DSANLS: Accelerating Distributed Nonnegative Matrix Factorization via Sketching

Technical Report

Yuqiu Qian*[1], Conghui Tan†[2], Nikos Mamoulis‡[3], and David W. Cheung§[1]

[1]Department of Computer Science, The University of Hong Kong
[2]Department of SEEM, The Chinese University of Hong Kong
[3]Department of Computer Science & Engineering, University of Ioannina

Abstract

Nonnegative matrix factorization (NMF) has been successfully applied in different fields, such as text mining, image processing, and video analysis. NMF is the problem of determining two nonnegative low rank matrices U and V, for a given input matrix M, such that $M \approx UV^\top$. There is an increasing interest in parallel and distributed NMF algorithms, due to the high cost of centralized NMF on large matrices. In this paper, we propose a *distributed sketched alternating nonnegative least squares* (DSANLS) framework for NMF, which utilizes a matrix sketching technique to reduce the size of nonnegative least squares subproblems in each iteration for U and V. We design and analyze two different random matrix generation techniques and two subproblem solvers. Our theoretical analysis shows that DSANLS converges to the stationary point of the original NMF problem and it greatly reduces the computational cost in each subproblem as well as the communication cost within the cluster. DSANLS is implemented using MPI for communication, and tested on both dense and sparse real datasets. The results demonstrate the efficiency and scalability of our framework, compared to the state-of-art distributed NMF MPI implementation. Its implementation is available at `https://github.com/qianyuqiu79/DSANLS`.

1 Introduction

Nonnegative matrix factorization (NMF) is a technique for discovering nonnegative latent factors and/or performing dimensionality reduction. NMF finds applications in text mining [34], image/video processing [21], and analysis of social networks [38]. Unlike general matrix factorization (MF), NMF restricts the two output matrix factors to be nonnegative. Nonnegativity is inherent in the feature space of many real-world applications, therefore the resulting factors can have a natural interpretation. Specifically, the goal of NMF is to decompose a huge matrix $M \in \mathbb{R}_+^{m \times n}$ into the

*yqqian@cs.hku.hk
†chtan@se.cuhk.edu.hk
‡nikos@cs.uoi.gr
§dcheung@cs.hku.hk

product of two matrices $U \in \mathbb{R}_+^{m \times k}$ and $V \in \mathbb{R}_+^{n \times k}$ such that $M \approx UV^\top$. $\mathbb{R}_+^{m \times n}$ denotes the set of $m \times n$ matrices with nonnegative real values, and k is a user-specified dimensionality, where typically $k \ll m, n$.

Generally, NMF can be defined as an optimization problem [23] as follows:

$$\min_{U \in \mathbb{R}_+^{m \times k}, V \in \mathbb{R}_+^{n \times k}} \left\| M - UV^\top \right\|_F, \tag{1}$$

where $\|X\|_F = \left(\sum_{ij} x_{ij}^2 \right)^{1/2}$ is the Frobenius norm of X. However, Problem (1) is hard to solve directly because it is non-convex. Therefore, almost all NMF algorithms leverage two-block coordinate descent schemes. That is, they optimize just over one of the two factors, U or V, while keeping the other fixed [8]. By fixing V, we can optimize U by solving a nonnegative least squares (NLS) subproblem:

$$\min_{U \in \mathbb{R}_+^{m \times k}} \left\| M - UV^\top \right\|_F. \tag{2}$$

Modern data analysis tasks apply on big matrix data with increasing scale and dimensionality. Examples include community detection in a billion-node social network, background separation on a 4K video in which every frame has approximately 27 million rows [17], text mining on a bag-of-words matrix with millions of words. The volume of data is anticipated to increase in the 'big data' era, making it impossible to store the whole matrix in the main memory throughout NMF. Therefore, there is a need for high-performance and scalable distributed NMF algorithms.

In this paper, we propose a distributed framework for NMF. We choose MPI[1]/C for our distributed implementation for efficiency, generality and privacy reasons. MPI/C does not require reading/writing data to/from disk or global shuffles of data matrix entries, as Spark or MapReduce do. Nodes can collaborate without sharing their local input data, which is important for applications that involve sensitive data and have privacy considerations. Besides, high performance numerical computing routines like MKL[2] can be leveraged. The state-of-art implementation of distributed NMF is MPI-FAUN [17], a general framework that iteratively solves nonnegative least squares (NLS) subproblems for U and V. The main idea behind MPI-FAUN is to exploit the independence of local updates for rows of U and V, in order to minimize the communication requirements of matrix multiplication operations within the NMF algorithms.

Our idea is to speed up distributed NMF in a new, orthogonal direction: by reducing the problem size of each NLS subproblem within NMF, which in turn decreases the overall computation cost. In a nutshell, we reduce the size of each NLS subproblem, by employing a *matrix sketching* technique: the involved matrices in the subproblem are multiplied by a specially designed random matrix at each iteration, which greatly reduces their dimensionality. As a result, the computational cost of each subproblem drops.

However, applying matrix sketching comes with several issues. First, although the size of each subproblem is significantly reduced, sketching involves matrix multiplication which brings computational overhead. Second, unlike in a single machine setting, the data are distributed to different nodes, which may have to communicate extensively in a poorly designed solution. In particular, each node only retains part of both the input matrix and the generated approximate matrices, causing difficulties due to data dependencies in the computation process. Besides, the generated random

[1] Message Passing Interface

[2] Intel® Math Kernel Library

matrices should be the same for all nodes in every iteration, while broadcasting the random matrix to all nodes brings severe communication overhead and can become the bottleneck of distributed NMF. Furthermore, after reducing each original subproblem to a sketched random new subproblem, it is not clear whether the algorithm still converges and whether it converges to stationary points of the original NMF problem.

Our *distributed sketched alternating nonnegative least squares* (DSANLS) overcomes these problems. First, the extra computation cost due to sketching is reduced with a proper choice of the random matrices. Second, the same random matrices used for sketching are generated independently at each node, thus there is no need for communication of random matrices between nodes during distributed NMF. Having the complete random matrix at each node, an NMF iteration can be done locally with the help of a matrix multiplication rule with proper data partitioning. Therefore, our matrix sketching approach reduces not only the computational, but also the communication cost. Moreover, due to the fact that sketching also *shifts* the optimal solution of each original NMF subproblem, we propose subproblem solvers paired with theoretical guarantees of their convergence to a stationary point of the original subproblems.

Our contributions can be summarized as follows:

- We propose DSANLS, a novel high-performance distributed NMF algorithm. DSANLS is the first distributed NMF algorithm that leverages matrix sketching to reduce the problem size of each NLS subproblem and can be applied to both dense and sparse input matrices with a convergence guarantee.
- We propose a novel and specially designed subproblem solver (*proximal coordinate descent*), which helps DSANLS to converge faster. We also discuss the use of *projected gradient descent* as subproblem solver, showing that it is equivalent to stochastic gradient descent (SGD) on the original (non-sketched) NLS subproblem.
- We present a detailed theoretical analysis of DSANLS, and prove that DSANLS converges to a stationary point of the original NMF problem. This convergence proof is novel and non-trivial because of the involvement of matrix sketching at each iteration.
- We conduct an experimental evaluation using several (dense and sparse) real datasets, which demonstrates the efficiency and scalability of DSANLS.

The remainder of the paper is organized as follows. Section 2 briefly discusses the properties of NMF, reviews NMF algorithms and distributed NMF techniques, and introduces the matrix sketching technique. Our DSANLS algorithm is presented in Section 3. Detailed theoretical analysis of DSANLS algorithm is discussed in Section 4. Section 5 evaluates DSANLS. Finally, Section 6 concludes the paper.

2 Background and Related Work

2.1 Properties of NMF

In this paper, we focus on solving problem (1), which has several properties. First, general NMF is NP-hard [41], so typical methods aim at finding an approximate solution (i.e., a local optimum). The second property is that the search space of NMF has numerous local minimums, so the results of different algorithms may vary significantly [14]. Third, choosing the best value for the factorization

rank k is quite hard. Widely-used approaches are: trial and error, estimation using SVD, and the use of experts' insights [43].

2.2 NMF Algorithms

Almost all NMF algorithms leverage a two-block coordinate descent scheme (exact or inexact). That is, they optimize just over one of the two factors, U or V, while keeping the other fixed [8]. The reason is that the original problem (1) is non-convex, so it is hard to solve it directly. However, by fixing V, we can solve a convex subproblem:

$$\min_{U \in \mathbb{R}_+^{m \times k}} \left\| M - UV^\top \right\|_F, \tag{3}$$

More precisely, (3) is a nonnegative least squares (NLS) problem. Similarly, if we fix U, the problem becomes:

$$\min_{V \in \mathbb{R}_+^{n \times k}} \left\| M^\top - VU^\top \right\|_F. \tag{4}$$

Algorithm 1: Two-Block Coordinate Descent: Framework of Most NMF Algorithms

initialize $U_0 \geq 0$, $V_0 \geq 0$;
for $t = 0$ ***to*** $T - 1$ **do**
  $U_{t+1} \leftarrow$ update(M, U_t, V_t);
  $V_{t+1} \leftarrow$ update(M, U_{t+1}, V_t);
end
return U_T *and* V_T

The first widely used update rule is Multiplicative Updates (MU), which was first applied for solving NLS problems in [6]. Later, MU was rediscovered and used for NMF in [23]. MU is based on the majorization-minimization framework. Its application guarantees that the objective function monotonically decreases [6, 23].

Another extensively studied method is alternating nonnegative least squares (ANLS), which represents a class of methods where the subproblems for U and V are solved exactly following the framework described in Algorithm 1. ANLS is guaranteed to converge to a stationary point [11] and has been shown to perform very well in practice with active set [18, 20], projected gradient [26], quasi-Newton [46], or accelerated gradient [13] methods as the subproblem solver.

Hierarchical alternating least squares (HALS) [4] solves each NLS subproblem using an exact coordinate descent method that updates one individual column of U at a time. The optimal solutions of the corresponding subproblems can be written in a closed form.

2.3 Distributed NMF

Parallel NMF algorithms are well studied in the literature [15, 39]. However, different from a parallel, single machine setting, in a distributed setting, data sharing and communication have considerable cost. Therefore, we need specialized NMF algorithms for massive scale data handling in a distributed environment. The first method in this direction [27] is based on the MU algorithm. It mainly focuses on sparse matrices and applies a careful partitioning of the data in order to maximize data locality and parallelism. Later, CloudNMF [25], a MapReduce-based NMF algorithm similar

to [27], was implemented and tested on large-scale biological datasets. Another distributed NMF algorithm [45] leverages block-wise updates for local aggregation and parallelism. It also performs frequent updates using whenever possible the most recently updated data, which is more efficient than traditional concurrent counterparts. Apart from MapReduce implementations, Spark is also attracting attention for its advantage in iterative algorithms, e.g., using MLlib [31]. Finally, there are implementations using X10 [12] and on GPU [30].

The most recent and related work in this direction is MPI-FAUN [16, 17], which is the first implementation of NMF using MPI for interprocessor communication. MPI-FAUN is flexible and can be utilized for a broad class of NMF algorithms that iteratively solve NLS subproblems including MU, HALS, and ANLS/BPP. MPI-FAUN exploits the independence of local update computation for rows of U and V to apply communication-optimal matrix multiplication. In a nutshell, the full matrix M is split across a two-dimensional grid of processors and multiple copies of both U and V are kept at different nodes, in order to reduce the communication between nodes during the iterations of NMF algorithms.

2.4 Matrix Sketching

Matrix sketching is a technique that has been previously used in numerical linear algebra [10], statistics [36] and optimization [37]. Its basic idea is described as follows. Suppose we need to find a solution x to the equation:

$$Ax = b, \quad (A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m). \tag{5}$$

Instead of solving this equation directly, in each iteration of matrix sketching, a random matrix $S \in \mathbb{R}^{d \times m}$ $(d \ll m)$ is generated, and we instead solve the following problem:

$$(SA)x = Sb. \tag{6}$$

Obviously, the solution of (5) is also a solution to (6), but not vice versa. However, the problem size has now decreased from $m \times n$ to $d \times n$. With a properly generated random matrix S and an appropriate method to solve subproblem (6), it can be guaranteed that we will progressively approach the solution to (5) by iteratively applying this sketching technique.

To the best of our knowledge, there is only one piece of previous work [42] which incorporates dual random projection into the NMF problem, in a centralized environment, sharing similar ideas as SANLS, the centralized version of our DSANLS algorithm. However, Wang et al. [42] did not provide an efficient subproblem solver, and their method was less effective than non-sketched methods in practical experiments. Besides, data sparsity was not taken into consideration in their work. Furthermore, no theoretical guarantee was provided for NMF with dual random projection. In short, SANLS is not same as [42] and DSANLS is much more than a distributed version of [42]. The methods that we propose in this paper are efficient in practice and have strong theoretical guarantees.

3 DSANLS: Distributed Sketched ANLS

As opposed to ANLS-based methods, which compute an optimal solution for each NLS subproblem (as mentioned in Section 2.1), our *Distributed Sketched ANLS* approach reduces the size of each NLS subproblem using matrix sketching and solves it approximately. We now present the details of our method.

3.1 Notations

For a matrix A, we use $A_{i:j}$ to denote the entry at the i-th row and j-th column of A. Besides, either i or j can be omitted to denote a column or a row, i.e., $A_{i:}$ is the i-th row of A, and $A_{:j}$ is its j-th column. Furthermore, i or j can be replaced by a subset of indices. For example, if $I \subset \{1, 2, \ldots, m\}$, $A_{I:}$ denotes the sub-matrix of A formed by all rows in I, whereas $A_{:J}$ is the sub-matrix of A formed by all columns in a subset $J \subset \{1, 2, \ldots, n\}$.

3.2 Data Partitioning

Assume there are N computing nodes in the cluster. We partition the row indices $\{1, 2, \ldots, m\}$ of the input matrix M into N disjoint sets $I_1, I_2, \ldots, I_N$, where $I_r \subset \{1, 2, \ldots, m\}$ is the subset of rows assigned to node r, as in [27]. Similarly, we partition the column indices $\{1, 2, \ldots, n\}$ into disjoint sets $J_1, J_2, \ldots, J_N$ and assign column set J_r to node r. The number of rows and columns in each node are near the same in order to achieve load balancing, i.e., $|I_r| \approx m/N$ and $|J_r| \approx n/N$ for each node r. The factor matrices U and V are also assigned to nodes accordingly, i.e., node r stores and updates $U_{I_r:}$ and $V_{J_r:}$ as shown in Figure 1.

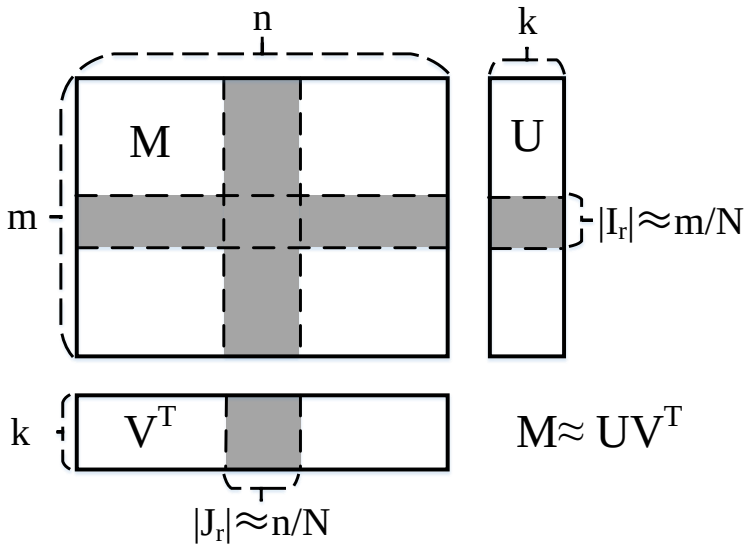


Figure 1: Data partitioning to N nodes

Data partitioning in distributed NMF differs from that in parallel NMF. Previous works on parallel NMF [15, 39] choose to partition U and V along the long dimension, but we adopt the row-partitioning of U and V as in [27]. To see why, take the U-subproblem (3) as an example and observe that it is row-independent in nature, i.e., the r-th row block of its solution $U_{I_r:}$ is given by

$$U_{I_r:} = \underset{U_{I_r:} \in \mathbb{R}_+^{|I_r| \times k}}{\arg\min} \left\| M_{I_r:} - U_{I_r:} V^\top \right\|_F^2 \tag{7}$$

and thus can be solved independently without referring to any other row blocks of U. The same holds for the V-subproblem. In addition, no communication is needed concerning M when solving (7) because $M_{I_r:}$ is already present in node r.

On the other hand, solving (7) requires the entire V of size $n \times k$, meaning that every node needs to gather V from all other nodes. This process can easily be the bottleneck of a naive distributed ANLS implementation. As we will explain shortly, our DSALNS algorithm alleviates this problem, since we use a sketched matrix of reduced size instead of the original complete matrix V.

3.3 SANLS: Sketched ANLS

To better understand DSANLS, we first introduce the Sketched ANLS (SANLS), i.e., a centralized version of our algorithm. Recall that, at each step of ANLS, either U or V is fixed and we solve a nonnegative least square problem (3) over the other variable. Intuitively, it is unnecessary to solve this subproblem with high accuracy, because we may not have reached the optimal solution for the fixed variable so far. Hence, when the fixed variable changes in the next step, its accurate solution from the previous step will not be optimal anymore and will have to be re-computed. Our idea is to apply matrix sketching for each subproblem, in order to obtain an approximate solution for it at a much lower computational and communication cost.

Specifically, suppose we are at the t-th iteration of ANLS, and our current estimations for U and V are U^t and V^t respectively. We must solve subproblem (3) in order to update U^t to a new matrix U^{t+1}. We apply matrix sketching to the residual term of subproblem (3). The subproblem now becomes:

$$\min_{U \in \mathbb{R}_+^{m \times k}} \left\| MS^t - U\left(V^{t\top} S^t\right) \right\|_F^2, \tag{8}$$

where $S^t \in \mathbb{R}^{n \times d}$ is a randomly-generated matrix. Hence, the problem size decreases from $n \times k$ to $d \times k$. d is chosen to be much smaller than n, in order to sufficiently reduce the computational cost[3]. Similarly, we transform the V-subproblem into

$$\min_{V \in \mathbb{R}_+^{n \times k}} \left\| M^\top S'^t - V\left(U^{t\top} S'^t\right) \right\|_F^2, \tag{9}$$

where $S'^t \in \mathbb{R}^{m \times d'}$ is also a random matrix with $d' \ll m$.

3.4 DSANLS: Distributed SANLS

Now, we come to our proposal: the distributed version of SANLS called DSANLS. Since the U-subproblem (8) is the same as the V-subproblem (9) in nature, here we restrict our attention to the U-subproblem. The first observation about subproblem (8) is that it is still row-independent, thus node r only needs to solve

$$\min_{U_{I_r:} \in \mathbb{R}_+^{|I_r| \times k}} \left\| \left(MS^t\right)_{I_r:} - U_{I_r:}\left(V^{t\top} S^t\right) \right\|_F^2.$$

For simplicity, we denote

$$A_r^t \triangleq \left(MS^t\right)_{I_r:} \quad \text{and} \quad B^t \triangleq V^{t\top} S^t, \tag{10}$$

[3]However, we should not choose an extremely small d, otherwise the the size of sketched subproblem would become so small that it can hardly represent the original subproblem, preventing NMF from converging to a good result. In practice, we can set $d = 0.1n$ for medium-sized matrices and $d = 0.01n$ for large matrices if $m \approx n$. When m and n differ a lot, e.g., $m \ll n$ without loss of generality, we should not apply sketching technique to the V subproblem (since solving the U subproblem is much more expensive) and simply choose $d = m \ll n$.

and the above subproblem can be written as:

$$\min_{U_{I_r:}\in\mathbb{R}_+^{|I_r|\times k}} \left\|A_r^t - U_{I_r:}B^t\right\|_F^2. \tag{11}$$

Thus, node r needs to know matrices A_r^t and B^t in order to solve the subproblem.

For A_r^t, by applying matrix multiplication rules, we get

$$A_r^t = \left(MS^t\right)_{I_r:} = M_{I_r:}S^t$$

Therefore, if S^t is stored at node r, A_r^t can be computed without any communication.

On the other hand, computing $B^t = \left(V^{t\top}S^t\right)$ requires communication across the whole cluster, since the rows of V^t are distributed across different nodes. Fortunately, if we assume that S^t is stored at all nodes again, we can compute B^t in a much cheaper way. Following block matrix multiplication rules, we can rewrite B^t as:

$$\begin{aligned} B^t &= V^{t\top}S^t \\ &= \left[\left(V_{J_1:}^t\right)^\top \ \cdots \ \left(V_{J_N:}^t\right)^\top\right] \begin{bmatrix} S_{J_1:}^t \\ \vdots \\ S_{J_N:}^t \end{bmatrix} \\ &= \sum_{r=1}^{N} \left(V_{J_r:}^t\right)^\top S_{J_r:}^t. \end{aligned}$$

Note that the summand $\bar{B}_r^t \triangleq \left(V_{J_r:}^t\right)^\top S_{J_r:}^t$ is a matrix of size $k \times d$ and can be computed locally. As a result, communication is only needed for summing up the matrices $\bar{B}_r^t$ of size $k \times d$ by using MPI all-reduce operation, which is much cheaper than transmitting the whole V_t of size $n \times k$.

Now, the only remaining problem is the transmission of S^t. Since S^t can be dense, even larger than V^t, broadcasting it across the whole cluster can be quite expensive. However, it turns out that we can avoid this. Recall that S^t is a randomly-generated matrix; each node can generate exactly the same matrix, if we use the same pseudo-random generator and the same seed. Therefore, we only need to broadcast the random seed, which is just an integer, at the beginning of the whole program. This ensures that each node generates exactly the same random number sequence and hence the same random matrices S^t at each iteration.

In short, the communication cost of each node is reduced from $\mathcal{O}(nk)$ to $\mathcal{O}(dk)$ by adopting our sketching technique for the U-subproblem. Likewise, the communication cost of each V-subproblem is decreased from $\mathcal{O}(mk)$ to $\mathcal{O}(d'k)$. The general framework of our DSANLS algorithm is listed in Algorithm 2.

3.5 Generation of Random Matrices

A key problem in Algorithm 2 is how to generate random matrices $S^t \in \mathbb{R}^{n\times d}$ and $S'^t \in \mathbb{R}^{m\times d'}$. Here we focus on generating a random $S^t \in \mathbb{R}^{d\times n}$ satisfying Assumption 1. The reason for choosing such a random matrix is that the corresponding sketched problem would be equivalent to the original problem on expectation; we will prove this in Section 3.6.

Algorithm 2: Distributed SANLS on Node r

Initialize $U^0_{I_r:}$, $V^0_{J_r:}$
Broadcast the random seed
for $t = 0$ ***to*** $T-1$ **do**
  Generate random matrix $S^t \in \mathbb{R}^{n\times d}$
  Compute $A^t_r \leftarrow M_{I_r:}S^t$
  Compute $\bar{B}^t_r \leftarrow \left(V^t_{J_r:}\right)^\top S^t_{J_r:}$
  All-Reduce: $B^t \leftarrow \sum_{i=1}^N \bar{B}^t_i$
  Update $U^{t+1}_{I_r:}$ by solving $\min_{U_{I_r:}} \|A^t_r - U_{I_r:}B^t\|$
  Generate random matrix $S'^t \in \mathbb{R}^{m\times d'}$
  Compute $A'^t_r \leftarrow \left(M_{:J_r}\right)^\top S'^t$
  Compute $\bar{B}'^t_r \leftarrow \left(U^t_{I_r:}\right)^\top S'^t_{I_r:}$
  All-Reduce: $B'^t \leftarrow \sum_{i=1}^N \bar{B}'^t_i$
  Update $V^{t+1}_{J_r:}$ by solving $\min_{V_{J_r:}} \|A'^t_r - V_{J_r:}B'^t\|$
end
return $U^T_{I_r:}$ *and* $V^T_{J_r:}$

Assumption 1. *Assume the random matrices are normalized and have bounded variance, i.e., there exists a constant σ^2 such that*

$$\mathbb{E}\left[S^tS^{t\top}\right] = I \quad \textit{and} \quad \mathbb{V}\left[S^tS^{t\top}\right] \le \sigma^2$$

for all t, where I is the identity matrix.

Different options exist for such matrices, which have different computation costs in forming sketched matrices $A^t_r = M_{I_r:}S^t$ and $\bar{B}^t_r = \left(V^t_{J_r:}\right)^\top S^t_{J_r:}$. Since $M_{I_r:}$ is much larger than $V^t_{J_r:}$ and thus computing A^t_r is more expensive, we only consider the cost of constructing A^t_r here.

The most classical choice for a random matrix is one with i.i.d. Gaussian entries having mean 0 and variance $1/d$. We can show that:

$$\mathbb{E}\left[\left(S^tS^{t\top}\right)_{i:j}\right] = \mathbb{E}\left[\sum_{l=1}^d S_{i:l}S_{j:l}\right] = \sum_{l=1}^d \mathbb{E}\left[S_{i:l}S_{j:l}\right] = \begin{cases} d\times\frac{1}{d} = 1, & \text{if } i=j,\\ d\times 0 = 0, & \text{otherwise}\end{cases}$$

which means that $\mathbb{E}\left[S^tS^{t\top}\right] = I$. Besides, Gaussian random matrix has bounded variance because Gaussian distribution has finite fourth-order moment. However, since each entry of such a matrix is totally random and thus no special structure exists in S^t, matrix multiplication will be expensive. That is, when given $M_{I_r:}$ of size $|I_r|\times n$, computing its sketched matrix $A^t_r = M_{I_r:}S^t$ requires $\mathcal{O}(|I_r|nd)$ basic operations.

A seemingly better choice for S^t would be a *subsampling* random matrix. Each column of such random matrix is uniformly sampled from $\{e_1, e_2, \ldots, e_n\}$ without replacement, where $e_i \in \mathbb{R}^n$ is the i-th canonical basis vector (i.e., a vector having its i-th element 1 and all others 0). We can easily show that such an S^t also satisfies $\mathbb{E}\left[S^tS^{t\top}\right] = I$ and the variance $\mathbb{V}\left[S^tS^{t\top}\right]$ is bounded, but this time constructing the sketched matrix $A^t_r = M_{I_r:}S^t$ only requires $\mathcal{O}\left(|I_r|d\right)$. Hence, a subsampling random matrix would be favored over a Gaussian random matrix by most applications, especially for very large-scale problems. On the other hand, we observed in our experiments that a Gaussian

random matrix can result in a faster per-iteration convergence rate, because each column of the sketched matrix A_r^t contains entries from multiple columns of the original matrix and thus is more informative. Hence, it would be better to use a Gaussian matrix when the sketch size d is small and thus a $\mathcal{O}(|I_r|nd)$ complexity is acceptable, or when the network speed of the cluster is poor, hence we should trade more local computation cost for less communication cost.

Although we only test two representative types of random matrices (i.e., Gaussian and subsampling random matrices), our framework is readily applicable for other choices, such as subsampled randomized Hadamard transform (SRHT) [1, 28] and count sketch [3, 5, 35]. The choice of random matrices is not the focus of this paper and left for future investigation.

3.6 Solving Subproblems

Before describing how to solve subproblem (11), let us make an important observation. As discussed in Section 2.4, the sketching technique has been applied in solving linear systems before. For a linear system, it is usually assumed that there exists a solution x^* such that $Ax^* = b$ exactly holds. Hence, x^* is also the solution to the sketched problem, namely, $(SA)x^* = Sb$. However, the situation is different in matrix factorization. Note that for the distributed matrix factorization problem we usually have

$$\min_{U_{I_r:}\in\mathbb{R}_+^{|I_r|\times k}} \left\| M_{I_r:} - U_{I_r:}V^{t\top} \right\|_F^2 \neq 0.$$

So, for the sketched subproblem (11), which can be equivalently written as

$$\min_{U_{I_r:}\in\mathbb{R}_+^{|I_r|\times k}} \left\| \left(M_{I_r:} - U_{I_r:}V^{t\top} \right) S^t \right\|_F^2,$$

the non-zero entries of the residual matrix $\left(M_{I_r:} - U_{I_r:}V^{t\top}\right)$ will be scaled by the matrix S^t at different levels. As a consequence, the optimal solution will be shifted because of sketching. This fact alerts us that for SANLS, we need to update U^{t+1} by exploiting the sketched subproblem (11) to step towards the true optimal solution and avoid convergence to the solution of the sketched subproblem.

3.6.1 Projected Gradient Descent

A natural method is to use *one step* of projected gradient descent for the sketched subproblem:

$$\begin{aligned} U_{I_r:}^{t+1} &= \max\left\{ U_{I_r:}^t - \eta_t \nabla_{U_{I_r:}} \left\| A_r^t - U_{I_r:}B^t \right\|_F^2 \Big|_{U_{I_r:}=U_{I_r:}^t},\ 0 \right\} \\ &= \max\left\{ U_{I_r:}^t - 2\eta_t \left[U_{I_r:}^t B^t B^{t\top} - A_r^t B^{t\top} \right],\ 0 \right\}, \end{aligned} \tag{12}$$

where $\eta_t > 0$ is the step size and $\max\{\cdot,\cdot\}$ denotes the entry-wise maximum operation.

To exploit the nature of this algorithm, we further expand the gradient:

$$\begin{aligned} \nabla_{U_{I_r:}} \left\| A_r^t - U_{I_r:}B^t \right\|_F^2 &= 2\left[U_{I_r:}B^tB^{t\top} - A_r^tB^{t\top} \right] \\ \overset{(10)}{=} 2\left[U_{I_r:}\left(V^{t\top}S^t\right)\left(V^{t\top}S^t\right)^\top - \left(M_{I_r:}S^t\right)\left(V^{t\top}S^t\right)^\top \right] \\ =2\left[U_{I_r:}V^{t\top}\left(S^tS^{t\top}\right)V^t - M_{I_r:}\left(S^tS^{t\top}\right)V^t \right]. \end{aligned}$$

By taking the expectation of the above equation, and using the fact $\mathbb{E}\left[S^t S^{t\top}\right] = I$, we have:

$$\mathbb{E}\left[\nabla_{U_{I_r:}} \left\|A_r^t - U_{I_r:}B^t\right\|_F^2\right] = 2\left[U_{I_r:}V^{t\top}V^t - M_{I_r:}V^t\right] = \nabla_{U_{I_r:}} \left\|M_{I_r:} - U_{I_r:}V^{t\top}\right\|_F^2,$$

which means that the gradient of the sketched subproblem is equivalent to the gradient of the original problem on expectation[4]. Therefore, such a step of gradient descent can be interpreted as a (generalized) *stochastic gradient descent* (SGD) [32] method on the original subproblem. Thus, according to the theory of SGD, we naturally require the step sizes $\{\eta_t\}$ to be diminishing, i.e., $\eta_t \to 0$ as t increases.

In the gradient descent step (12), the computational cost mainly comes from two matrix multiplications: $B^t B^{t\top}$ and $A_{t,r}B^{t\top}$. Note that A_r^t and B^t are of sizes $|I_r| \times d$ and $k \times d$ respectively, thus the gradient descent step takes $\mathcal{O}\left(kd(|I_r| + k)\right)$ in total.

3.6.2 Regularized Coordinate Descent

However, it is well known that the gradient descent method converges slowly when solving NMF subproblems, while the coordinate descent method, namely the HALS method for NMF, is quite efficient [8]. Still, because of its very fast convergence, HALS should not be applied to the sketched subproblem, since it converges to the optimal solution of the subproblem after just a few iterations. As we have discussed in the beginning of Section 3.6, this is undesirable because it shifts the solution away from the true optimal solution. Therefore, we need to develop a method which resembles HALS but will not converge towards the solutions of the sketched subproblems.

To achieve this, we add a regularization term to the sketched subproblem (11). The new subproblem is:

$$\min_{U_{I_r:} \in \mathbb{R}_+^{|I_r| \times k}} \left\|A_r^t - U_{I_r:}B^t\right\|_F^2 + \mu_t \left\|U_{I_r:} - U_{I_r:}^t\right\|_F^2, \tag{13}$$

where $\mu_t > 0$ is a parameter. This regularization is reminiscent to the proximal point method [40] in optimization. Although the objective function is changed, the real effect of the regularization term is to control the step size, and thus it does not change the solution to which the algorithm ultimately converges. Therefore, parameter μ_t plays a role similar to $1/\eta_t$ in projected gradient descent; we require $\mu_t \to +\infty$ to enforce the convergence of the whole algorithm, e.g., $\mu_t = t$.

To solve (13) efficiently, we apply coordinate descent. At each step, only one column of $U_{I_r:}$, say $U_{I_r,j}$ where $j \in \{1, 2, \ldots, k\}$, is updated:

$$\min_{U_{I_r:j} \in \mathbb{R}_+^{|I_r|}} \left\|A_r^t - U_{I_r:j}B_{j:}^t - \sum_{l \neq j} U_{I_r:l}B_{l:}^t\right\|_F^2 + \mu_t \left\|U_{I_r:j} - U_{I_r:j}^t\right\|_2^2.$$

It is not hard to see that the above problem is still row-independent, which means that each entry of the row vector $U_{I_r:j}$ can be solved independently. For example, for any $i \in I_r$, the solution of $U_{i:j}^{t+1}$ is given by:

[4]We generalize such property as Lemma 2, shown in Appendix B.1.

$$U_{i:j}^{t+1} = \arg\min_{U_{i:j}\geq 0} \left\| \left(A_r^t\right)_{i:} - U_{i:j}B_{j:}^t - \sum_{l\neq j} U_{i:l}B_{l:}^t \right\|_2^2 + \mu_t \left\| U_{i:j} - U_{i:j}^t \right\|_2^2$$
$$= \max\left\{ \frac{\mu_t U_{i:j}^t + \left(A_r^t\right)_{i:} B_{j:}^{t\top} - \sum_{l\neq j} U_{i:l}B_{l:}^t B_{j:}^{t\top}}{B_{j:}^t B_{j:}^{t\top} + \mu_t}, 0 \right\}. \tag{14}$$

At each step of coordinate descent, we choose the column j from $\{1, 2, \ldots, k\}$ successively. When updating column j at iteration t, the columns $l < j$ have already been updated and thus $U_{I_r:l} = U_{I_r:l}^{t+1}$, while the columns $l > j$ are old so $U_{I_r:l} = U_{I_r:l}^t$. Based on this, (14) can be equivalently written in vector form as:

$$U_{I_r:j}^{t+1} = \max\left\{ \frac{\mu_t U_{I_r:j}^t + A_r^t B_{j:}^{t\top} - \sum_{l=1}^{j-1} B_{l:}^t B_{j:}^{t\top} U_{I_r:l}^{t+1} - \sum_{l=j+1}^{k} B_{l:}^t B_{j:}^{t\top} U_{I_r:l}^t}{B_{j:}^t B_{j:}^{t\top} + \mu_t}, 0 \right\}.$$

The complete coordinate descent algorithm for the U-subproblem is summarized in Algorithm 3. When updating column j, computing the matrix-vector multiplication $A_r^t B_{j:}^{t\top}$ takes $\mathcal{O}(d|I_r|)$. The whole inner loop takes $\mathcal{O}\left(k\left(d + |I_r|\right)\right)$ because one vector dot product of length d is required for computing each summand and the summation itself needs $\mathcal{O}\left(k|I_r|\right)$. Considering that there are k columns in total, the overall complexity of coordinate descent is $\mathcal{O}\left(k((k+d)\,|I_r| + kd)\right)$. Typically, we choose $d > k$, so the complexity can be simplified to $\mathcal{O}\left(kd\left(|I_r| + k\right)\right)$, which is the same as that of gradient descent.

Since we find that the regularized coordinate descent is much more efficient than projected gradient descent, we adopt it as the default subproblem solver within DSANLS.

Algorithm 3: Regularized Coordinate Descent for Local Subproblem (11) on Node r

Parameter: $\mu_t > 0$

for $j = 1$ ***to*** k **do**
  $T \leftarrow \mu_t U_{I_r:j}^t + A_r^t B_{j:}^{t\top}$
  for $l = 1$ ***to*** $j-1$ **do**
    $T \leftarrow T - \left(B_{l:}^t B_{j:}^{t\top}\right) U_{I_r:l}^{t+1}$
  end
  for $l = j+1$ ***to*** k **do**
    $T \leftarrow T - \left(B_{l:}^t B_{j:}^{t\top}\right) U_{I_r:l}^t$
  end
  $U_{I_r:j}^{t+1} \leftarrow \max\left\{T / \left(B_{j:}^t B_{j:}^{t\top} + \mu_t\right), 0\right\}$
end
return $U_{I_r:}^{t+1}$

4 Theoretical Analysis

Both complexity and convergence analyses of DSANLS are provided in this section.

4.1 Complexity Analysis

We now analyze the computational and communication costs of our DSANLS algorithm, when using subsampling random sketch matrices. The computational complexity at each node is:

$$\begin{aligned}&\mathcal{O}\big(\overbrace{d}^{\text{generating } S^t} + \overbrace{|I_r|d}^{\text{constructing } A_r^t \text{ and } B^t} + \overbrace{kd(|I_r|+k)}^{\text{solving subproblem}}\big)\\ &= \mathcal{O}\left(kd(|I_r|+k)\right) \approx \mathcal{O}\left(kd\left(\frac{m}{N}+k\right)\right)\end{aligned} \tag{15}$$

Moreover, as we have shown in Section 3.4, the communication cost of DSANLS is $\mathcal{O}(kd)$.

On the other hand, for a classical implementation of distributed HALS [7], the computational cost is

$$\mathcal{O}\left(kn\left(|I_r|+k\right)\right) \approx \mathcal{O}\left(kn\left(\frac{m}{N}+k\right)\right) \tag{16}$$

and the communication cost is $\mathcal{O}(kn)$ due to the all-gathering of V^t's.

Comparing the above quantities, we observe an $n/d \gg 1$ speedup of our DSANLS algorithm over HALS in both computation and communication. However, we empirically observed that DSANLS has a slower per-iteration convergence rate (i.e., it needs more iterations to converge). Still, as we will show in the next section, in practice, DSANLS is superior to alternative distributed NMF algorithms, after taking all factors into account.

4.2 Convergence Analysis

Here we provide theoretical convergence guarantees for the proposed SANLS and DSANLS algorithms. We show that SANLS and DSANLS converge to a stationary point.

4.2.1 Assumptions

To establish convergence result, Assumption 2 is needed first.

Assumption 2. *Assume all the iterates U^t and V^t have uniformly bounded norms, which means that there exists a constant R such that*

$$\|U^t\|_F \le R \quad \textit{and} \quad \|V^t\|_F \le R$$

for all t.

We experimentally observed that this assumption holds in practice, as long as the step sizes used are not too large. Besides, Assumption 2 can also be enforced by imposing additional constraints, e.g.:

$$U_{i:l} \le \sqrt{2\|M\|_F} \quad \text{and} \quad V_{j:l} \le \sqrt{2\|M\|_F} \quad \forall i,j,l, \tag{17}$$

with which we have $R = \max\{m,n\}k\sqrt{2\|M\|_F}$. Such constraints can be very easily handled by both of our projected gradient descent and regularized coordinate descent solvers. Lemma 1 shows that imposing such extra constraints does not prevent us from finding the global optimal solution. The lemma is proved in Appendix A.

Lemma 1. *If the optimal solution to the original problem* (1) *exists, there is at least one global optimal solution in the domain* (17).

4.2.2 Convergence Theorem

Based on Assumptions 1 (see Section 3.5) and 2, we now can formally show our main convergence result:

Theorem 1. *Under Assumptions 1 and 2, if the step sizes satisfy*

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty,$$

for projected gradient descent, or

$$\sum_{t=1}^{\infty} 1/\mu_t = \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} 1/\mu_t^2 < \infty,$$

for regularized coordinate descent, then SANLS and DSANLS with either sub-problem solver will converge to a stationary point of problem (1) *with probability 1.*

Here stationary point means a point whose projected gradient is zero. In convex optimization problems, stationary points must be globally optimal solutions. Although our problem is non-convex and hence its stationary points do not necessarily correspond to global optima, considering that it is NP-hard to find a global optimal solution for a general non-convex problems, convergence to a stationary point is already the best theoretical result that one can hope for. The proof of Theorem 1 can be found in Appendix B.

5 Experimental Evaluation

This section includes an experimental evaluation of our algorithm on both dense and sparse real data matrices.

5.1 Datasets

We use real public datasets corresponding to different NMF tasks in our evaluation. Their statistics are summarized in Table 1.

Table 1: Statistics of Datasets

Task	Dataset	# rows	# columns	Non-zero values	sparsity
Video analysis	BOATS	216,000	300	64,800,000	0%
Image processing	MIT CBCL FACE	2,429	361	876,869	0%
	MNIST	70,000	784	10,505,375	80.86%
	GISETTE	13,500	5,000	8,770,559	87.01%
Text mining	Reuters(RCV1)	804,414	47,236	60,915,113	99.84%
Community detection	DBLP Collaboration Network	317,080	317,080	2,416,812	99.9976%

Video Analysis. NMF can be used on video data for background subtraction (i.e., to detect moving objects) [19]. We here use BOATS[5] video dataset [2], which includes boats moving through

[5] `http://visal.cs.cityu.edu.hk/downloads/`

water. The video has 15 fps and it is saved as a sequence of png files, whose format is RGB with a frame size of 360×200. We use 'Boats2' which contains one boat close to the camera for 300 frames and reshape the matrix such that every RGB frame is a column of our matrix; the final matrix is dense with size $216,000 \times 300$.

Image Processing. The first dataset we use for this application is MIT CBCL FACE DATABASE[6] as in [22]. To form the vectorized matrix, whose size is 2429×361, we use all 2,429 face images (each with 19×19 pixels) in the original training set. The second dataset is MINST[7], which is a widely used handwritten digits dataset. We use all 70,000 samples including both training and test set, and form the vectorized matrix. The third dataset is GISETTE[8], from another handwritten digit recognition problem. This dataset is one of the five datasets used in the NIPS 2003 feature selection challenge. We use all pictures in the training, validation, and test datasets and form the vectorized matrix, whose size is $5,000 \times 13,500$.

Text Mining. We use the Reuters document corpora[9] as in [44]. Reuters Corpus Volume I (RCV1) [24] is an archive of over 800,000 manually categorized newswire stories made available by Reuters, Ltd. for research purposes. It has 804,414 samples and 47,236 features. Non-zero values contain cosine-normalized, log TF-IDF vectors. A nearly chronological split is also utilized, which follows the official LYRL2004 chronological split.

Community Detection. We use the DBLP collaboration network[10]. It is a co-authorship graph where two authors are connected if they have published at least one paper together. We convert it to an adjacency matrix, which has 2,416,812 non-zero values, taking 0.0024% of the whole matrix.

5.2 Setup

We conduct our experiments on the Linux cluster of our institute with a total of 96 nodes. Each node contains 8-core Intel(R) Core(TM) i7-3770 CPU @ 1.60GHz cores and 16 GB of memory. Our algorithm is implemented in C using the Intel Math Kernel Library (MKL) and Message Passing Interface (MPI). We use 10 nodes by default. Since tuning the factorization rank k is outside the scope of this paper, we use 100 as default value of k. Because of the large sizes of RCV1 and DBLP, we only use subsampling random matrices for them, as the use of Gaussian random matrices is too slow.

We evaluate DSANLS with subsampling and Gaussian random matrices, denoted by DSANLS/S and DSANLS/G, respectively, using regularized coordinate descent as the default subproblem solver. As mentioned in [16, 17], it is unfair to compare with a Hadoop implementation, as Hadoop is not designed for high performance computing of iterative numerical algorithms. Although Spark is more appropriate than Hadoop for iterative algorithms, [9] further shows that Spark implementations incur significant overheads due to task scheduling, task start delays, and idle time caused by Spark stragglers, and it is usually around 4x slower compared to MPI implementations. Therefore, we only compare DSANLS with MPI-FAUN[11] (all MPI-FAUN-MU, MPI-FAUN-HALS, and MPI-FAUN-

[6] `http://cbcl.mit.edu/software-datasets/FaceData2.html`

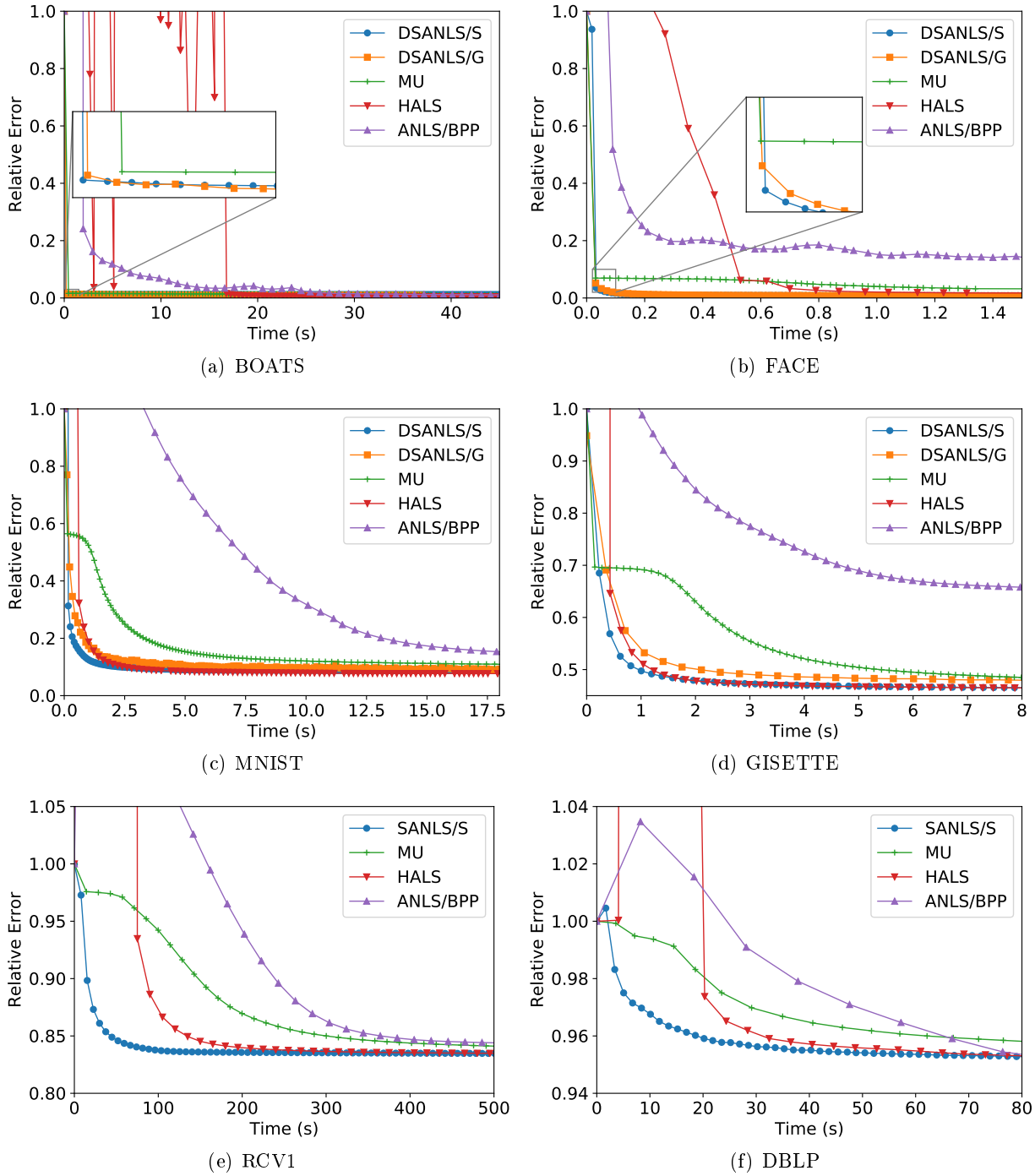
[7] `http://yann.lecun.com/exdb/mnist/`

[8] `http://clopinet.com/isabelle/Projects/NIPS2003/#challenge`

[9] we use the second version RCV1-v2, which can be found in `http://jmlr.csail.mit.edu/papers/volume5/lewis04a/`

[10] `http://snap.stanford.edu/data/com-DBLP.html`

[11] public code available at `https://github.com/ramkikannan/nmflibrary`



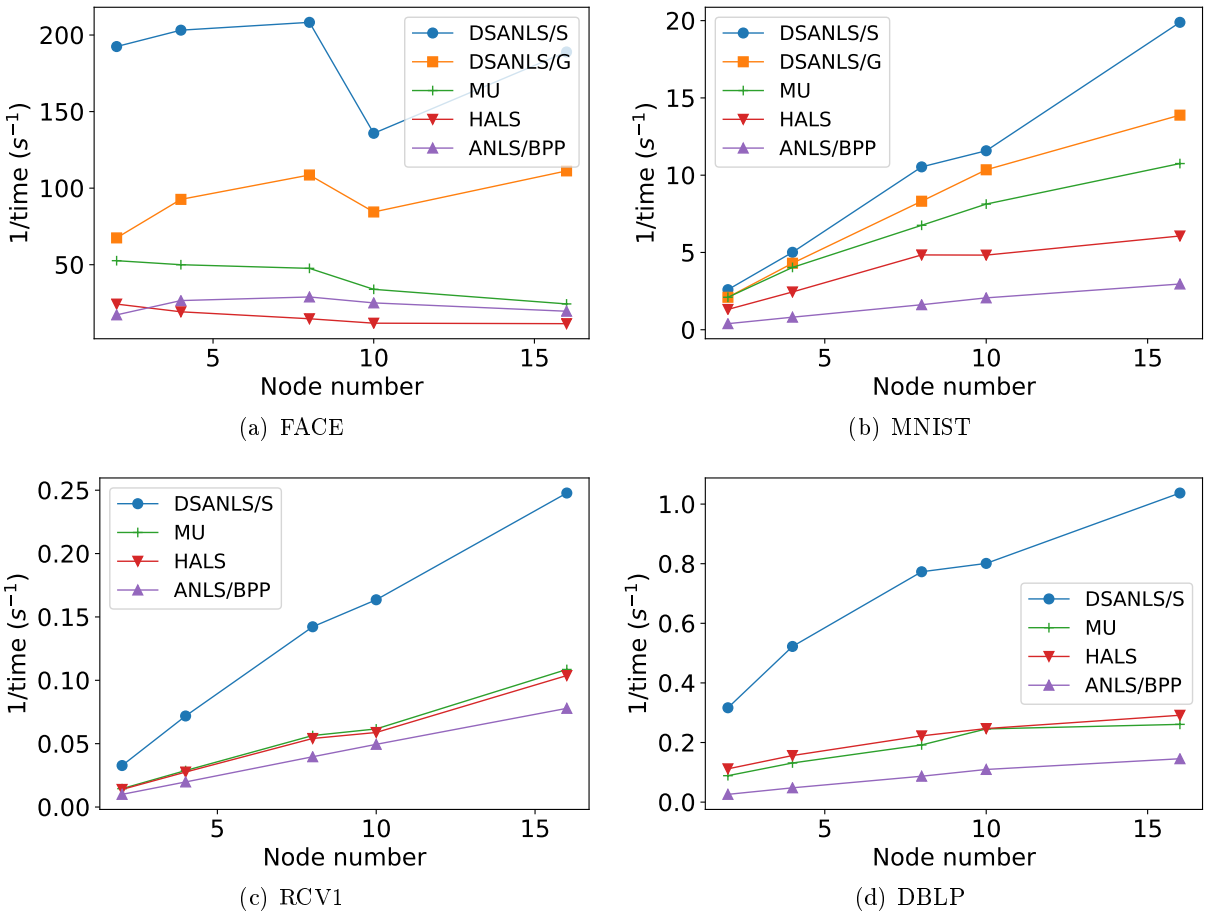
(a) BOATS (b) FACE

(c) MNIST (d) GISETTE

(e) RCV1 (f) DBLP

Figure 2: Relative error over time

ABPP implementations), which is the first and the state-of-the-art C++/MPI implementation with MKL and Armadillo. For parameters pc and pr in MPI-FAUN, we use the optimal values for each dataset, according to the recommendations in [16, 17].



(a) FACE (b) MNIST

(c) RCV1 (d) DBLP

Figure 3: Reciprocal of per-iteration time as a function of cluster size

5.3 Results

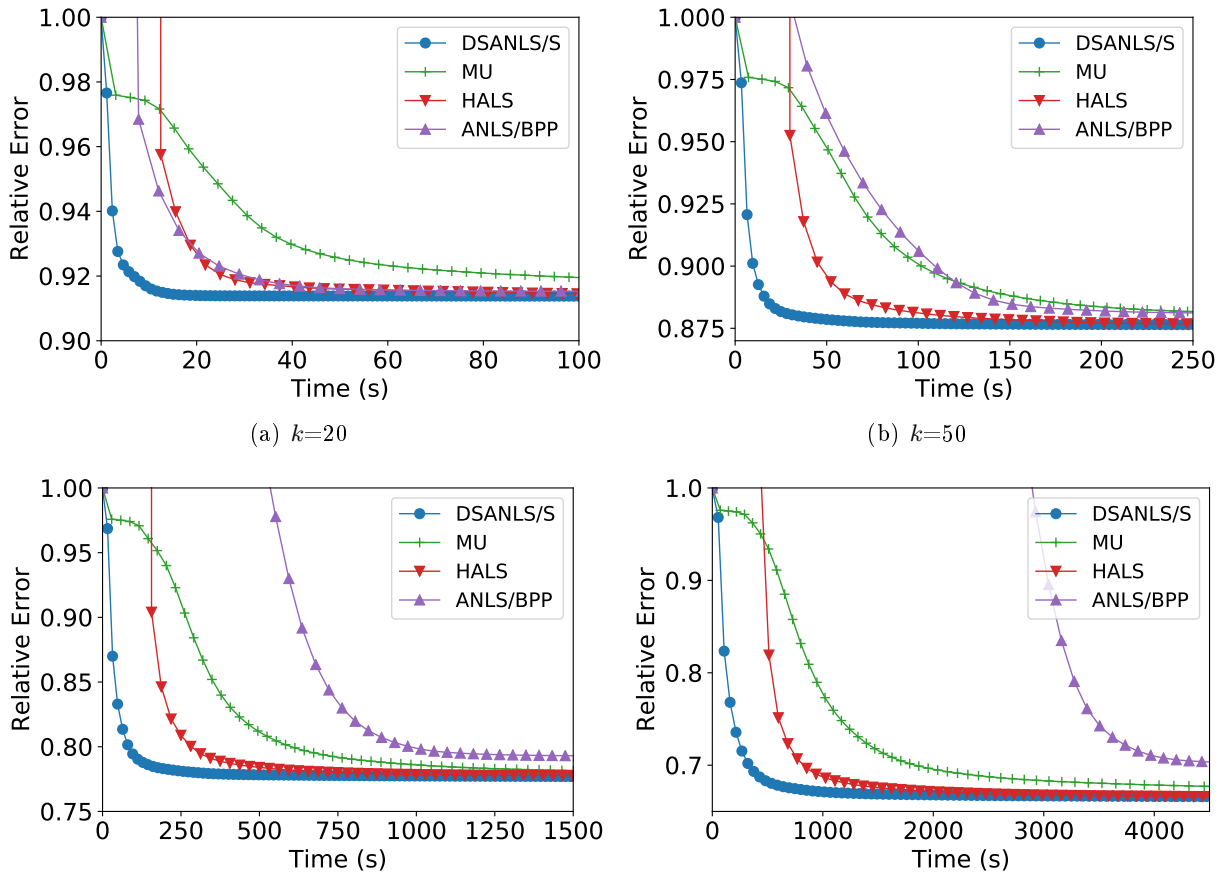
5.3.1 Performance Comparison

We use the relative error of the low rank approximation compared to the original matrix to measure the effectiveness of NMF by DSANLS with MPI-FAUN. This error measure was been widely used in previous work [16, 17, 19] and is formally defined as

$$\left\|M - UV^\top\right\|_F / \|M\|_F .$$

Since the time for each iteration is significantly reduced by our proposed DSANLS compared to MPI-FAUN, in Figure 2, we show the relative error over time for DSANLS and MPI-FAUN implementations of MU, HALS, and ANLS/BPP on the 6 real public datasets. Observe that DSANLS/S performs best in all 6 datasets, although DSANLS/G has faster per-iteration convergence rate. MU converges relatively slowly and usually has a bad convergence result; on the other hand HALS may oscillate in the early rounds[12], but converges quite fast and to a good solution. Surprisingly, al-

[12]HALS does not guarantee the objective function to decrease monotonically.



(a) k=20 (b) k=50

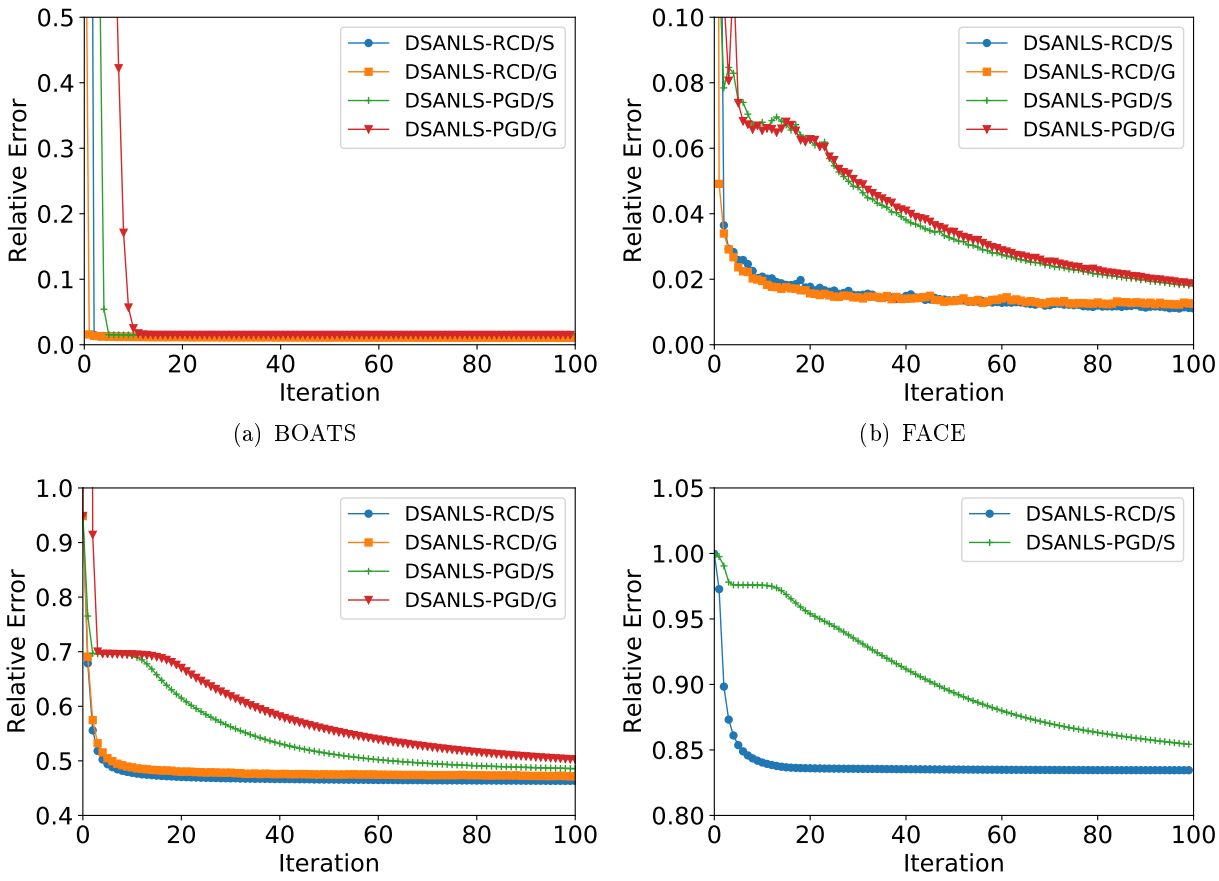
(c) k=200 (d) k=500

Figure 4: Relative error over time, varying k value

though ANLS/BPP is considered to be the state-of-art NMF algorithm, it does not perform well in all 6 datasets. As we will see, this is due to its high per-iteration cost.

5.3.2 Scalability Comparison

We vary the number of nodes used in the cluster from 2 to 16 and record the average time for 100 iterations of each algorithm. Figure 3 shows the reciprocal of per-iteration time as a function of the number of nodes used. All algorithms exhibit good scalability for all datasets (nearly a straight line), except for FACE (i.e., Figure 3(a)). FACE is the smallest dataset, whose number of columns is 300, while k is set to 100 by default. When n/N is smaller than k, the complexity is dominated by k, hence, increasing the number of nodes does not reduce the computational cost, but may increase the communication overhead. In general, we can observe that DSANLS/Subsampling has the lowest per-iteration cost compared to all other algorithms, and DSANLS/Gaussian has similar cost to MU and HALS. ANLS/BPP has the highest per-iteration cost, explaining the bad performance of ANLS/BPP in Figure 2.



(a) BOATS (b) FACE

(c) GISETTE (d) RCV1

Figure 5: Relative error per-iteration of different subproblem solvers

5.3.3 Performance Varying the Value of k

Although tuning the factorization rank k is outside the scope of this paper, we compare the performance of DSANLS with MPI-FAUN varying the value of k from 20 to 500 on RCV1. Observe from Figures 4 and 2(e) that DSANLS outperforms the state-of-art algorithms for all values of k. Naturally, the relative error of all algorithms decreases with the increase of k, but they also take longer to converge.

5.3.4 Comparison with Projected Gradient Descent

In Section 3.6, we claimed that our regularized coordinate descent approach (denoted as DSANLS-RCD) is faster than projected gradient descent (also presented in the same section, denoted as DSANLS-PGD). Figure 5 confirms the difference in the convergence rate of the two approaches regardless the choice of the random matrix generation approach.

6 Conclusion

In this paper, we presented a novel distributed NMF algorithm that can be used for scalable analytics of high dimensional matrix data. Our approach follows the general framework of ANLS, but utilizes matrix sketching to reduce the problem size of each NLS subproblem. We discussed and compared two different approaches for generating random matrices (i.e. Gaussian and subsampling random matrices). Then, we presented two subproblem solvers for our general framework, and theoretically proved their convergence. We analyzed the per-iteration computational and communication cost of our approach and its convergence, showing its superiority compared to the previous state-of-the-art. Our experiments on several real datasets show that our method converges fast to an accurate solution and scales well with the number of cluster nodes used. In the future, we plan to study the application of DSANLS to dense or sparse tensors.

References

[1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *STOC*, pages 557–563. ACM, 2006.

[2] A. B. Chan, V. Mahadevan, and N. Vasconcelos. Generalized stauffer–grimson background subtraction for dynamic scenes. *Machine Vision and Applications*, 22(5):751–766, 2011.

[3] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

[4] A. Cichocki and P. Anh-Huy. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[5] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC*, pages 81–90. ACM, 2013.

[6] M. E. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ect. *IEEE transactions on medical imaging*, 5(2):61–66, 1986.

[7] J. P. Fairbanks, R. Kannan, H. Park, and D. A. Bader. Behavioral clusters in dynamic graphs. *Parallel Computing*, 47:38–50, 2015.

[8] N. Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.

[9] A. Gittens, A. Devarakonda, E. Racah, M. Ringenburg, L. Gerhardt, J. Kottalam, J. Liu, K. Maschhoff, S. Canon, J. Chhugani, et al. Matrix factorization at scale: a comparison of scientific data analytics in spark and c+ mpi using three case studies. *arXiv preprint arXiv:1607.01335*, 2016.

[10] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[11] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear gauss–seidel method under convex constraints. *Operations research letters*, 26(3):127–136, 2000.

[12] D. Grove, J. Milthorpe, and O. Tardieu. Supporting array programming in x10. In *ARRAY*, page 38. ACM, 2014.

[13] N. Guan, D. Tao, Z. Luo, and B. Yuan. Nenmf: an optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

[14] K. Huang, N. D. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2014.

[15] K. Kanjani. Parallel non negative matrix factorization for document clustering. *CPSC-659 (Parallel and Distributed Numerical Algorithms) course. Texas A&M University, Tech. Rep*, 2007.

[16] R. Kannan, G. Ballard, and H. Park. A high-performance parallel algorithm for nonnegative matrix factorization. In *PPoPP*, page 9. ACM, 2016.

[17] R. Kannan, G. Ballard, and H. Park. Mpi-faun: An mpi-based framework for alternating-updating nonnegative matrix factorization. *arXiv preprint arXiv:1609.09154*, 2016.

[18] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM journal on matrix analysis and applications*, 30(2):713–730, 2008.

[19] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2014.

[20] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[21] I. Kotsia, S. Zafeiriou, and I. Pitas. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Transactions on Information Forensics and Security*, 2(3):588–595, 2007.

[22] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[23] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.

[24] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *JMLR*, 5(Apr):361–397, 2004.

[25] R. Liao, Y. Zhang, J. Guan, and S. Zhou. Cloudnmf: a mapreduce implementation of nonnegative matrix factorization for large-scale biological datasets. *Genomics, proteomics & bioinformatics*, 12(1):48–51, 2014.

[26] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[27] C. Liu, H.-c. Yang, J. Fan, L.-W. He, and Y.-M. Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *WWW*, pages 681–690. ACM, 2010.

[28] Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *NIPS*, pages 369–377, 2013.

[29] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *NIPS*, pages 2283–2291, 2013.

[30] E. Mejía-Roa, D. Tabas-Madrid, J. Setoain, C. García, F. Tirado, and A. Pascual-Montano. Nmf-mgpu: non-negative matrix factorization on multi-gpu systems. *BMC bioinformatics*, 16(1):1, 2015.

[31] X. Meng, J. Bradley, B. Yuvaz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, et al. Mllib: Machine learning in apache spark. *JMLR*, 17(34):1–7, 2016.

[32] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[33] J. Neveu. *Discrete-parameter martingales*, volume 10. Elsevier, 1975.

[34] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons. Text mining using non-negative matrix factorizations. In *SDM*, pages 452–456. SIAM, 2004.

[35] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *ACM SIGKDD*, pages 239–247. ACM, 2013.

[36] M. Pilanci and M. J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *JMLR*, pages 1–33, 2015.

[37] M. Pilanci and M. J. Wainwright. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

[38] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011.

[39] S. A. Robila and L. G. Maciak. A parallel unmixing algorithm for hyperspectral images. In *Optics East*, pages 63840F–63840F. International Society for Optics and Photonics, 2006.

[40] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[41] S. A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[42] F. Wang and P. Li. Efficient nonnegative matrix factorization with random projections. In *SDM*, pages 281–292. SIAM, 2010.

[43] Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *TKDE*, 25(6):1336–1353, 2013.

[44] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273. ACM, 2003.

[45] J. Yin, L. Gao, and Z. M. Zhang. Scalable nonnegative matrix factorization with block-wise updates. In *ECML-PKDD*, pages 337–352. Springer, 2014.

[46] R. Zdunek and A. Cichocki. Non-negative matrix factorization with quasi-newton optimization. In *ICAISC*, pages 870–879. Springer, 2006.

A Proof of Lemma 1

Proof of Lemma 1. Suppose (U^*, V^*) is the global optimal solution but fails to satisfy (17). If there exist indices i, j, l such that $U^*_{i:l} \cdot V^*_{j:l} > 2\|M\|_F$, then

$$\left\|M - U^* V^{*\top}\right\|_F^2 \geq \left(U^*_{i:l} \cdot V^*_{j:l} - M_{i:j}\right)^2 > (2\|M\|_F - \|M\|_F)^2 \geq \|M\|_F^2.$$

However, simply choosing $U = 0$ and $V = 0$ will yield a smaller error $\|M\|_F^2$, which contradicts the fact that (U^*, V^*) is optimal. Therefore, if we define $\alpha_l = \max_i U^*_{i:l}$ and $\beta_l = \max_j V^*_{j:l}$, we must have $\alpha_l \cdot \beta_l \leq 2\|M\|_F$ for each l. Now we construct a new solution $(\overline{U}, \overline{V})$ by:

$$\overline{U}_{i:l} = U^*_{i:l} \cdot \sqrt{\beta_l/\alpha_l} \quad \text{and} \quad \overline{V}_{j:l} = V^*_{j:l} \cdot \sqrt{\alpha_l/\beta_l}.$$

Note that

$$\overline{U}_{i:l} \leq \alpha_l \cdot \sqrt{\beta_l/\alpha_l} = \sqrt{\alpha_l \cdot \beta_l} \leq \sqrt{2\|M\|_F},$$
$$\overline{V}_{j:l} \leq \beta_l \cdot \sqrt{\alpha_l/\beta_l} = \sqrt{\alpha_l \cdot \beta_l} \leq \sqrt{2\|M\|_F},$$

so $(\overline{U}, \overline{V})$ satisfy (17). Besides,

$$\begin{aligned}
\left\|M - \overline{U}\,\overline{V}^\top\right\|_F^2 &= \sum_{i,j} \left(M_{i:j} - \sum_l \overline{U}_{i:l}\overline{V}_{j:l}\right)^2 \\
&= \sum_{i,j} \left(M_{i:j} - \sum_l U^*_{i:l} \cdot \sqrt{\beta_l/\alpha_l} \cdot V^*_{j:l} \cdot \sqrt{\alpha_l/\beta_l}\right)^2 \\
&= \sum_{i,j} \left(M_{i:j} - \sum_l U^*_{i:l} \cdot V^*_{j:l}\right)^2 = \|M - U^* V^{*\top}\|_F^2,
\end{aligned}$$

which means that $(\overline{U}, \overline{V})$ is also an optimal solution. In short, for any optimal solution of (1) outside the domain (17), there exists a corresponding global optimal solution satisfying (17), which further means that there exists at least one optimal solution in the domain (17). □

B Proof of the Main Theorem

For simplicity, we denote $f(U,V) = \|M - UV^\top\|_F^2$, $\tilde{f}_S = \|MS - U(V^\top S)\|_F^2$, and $\tilde{f}'_{S'} = \|M^\top S' - V(U^\top S')\|_F^2$. Let G^t and $\tilde{G}^t$ denote the gradients of the above quantities, i.e.,

$$G^t \triangleq \nabla_U f(U,V^t)\big|_{U=U^t}, \quad \tilde{G}^t \triangleq \nabla_U \tilde{f}_{S^t}(U,V^t)\Big|_{U=U^t},$$
$$G'^t \triangleq \nabla_V f(U^{t+1},V)\big|_{V=V^t}, \quad \tilde{G}'^t \triangleq \nabla_V \tilde{f}'_{S'^t}(U^{t+1},V)\Big|_{V=V^t}.$$

Besides, let

$$\Delta^t \triangleq \frac{1}{\eta_t}\left(U^t - U^{t+1}\right) \quad \text{and} \quad \Delta'^t \triangleq \frac{1}{\eta_t}\left(V^t - V^{t+1}\right).$$

B.1 Preliminary Lemmas

To prove Theorem 1, we need following lemmas (which are proved in Appendix B.3):

Lemma 2. *Under Assumption 1 and 2, conditioned on U^t and V^t, $\tilde{G}^t$ and $\tilde{G}'^t$ are unbiased estimators of G^t and G'^t respectively with uniformly bounded variance.*

Lemma 3. *Assume X is a nonnegative random variable with mean μ and variance σ^2, and $c \geq 0$ is a constant. Then we have*

$$\mathbb{E}\left[\min\{X,c\}\right] \geq \min\left\{c, \frac{\mu}{2}\right\} \cdot \left(1 - \frac{4\sigma^2}{4\sigma^2 + \mu^2}\right). \tag{18}$$

Lemma 4. *Define the function*

$$\phi(x,y,z) = \min\left\{|xy|, y^2/2\right\} \cdot \left(1 - \frac{4z^2}{4z^2 + y^2}\right) \geq 0. \tag{19}$$

Conditioned on U^t and V^t, there exists an uniform constant $\sigma'^2 > 0$ such that

$$\mathbb{E}[G^t_{i:l} \cdot \Delta^t_{i:l}] \geq \phi\left(U^t_{i:l}/\eta_t, G^t_{i:l}, \sigma'^2\right) \tag{20}$$

and

$$\mathbb{E}[G'^t_{j:l} \cdot \Delta'^t_{j:l}] \geq \phi\left(V^t_{j:l}/\eta_t, G'^t_{j:l}, \sigma'^2\right)$$

for any i, j, l.

Lemma 5 (Supermartingale Convergence Theorem, [33])**.** *Let Y_t, Z_t and W_t, $t = 0, 1, \ldots$, be three sequences of random variables and let $\mathcal{F}_t$, $t = 0, 1, \ldots$, be sets of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Suppose that*

1. *The random variables Y_t, Z_t and W_t are nonnegative, and are functions of the random variables in $\mathcal{F}_t$.*
2. *For each t, we have*
$$\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq Y_t - Z_t + W_t.$$
3. *There holds, with probability 1, $\sum_{t=0}^{\infty} W_t < \infty$.*

Then we have $\sum_{t=0}^{\infty} Z_t < \infty$, and the sequence Y_t converges to a nonnegative random variable Y, with probability 1.

Lemma 6 ([29]). *For two nonnegative scalar sequences $\{a_t\}$ and $\{b_t\}$, if $\sum_{t=0}^{\infty} a_k = \infty$ and $\sum_{t=0}^{\infty} a_t b_t < \infty$, then*

$$\liminf_{t\to\infty} b_t = 0.$$

Furthermore, if $|b_{t+1} - b_t| \le B \cdot a_t$ for some constant $B > 0$, then

$$\lim_{t\to\infty} b_t = 0.$$

B.2 Proof of Theorem 1

Proof of Theorem 1. Let us first focus on projected gradient descent. By conditioning on U^t and V^t, we have

$$\begin{aligned} f(U^{t+1}, V^t) =& \left\| M - U^{t+1} V^{t\top} \right\|_F^2 = \left\| M - \left(U^t - \eta_t \Delta^t\right) V^{t\top} \right\|_F^2 = \left\| \left(M - U^t V^{t\top}\right) - \eta_t \Delta^t V^{t\top} \right\|_F^2 \\ =& \left\| M - U^t V^{t\top} \right\|_F^2 - 2\eta_t \left(M - U^t V^{t\top}\right) \cdot \left(\Delta^t V^{t\top}\right) + \eta_t^2 \|\Delta^t V^{t\top}\|_F^2 \\ =& f(U^t, V^t) - 2\eta_t \left(M - U^t V^{t\top}\right) \cdot \left(\Delta^t V^{t\top}\right) + \eta_t^2 \|\Delta^t V^{t\top}\|_F^2. \end{aligned} \tag{21}$$

For the second term of (21), note that

$$2\left(M - U^t V^{t\top}\right) \cdot \left(\Delta^t V^{t\top}\right) = 2\mathrm{tr}\left[\left(M - U^t V^{t\top}\right) V^t \Delta^{t\top}\right] = \mathrm{tr}\left[G^t \Delta^{t\top}\right] = \sum_{i,l} G_{i:l}^t \cdot \Delta_{i:l}^t$$

By taking expectation and using Lemma 4, we obtain:

$$\mathbb{E}\left[2\left(M - U^t V^{t\top}\right) \cdot \left(\Delta^t V^{t\top}\right)\right] = \sum_{i,l} \mathbb{E}\left[G_{i:l}^t \cdot \Delta_{i:l}^t\right] \ge \sum_{i,l} \phi\left(U_{i:l}^t/\eta_t, G_{i:l}^t, \sigma'^2\right).$$

For simplicity, we will use the notation

$$\Phi(U^t/\eta_t, G^t) \triangleq \sum_{i,l} \phi\left(U_{i:l}^t/\eta_t, G_{i:l}^t, \sigma'^2\right).$$

For the third term of (21), we can bound it in the following way:

$$\begin{aligned} \|\Delta^t V^{t\top}\|_F^2 \le& \|\Delta^t\|_F^2 \cdot \|V^t\|_F^2 \le \|\tilde{G}^t\|_F^2 \cdot \|V^t\|_F^2 \\ =& \left\| 2(U^t V^{t\top} - M)(S^t S^{t\top}) V^t \right\|_F^2 \cdot \|V^t\|_F^2 \\ \le& 4\|M - U^t V^{t\top}\|_F^2 \cdot \|S^t S^{t\top}\|_F^2 \cdot \|V^t\|_F^4 \\ \le& 8\left(\|M\|_F^2 + \|U^t\|_F^2 \cdot \|V^t\|_F^2\right) \cdot \|S^t S^{t\top}\|_F^2 \cdot \|V^t\|_F^4 \\ \le& 8\left(\|M\|_F^2 + R^4\right) R^4 \cdot \|S^t S^{t\top}\|_F^2, \end{aligned}$$

where in the last inequality we have applied Assumption 2. If we take expectation, we have

$$\begin{aligned}\mathbb{E}\|\Delta^t V^{t\top}\|_F^2 \leq & 8\left(\|M\|_F^2+R^4\right)R^4\cdot\mathbb{E}\|S^tS^{t\top}\|_F^2\\ \leq & 8\left(\|M\|_F^2+R^4\right)R^4\cdot\left(\left\|\mathbb{E}[S^tS^{t\top}]\right\|^2+\mathbb{V}[S^tS^{t\top}]\right)\\ \leq & 8\left(\|M\|_F^2+R^4\right)R^4\cdot\left(n+\sigma^2\right),\end{aligned}$$

where mean-variance decomposition have been applied in the second inequality, and Assumption 1 was used in the last line. For convenience, we will use

$$\Gamma \triangleq 8\left(\|M\|_F^2+R^4\right)R^4\cdot\left(n+\sigma^2\right)\geq 0$$

to denote this constant later on.

By combining all results, we can rewrite (21) as

$$\mathbb{E}\left[f(U^{t+1},V^t)\right]\leq f(U^t,V^t)-\eta_t\Phi\left(U^t/\eta_t,G^t\right)+\eta_t^2\Gamma.$$

Likewise, conditioned on U^{t+1} and V^t, we can prove a similar inequality for V:

$$\mathbb{E}\left[f(U^{t+1},V^{t+1})\right]\leq f(U^{t+1},V^t)-\eta_t\Phi\left(V^t/\eta_t,G'^t\right)+\eta_t^2\Gamma',$$

where $\Gamma'\geq 0$ is also some uniform constant. From definition, it is easy to see both $\Phi\left(U^t/\eta_t,G^t\right)$ and $\Phi\left(V^t/\eta_t,G'^t\right)$ are nonnegative. Along with condition the condition $\sum_{t=0}^{\infty}\eta_t^2<\infty$, we can apply the Supermartingale Convergence Theorem (Lemma 5) with

$$\begin{aligned}Y_{2t}=f(U^t,V^t),\quad & Y_{2t+1}=f(U^{t+1},V^t),\\ Z_{2t}=\Phi\left(U^t/\eta_t,G^t\right),\quad & Z_{2t+1}=\Phi\left(V^t/\eta_t,G'^t\right),\\ W_{2t}=\Gamma\eta_t^2,\quad & W_{2t+1}=\Gamma'\eta_t^2,\end{aligned}$$

and then conclude that both $\{f(U^{t+1},V^t)\}$ and $\{f(U^t,V^t)\}$ will converge to a same value, and besides:

$$\sum_{t=0}^{\infty}\eta_t\left[\Phi\left(U^t/\eta_t,G^t\right)+\Phi\left(V^t/\eta_t,G'^t\right)\right]<\infty,$$

with probability 1. In addition, it is not hard to verify that $\left|\Phi\left(U^{t+1}/\eta_{t+1},G^{t+1}\right)-\Phi\left(U^t/\eta_t,G^t\right)\right|\leq C\cdot\eta_t$ for some constant C because of the boundness of the gradients. Then, by Lemma 6, we obtain that

$$\lim_{t\to\infty}\Phi\left(U^t/\eta_t,G^t\right)=\lim_{t\to\infty}\sum_{i:l}\phi\left(U_{i:l}^t/\eta_t,G_{i:l}^t,\sigma'^2\right)\to 0.$$

Since each summand in the above is nonnegative, this equation further implies

$$\lim_{t\to\infty}\phi\left(U_{i:l}^t/\eta_t,G_{i:l}^t,\sigma'^2\right)\to 0$$

for all i and l. By looking into the definition of ϕ in (19), it is not hard to see that $\phi\left(U_{i:l}^t/\eta_t,G_{i:l}^t,\sigma'^2\right)\to 0$ if and only if $\min\left\{U_{i:l}^t/\eta_t,\ \left|G_{i:l}^t\right|\right\}\to 0$. Considering $\eta_t>0$ and $\eta_t\to 0$, we can conclude that

$$\lim_{t\to\infty}\min\left\{U_{i:l}^t,\ \left|G_{i:l}^t\right|\right\}\to 0$$

for all i, l, which means either the gradient $G^t_{i:l}$ converges to 0, or $U^t_{i:l}$ converges to the boundary 0. In other words, the projected gradient at (U^t, V^t) w.r.t U converges to 0 as $t \to \infty$. Likewise, we can prove

$$\lim_{t\to\infty} \min\left\{V^t_{j:l},\, \left|G'^t_{j:l}\right|\right\} \to 0,$$

in a similar way, which completes the proof of projected gradient descent.

The proof of regularized coordinate descent is similar to that of projected gradient descent, and hence we only include a sketch proof here. The key here is to establish an inequality similar to (21), but with the difference that just one column rather than whole U or V is changed every time. Take $U_{:1}$ as an example. An important observation is that when projection does not happen, we can rewrite (14) as $U^{t+1}_{:1} = U^t_{:1} - \tilde{G}_{:1}/(\tau_t + B^t_{j:}B^{t\top}_{j:})$, which means that the moving direction of regularized coordinate descent is the same as that of projected gradient descent, but with step size being $1/(\tau_t + B^t_{j:}B^{t\top}_{j:})$. Since both the expectation and variance of $B^t_{j:}B^{t\top}_{j:}$ are bounded, we will have $1/(\tau_t + B^t_{j:}B^{t\top}_{j:}) \approx 1/\tau_t$ when τ_t is large. Given these two reasons, we can out down a similar inequality as (21). The remaining proof just follows the one for projected gradient descent. □

B.3 Proof of Preliminary Lemmas

Proof of Lemma 2. Since the proof related to $\tilde{G}'^t$ is similar to $\tilde{G}^t$, here we only focus on the latter one.

First, let us write down the definition of G^t and $\tilde{G}^t$:

$$G^t = 2(U^tV^{t\top} - M)V^t \quad \text{and} \quad \tilde{G}^t = 2(U^tV^{t\top} - M)(S^tS^{t\top})V^t.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\tilde{G}^t] =& \mathbb{E}\left[2(U^tV^{t\top} - M)(S^tS^{t\top})V^t\right] = 2(U^tV^{t\top} - M)\,\mathbb{E}[S^tS^{t\top}]\,V^t \\ =& 2(U^tV^{t\top} - M)\,I\,V^t = 2(U^tV^{t\top} - M)V^t = G^t,\end{aligned}$$

which means $\tilde{G}^t$ is an unbiased estimator of G^t. Besides, its variance is uniformly bounded because

$$\begin{aligned}\mathbb{V}[\tilde{G}^t] \le& \mathbb{V}\left[2(U^tV^{t\top} - M)_{i:}(S^tS^{t\top})V^t_{:l}\right] \\ \le& 4\|M - U^tV^{t\top}\|_F^2 \cdot \mathbb{V}[S^tS^{t\top}] \cdot \|V^t\|_F^2 \\ \le& 8\left(\|M\|_F^2 + \|U^t\|_F^2\|V^t\|_F^2\right) \cdot \|V^t\|_F^2 \cdot \mathbb{V}[S^tS^{t\top}] \\ \le& 8\left(\|M\|_F^2 + R^4\right) R^2 \cdot \sigma^2,\end{aligned}$$

where both Assumption 1 and Assumption 2 are applied in the last line. □

Proof of Lemma 3. In this proof, we will use Cantelli's inequality:

$$\Pr(X \ge \mu + \lambda) \ge 1 - \frac{\sigma^2}{\sigma^2 + \lambda^2} \quad \forall \lambda < 0.$$

When $\mu = 0$, it is easy to see that the right-hand-side of (18) is 0. Considering that the left-hand-side is the expectation of a nonnegative random variable, (18) obviously holds in this case.

When $\mu > 0$ and $\mu \ge 2c$, by using the fact that X is nonnegative, we have

$$\mathbb{E}\left[\min\{X, c\}\right] \ge c \cdot \Pr(X \ge c).$$

Now we can apply Cantelli's inequality to bound $\Pr(X \geq c)$ with $\lambda = c - \mu < c - \mu/2 \leq 0$, and obtain:

$$\mathbb{E}\left[\min\{X, c\}\right] \geq c \cdot \left(1 - \frac{\sigma^2}{\sigma^2 + (\mu - c)^2}\right) \geq c \cdot \left(1 - \frac{\sigma^2}{\sigma^2 + (\mu - \mu/2)^2}\right) = c \cdot \left(1 - \frac{4\sigma^2}{4\sigma^2 + \mu^2}\right), \tag{22}$$

where in the second inequality we used the fact $c \leq \mu/2$ again.

When $\mu > 0$ but $\mu < 2c$, we have:

$$\mathbb{E}\left[\min\{X, c\}\right] \geq \mathbb{E}\left[\min\{X, \mu/2\}\right].$$

Now we can apply inequality (22) from the previous part with $c = \mu/2$, and thus

$$\mathbb{E}\left[\min\{X, c\}\right] \geq \mathbb{E}\left[\min\{X, \mu/2\}\right] \geq \frac{\mu}{2} \cdot \left(1 - \frac{4\sigma^2}{4\sigma^2 + \mu^2}\right),$$

which completes the proof. □

Proof of Lemma 4. We only focus on G^t and Δ^t. We first show that

$$G^t_{i:l} \cdot \tilde{G}^t_{i:l} \geq 0 \tag{23}$$

for any random matrix S^t. Note that

$$G^t_{i:l} = 2(U^t V^{t\top} - M)_{i:} V^t_{:l} \quad \text{and} \quad \tilde{G}^t_{i:l} = 2(U^t V^{t\top} - M)_{i:} (S^t S^{t\top}) V^t_{:l}.$$

Hence it would be sufficient if we can show that there holds $a^\top (S^t S^{t\top}) b \cdot a^\top b \geq 0$ for any vectors a and b:

$$a^\top (S^t S^{t\top}) b \cdot a^\top b = \operatorname{tr}\left(a^\top (S^t S^{t\top}) b b^\top a\right) = \operatorname{tr}\left(a a^\top (S^t S^{t\top}) b b^\top\right) \geq 0,$$

where the first equality is because $A \cdot B = \operatorname{tr}(AB^\top)$, the second equality is due to cyclic permutation invariant property of trace, and the last inequality is because all of $aa^\top$, $bb^\top$ and $S^t S^{t\top}$ are positive semi-definite matrices.

Now, let us consider the relationship between Δ^t and $\tilde{G}^t$:

$$\Delta^t = \frac{1}{\eta_t}\left(U^t - U^{t+1}\right) = \frac{1}{\eta_t}\left(U^t - \max\left\{U^t - \eta_t \tilde{G}^t,\, 0\right\}\right),$$

from which it can be shown that

$$\Delta^t_{i:l} = \min\left\{U^t_{i:l}/\eta_t,\, \tilde{G}^t_{i:l}\right\}. \tag{24}$$

When $G^t_{i:l} = 0$, it is easy to see that both sides of (20) become 0, and hence (20) holds.

When $G^t_{i:l} > 0$, from (23) we know that $\tilde{G}^t_{i:l} \geq 0$ regardless of the choice of S^t. From Lemma 2 we know that

$$\mathbb{E}[\tilde{G}^t_{i:l}] = G^t_{i:l}$$

and there exists a constant $\sigma'^2 \geq 0$ such that

$$\mathbb{V}[\tilde{G}^t_{i:l}] \leq \sigma'^2.$$

Since $U_{i:l}^t$ is a nonnegative constant here, we can apply Lemma 3 to (24) and conclude

$$\begin{aligned}\mathbb{E}[\Delta_{i:l}^t] &\geq \min\left\{U_{i:l}^t/\eta_t, G_{i:l}^t/2\right\} \cdot \left(1 - \frac{4\mathbb{V}[\tilde{G}_{i:l}^t]}{4\mathbb{V}[\tilde{G}_{i:l}^t] + \left(G_{i:l}^t\right)^2}\right) \\ &\geq \min\left\{U_{i:l}^t/\eta_t, G_{i:l}^t/2\right\} \cdot \left(1 - \frac{4\sigma'^2}{4\sigma'^2 + \left(G_{i:l}^t\right)^2}\right),\end{aligned}$$

from which (20) is obvious.

When $G_{i:l}^t < 0$, also from (23) we know that $\tilde{G}_{i:l}^t \leq 0$. Since $U_{i:l}^t$ is a nonnegative constant here, we always have

$$\Delta_{i:l}^t = \min\left\{U_{i:l}^t/\eta_t, \tilde{G}_{i:l}^t\right\} = \tilde{G}_{i:l}^t.$$

Therefore, by taking expectation and using Lemma 2, we obtain

$$\mathbb{E}[\Delta_{i:l}^t] = \mathbb{E}[\tilde{G}_{i:l}^t] = G_{i:l}^t,$$

and thus

$$\mathbb{E}\left[G_{i:l}^t \cdot \Delta_{i:l}^t\right] = \left(G_{i:l}^t\right)^2 > \frac{\left(G_{i:l}^t\right)^2}{2} \cdot \left(1 - \frac{4\sigma'^2}{4\sigma'^2 + \left(G_{i:l}^t\right)^2}\right)$$

for any constant σ', which means that (20) holds. □