary words and topics called power words and topics, and thus has very low computation, memory and communication costs. Extensive experiments on big data sets confirm that POBP is faster, lighter, and more accurate than other state-of-the-art parallel LDA algorithms, such as parallel fast Gibbs sampling (PFGS) [6], parallel sparse Gibbs sampling (PSGS) [21], Yahoo LDA (YLDA) [14], and parallel variational Bayes (PVB) [19]. Therefore,

POBP is very competitive for web-scale topic modeling applications, which require a high processing speed under limited resources or seek a high processing efficiency/cost performance. Since POBP can be interpreted within the EM framework, its basic idea can be generalized to speed up parallel batch or online EM algorithms for other latent variable models. Besides, the power law explanation may shed more light on building faster big learning algorithms such as deep learning algorithms with high performance computing systems [38].

Future work may include two parts. First, we still need to investigate the multi-core architecture (MCA) such as GPU clusters for big topic modeling in the shared memory systems [39]. We may avoid serious race conditions by dynamical scheduling of non-conflict subsets of vocabulary words and topics. Second, we need to study how to apply POBP in other parallel paradigms like in-memory Map-Reduce (Spark)² or Graph-Lab/Chi.³

ACKNOWLEDGEMENTS

This work is supported by NSFC (Grant No. 61272449, 61202029, 61003154, 61373092 and 61033013), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 12KJA520004), Innovative Research Team in Soochow University (Grant No. SDT2012B02) to JFY and JZ, and a GRF grant from RGC UGC Hong Kong (GRF Project No. 9041574), a grant from City University of Hong Kong (Project No. 7008026) to ZQL.

REFERENCES

- [1] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.
- [2] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, pp. 77–84, 2012.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Natl. Acad. Sci.*, vol. 101, pp. 5228–5235, 2004.

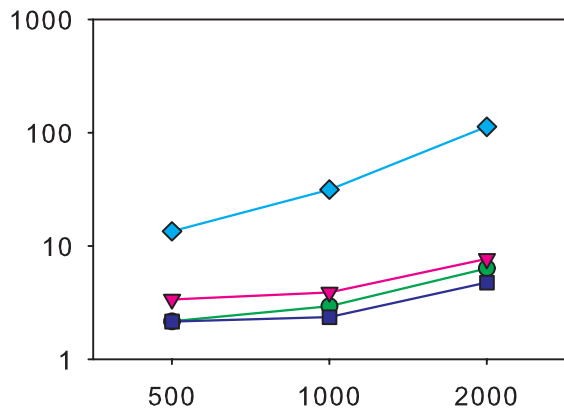
2. <http://spark.incubator.apache.org/>

3. <http://graphlab.org/>

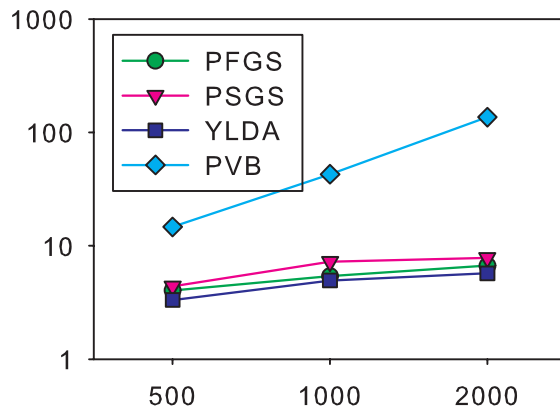
- [5] J. Zeng, W. K. Cheung, and J. Liu, "Learning topic models by belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1121–1134, 2013.
- [6] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed Gibbs sampling for latent Dirichlet allocation," in *KDD*, 2008, pp. 569–577.
- [7] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," pp. 937–946, 2009.
- [8] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "A new approach to speeding up topic modeling," *arXiv preprint arXiv:1204.0170*, 2012.
- [9] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *ICML*, 2013.
- [10] L. Bottou, *Online learning and stochastic approximations*. Cambridge University Press, 1998.
- [11] M. Hoffman, D. Blei, and F. Bach, "Online learning for latent Dirichlet allocation," in *NIPS*, 2010, pp. 856–864.
- [12] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "Online belief propagation for topic modeling," *arXiv preprint arXiv:1210.2179*, 2012.
- [13] F. Yan, N. Xu, and Y. Qi, "Parallel inference for latent Dirichlet allocation on graphics processing units," in *NIPS*, 2009, pp. 2134–2142.
- [14] A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. Smola, "Scalable inference in latent variable models," in *WSDM*, 2012, pp. 123–132.
- [15] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *J. Mach. Learn. Res.*, vol. 10, pp. 1801–1828, 2009.
- [16] Y. Wang, H. Bai, M. Stanton, W. Y. Chen, and E. Chang, "PLDA: Parallel latent Dirichlet allocation for large-scale applications," in *Algorithmic Aspects in Information and Management*, 2009, pp. 301–314.
- [17] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–18, 2011.
- [18] A. Smola and S. Narayanamurthy, "An architecture for parallel topic models," in *PVLDB*, 2010, pp. 703–710.
- [19] K. Zhai, J. Boyd-Graber, and N. Asadi, "Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce," in *WWW*, 2012, pp. 879–888.
- [20] M. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.
- [21] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *KDD*, 2009, pp. 937–946.
- [22] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "A new approach to speeding up topic modeling," p. arXiv:1204.0170 [cs.LG], 2012.
- [23] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," vol. 89, pp. 355–368, 1998.
- [24] P. Liang and D. Klein, "Online EM for unsupervised models," in *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 2009, pp. 611–619.
- [25] O. Cappé and E. Moulines, "Online expectation-maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B*, vol. 71, no. 3, pp. 593–613, 2009.
- [26] K. Zhai and J. Boyd-Graber, "Online latent Dirichlet allocation with infinite vocabulary," in *ICML*, 2013, pp. 561–569.
- [27] M. Wahabzada and K. Kersting, "Larger residuals, less work: Active document scheduling for latent Dirichlet allocation," in *ECML/PKDD*, 2011, pp. 475–490.
- [28] D. Mimno, M. D. Hoffman, and D. M. Blei, "Sparse stochastic inference for latent Dirichlet allocation," in *ICML*, 2012.
- [29] G. Elidan, I. McGraw, and D. Koller, "Residual belief propagation: Informed scheduling for asynchronous message passing," in *UAI*, 2006, pp. 165–173.
- [30] J. Zeng, X.-Q. Cao, and Z.-Q. Liu, "Residual belief propagation for topic modeling," in *ADMA*, 2012, pp. 739–752.
- [31] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.
- [32] J. Yan, Z.-Q. Liu, Y. Gao, and J. Zeng, "Communication-efficient parallel belief propagation for latent Dirichlet allocation," p. arXiv:1206.2190v1 [cs.LG], 2012.
- [33] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *UAI*, 2009, pp. 27–34.
- [34] N. de Freitas and K. Barnard, "Bayesian latent semantic analysis of multimedia databases," University of British Columbia, Tech. Rep., 2001.
- [35] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 198–207, 2008.
- [36] R. Sanders, "The pareto principle: Its use and abuse," *Journal of Product & Brand Management*, vol. 1, no. 2, pp. 37–40, 1992.
- [37] J. Zeng, "A topic modeling toolbox using belief propagation," *J. Mach. Learn. Res.*, vol. 13, pp. 2233–2236, 2012.
- [38] A. Coates, B. Huval, T. Wang, D. J. Wu, A. Y. Ng, and B. Catanzaro, "Deep learning with COTS HPC systems," in *ICML*, 2013.
- [39] Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin, "A fast parallel SGD for matrix factorization in shared memory systems," in *ACM Recommender Systems*, 2013.

Computation Cost Ratio/POBP

NYTIMES



PUBMED



WIKIPEDIA

