Alternating Mixing Stochastic Gradient Descent for Large-scale Matrix Factorization

Zhenhong Chen, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng
CAS Key Laboratory of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Introduction

Background

-- In the big data era, large scale matrix factorization (MF) has received much attention, e.g. recommender system.
-- Stochastic gradient descent (SGD) is one of the most popular algorithm to solve matrix factorization problem.
-- State-of-the-art distributed stochastic gradient descent methods: distributed SGD (DSGD), asynchronous SGD (ASGD), and iterative parameter mixing (IPM, also known as PSGD).

Motivation

-- IPM is elegant and easy to implement.
-- IPM outperforms DSGD and ASGD in many learning tasks such as learning conditional maximum entropy model and structured perceptron [1].
-- IPM was empirically shown to fails in matrix factorization [2]. Why the failure happens and how to get rid of it motivate this work.

Contributions

-- Theoretical analysis of the failure of IPM on MF.
-- Proposal of the alternating mixing SGD algorithm (AM-SGD).
-- Theoretical and empirical analysis of the proposed AM-SGD algorithm.

Failure of IPM on MF

MF formulation

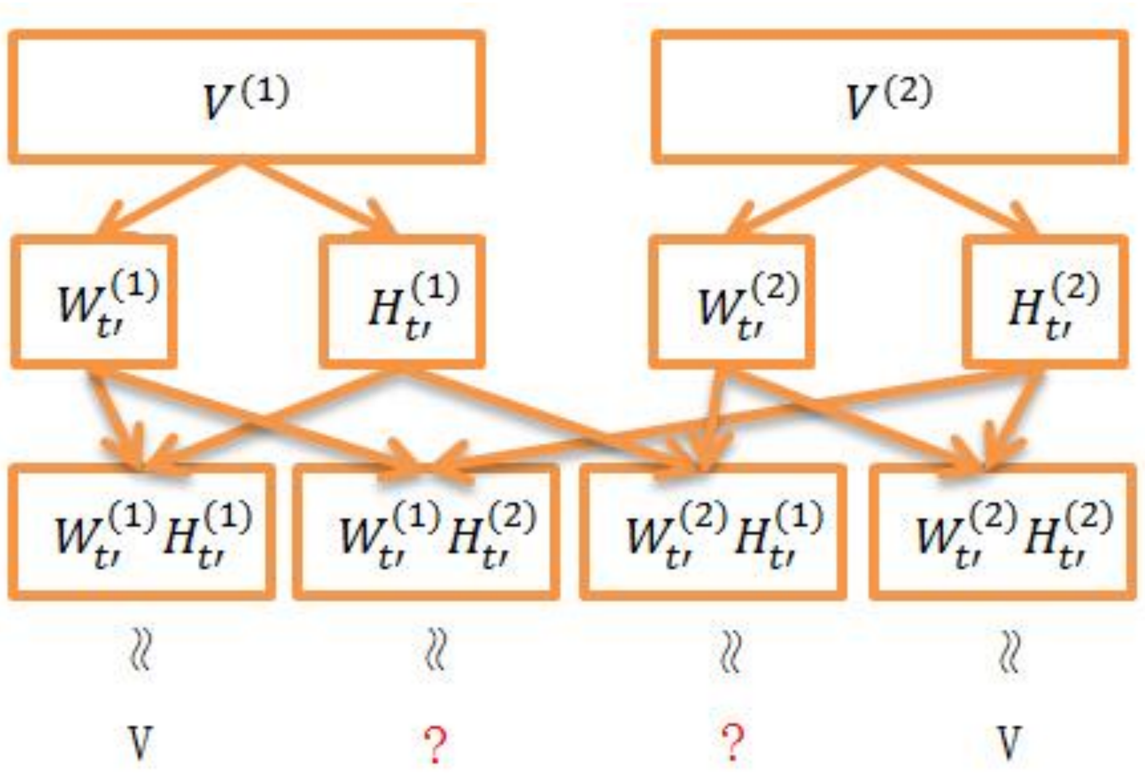
$$V \approx W \times H$$

IPM on MF

$$V = \cup V^{(i)}, \quad V^{(i)} \approx W_{t'}^{(i)} \times H_{t'}^{(i)}, \ \forall i = 1, \dots, d,$$

$$W_{t+1} = \frac{1}{d}\sum_1^d W_{t'}^{(i)}, \ H_{t+1} = \frac{1}{d}\sum_1^d H_{t'}^{(i)}$$

Failure Analysis

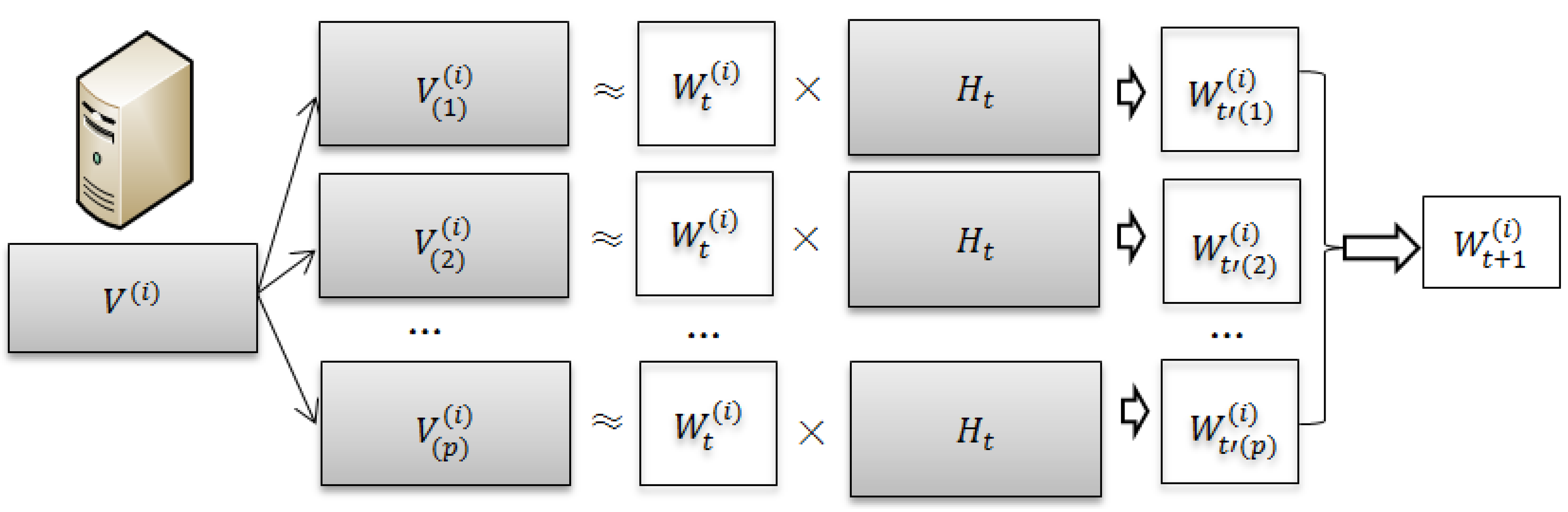


AM-SGD

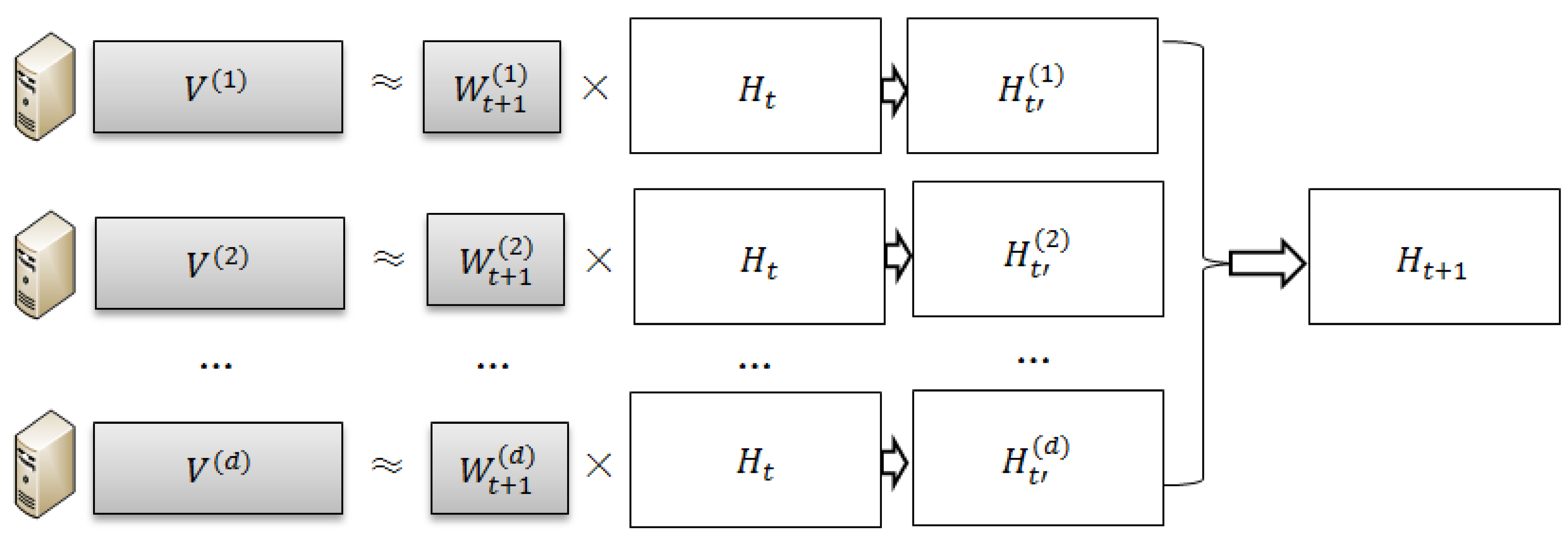
Data and Parameter partition

-- V and W are partitioned into $d \times 1$ blocks
-- each node c_i store $V^{(i)}, W^{(i)}$ and the whole H, $\forall i = 1, \dots, d$

Update $W_t^{(i)}$ with H_t fixed on node c_i, in parallel (with p threads)



Update H_t with $W_{t+1}^{(i)}$ fixed on node c_i, in parallel



Experimental Results

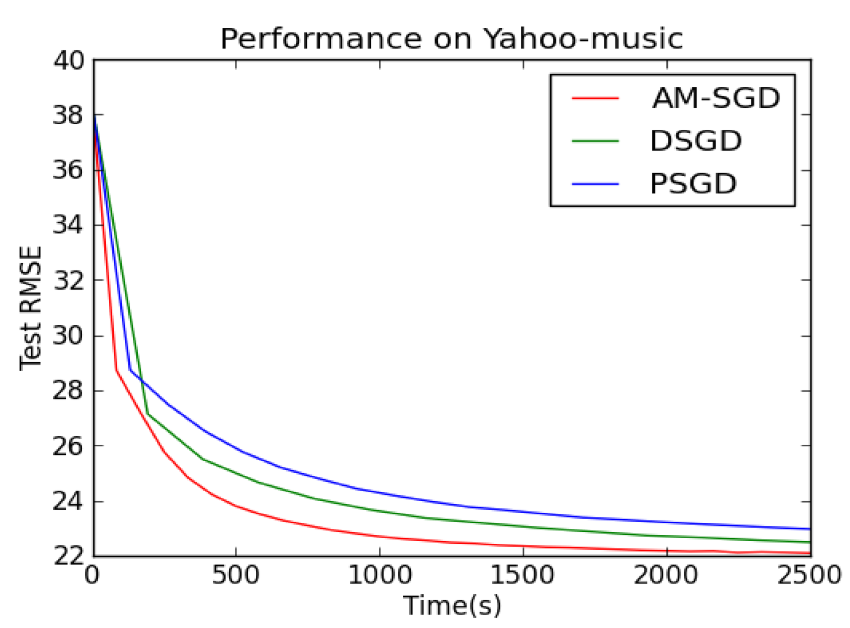
Platform

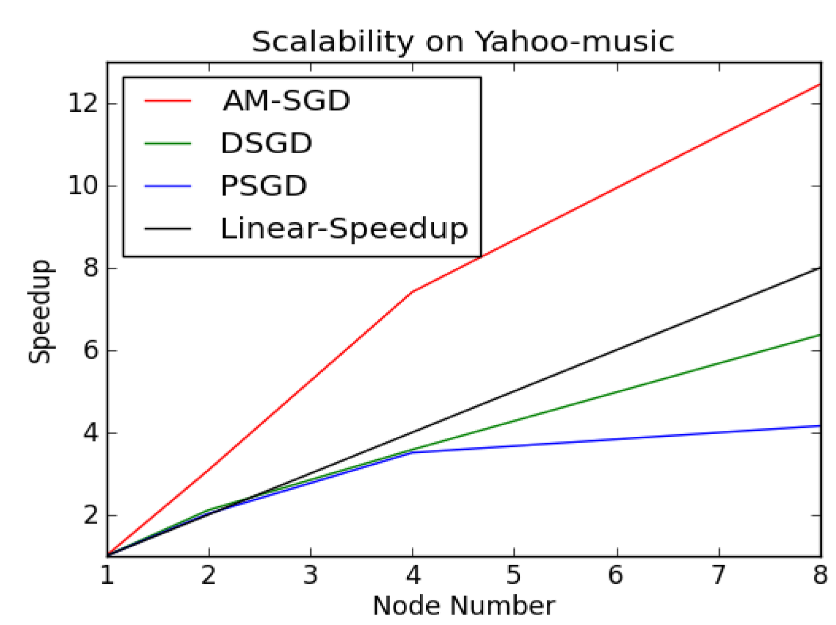
-- an MPI cluster, consists of 16 servers, each equipped with a four-core 2.30GHz AMD Opteron processor and 8GB RAM.

Data Sets

-- Netfilx, Yahoo-music, and a much large Synthetic data set.

Results on Yahoo-Music (rank K=100)





Analysis

-- AM-SGD outperforms PSGD and DSGD[2].
-- AM-SGD shows much superior scalability compared to PSGD and DSGD.

Conclusion

Conclusions

-- We found that the failure of PSGD for MF coms from the coupling of W and H in the optimization.
-- We propose an alternating parameter mixing algorithm, namely AM-SGD.
-- We proved that AM-SGD outperforms state-of-the-art SGD-based MF algorithms, i.e. PSGD and DSGD.
-- AM-SGD showed better scalability, thus is suitable for large-scale MF.

Future work

-- Comparing the convergence rate between AM-SGD and PSGD to further proved the effectiveness of AM-SGD.
-- Experimental results on large synthetic data to study the scalability.

References

1. K. B. Hall, S. Gilpin, and G. Mann, "Mapreduce/bigtable for distributed optimization," in *NIPS LCCC Workshop*, 2010.
2. R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis, "Large-scale matrix factorization with distributed stochastic gradient descent," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 69–77.