

An Accelerated Stochastic Gradient for Canonical Polyadic Decomposition

Ioanna Siaminou

*School of Electrical and Computer Engineering
Technical University of Crete*

Chania, Greece
isiaminou@isc.tuc.gr

Athanasios P. Liavas

*School of Electrical and Computer Engineering
Technical University of Crete*

Chania, Greece
liavas@telecom.tuc.gr

Abstract—We consider the problem of structured canonical polyadic decomposition. If the size of the problem is very big, then stochastic gradient approaches are viable alternatives to classical methods, such as Alternating Optimization and All-At-Once optimization. We extend a recent stochastic gradient approach by employing an acceleration step (Nesterov momentum) in each iteration. We compare our approach with state-of-the-art alternatives, using both synthetic and real-world data, and find it to be very competitive.

Index Terms—Tensor Factorization, Stochastic Optimization, CPD/PARAFAC, Nesterov Acceleration.

I. INTRODUCTION

The need to understand and analyze multi-dimensional data and their interdependencies has led to the extended use of tensors in diverse scientific fields, ranging, for example, from medicine to geodesy. An overview of tensor applications can be found in [1], [2].

Canonical Polyadic Decomposition (CPD) or Parallel Factor Analysis (PARAFAC) is a widely used model since, in many cases, it extracts meaningful structure from a given dataset.

Alternating Optimization (AO), All-at-Once Optimization (AOO), and Multiplicative Updates (MUs) are the workhorse methods towards the computation of the CPD [3], [4]. However, the very large size of the collected tensor data makes the implementation of these algorithms very computationally demanding.

Recently, various approaches have been introduced in order to deal with large-scale CPD problems. An obvious approach is the development and implementation of parallel algorithms (distributed or shared-memory) [5], [6], [7].

From a different perspective, stochastic gradient based algorithms have gained much attention, since they are relatively easy to implement, have lower computational cost, and can guarantee accurate solutions.

All members were supported by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code : T1EΔK – 03360).

A. Related Work

Sub-sampling of the target tensor \mathcal{X} using regular sampling techniques has been introduced in [8].

In [9], entries of the target tensor are sampled in a random manner and the respective blocks of the latent factors are updated at each iteration.

In [10] and [11], a distributed framework is employed, where smaller replicas of the target tensor are independently factored. The resulting factors of each independent decomposition are effectively merged at the end to obtain the final latent factors.

In [12] and [13], a set of fibers is randomly selected at each iteration and a stochastic proximal gradient step is performed. We improve upon the work of [13] by incorporating Nesterov acceleration at each iteration and a proximal term to deal with ill-conditioned cases.

B. Notation

Vectors and matrices are denoted by lowercase and uppercase bold letters, for example, \mathbf{x} and \mathbf{X} . Tensors are denoted by bold calligraphic capital letters, namely, \mathcal{X} . $\|\cdot\|_F$ denotes the Frobenius norm of the matrix or tensor argument. $\mathbf{A} \odot \mathbf{B}$ and $\mathbf{A} \otimes \mathbf{B}$ denote, respectively, the Khatri-Rao and the Hadamard product of matrices \mathbf{A} and \mathbf{B} . The outer product between vectors is denoted by the operator \circ . Finally, MATLAB notation is used when it seems appropriate.

II. CANONICAL POLYADIC DECOMPOSITION

Let tensor $\mathcal{X}^o \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ admit a rank- R factorization of the form

$$\mathcal{X}^o = [\mathbf{A}^{o(1)}, \dots, \mathbf{A}^{o(N)}] = \sum_{r=1}^R \mathbf{a}_r^{o(1)} \circ \dots \circ \mathbf{a}_r^{o(N)}, \quad (1)$$

where $\mathbf{A}^{o(i)} = [\mathbf{a}_1^{o(i)} \dots \mathbf{a}_R^{o(i)}] \in \mathbb{R}^{I_i \times R}$, for $i = 1, \dots, N$. In many cases, the latent factors have a special structure or satisfy a specific property, which is denoted as $\mathbf{A}^{o(i)} \in \mathbb{A}^i$.

The actual observation is corrupted by additive noise \mathcal{E} , thus, we observe tensor $\mathcal{X} = \mathcal{X}^o + \mathcal{E}$.

Estimates of $\mathbf{A}^{o(i)}$ can be obtained by computing matrices $\mathbf{A}^{(i)}$ that solve the optimization problem

$$\min_{\{\mathbf{A}^{(i)} \in \mathbb{A}^i\}_{i=1}^n} f_{\mathcal{X}}(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}), \quad (2)$$

where $f_{\mathcal{X}}$ is a function that measures the quality of the factorization, with a common choice being

$$f_{\mathcal{X}}(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}) = \left\| \mathcal{X} - \llbracket \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)} \rrbracket \right\|_F^2. \quad (3)$$

This optimization problem is nonconvex and, thus, difficult to solve. The matrix unfolding (or tensor matricization) has been very useful towards the solution of problem (3). More specifically, if $\widehat{\mathcal{X}} = \llbracket \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)} \rrbracket$, then the matrix unfolding for an arbitrary mode i is given by [14], [2]

$$\widehat{\mathbf{X}}^{(i)} = \mathbf{K}^{(i)} \mathbf{A}^{(i)T}, \quad (4)$$

where $\mathbf{K}^{(i)}$ is defined as

$$\mathbf{K}^{(i)} := \mathbf{A}^{(N)} \circledast \dots \circledast \mathbf{A}^{(i+1)} \circledast \mathbf{A}^{(i-1)} \circledast \dots \circledast \mathbf{A}^{(1)}. \quad (5)$$

It can be shown that, for $i = 1, \dots, N$,

$$f_{\mathcal{X}}(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}) = \left\| \mathbf{X}^{(i)} - \widehat{\mathbf{X}}^{(i)} \right\|_F^2. \quad (6)$$

These expressions form the basis of the CPD AO method. If, at iteration k , the estimated matrix factors have values $\mathbf{A}_k^{(j)}$, for $j = 1, \dots, N$, we can update $\mathbf{A}_k^{(i)}$ by solving the matrix least-squares problem

$$\mathbf{A}_{k+1}^{(i)} = \underset{\mathbf{A}^{(i)} \in \mathbb{A}^i}{\operatorname{argmin}} \left\| \mathbf{X}^{(i)} - \mathbf{K}_k^{(i)} \mathbf{A}^{(i)T} \right\|_F^2, \quad (7)$$

where

$$\mathbf{K}_k^{(i)} := \mathbf{A}_k^{(N)} \circledast \dots \circledast \mathbf{A}_k^{(i+1)} \circledast \mathbf{A}_k^{(i-1)} \circledast \dots \circledast \mathbf{A}_k^{(1)}. \quad (8)$$

The gradient of $f_{\mathcal{X}}$, with respect to $\mathbf{A}^{(i)}$, is given by

$$\begin{aligned} \nabla_{\mathbf{A}^{(i)}} f_{\mathcal{X}}(\mathbf{A}_k^{(1)}, \mathbf{A}_k^{(2)}, \dots, \mathbf{A}_k^{(N)}) \\ = \mathbf{A}_k^{(i)} \mathbf{K}_k^{(i)T} \mathbf{K}_k^{(i)} - \mathbf{X}^{(i)T} \mathbf{K}_k^{(i)}. \end{aligned} \quad (9)$$

Quantity $\mathbf{X}^{(i)T} \mathbf{K}_k^{(i)}$ is called Matricized Tensor Times Khatri-Rao Product (MTTKRP) and is the most computationally demanding part of the CPD AO algorithm. Thus, the development of efficient algorithms which do not compute a full MTTKRP during each iteration is of great interest.

III. STOCHASTIC GRADIENT CPD - BRASCPD

In [9], a stochastic gradient method has been developed for the unconstrained case, where, at each iteration, only a part of a factor is updated.

A fiber sampling technique has been developed in [12] and [13], where, at each iteration, a whole factor is updated. The scheme proposed in [13], named BrasCPD, can handle both unconstrained and constrained problems.

BrasCPD combines the fiber sampling technique with the AO algorithm. Assume, again, that, at the beginning of iteration k , the values of the estimated factors are $\mathbf{A}_k^{(j)}$, for $j = 1, \dots, N$. At iteration k , an index i is picked at random. Then, B^i mode- i fibers are sampled, indexed by $\mathcal{F}_k^i \subset \{1, 2, \dots, J^i\}$, where J^i denotes the number of rows of $\mathbf{X}^{(i)}$, and a smaller problem is constructed, namely,

$$\min_{\mathbf{A}^{(i)} \in \mathbb{A}^i} f_k^{(i)}(\mathbf{A}^{(i)}), \quad (10)$$

where

$$f_k^{(i)}(\mathbf{A}^{(i)}) = \left\| \mathbf{X}^{(i)}(\mathcal{F}_k^i, :) - \mathbf{K}_k^{(i)}(\mathcal{F}_k^i, :) \mathbf{A}^{(i)T} \right\|_F^2.$$

BrasCPD performs a proximal gradient step and updates $\mathbf{A}_k^{(i)}$ as

$$\mathbf{A}_{k+1}^{(i)} = \operatorname{prox}_{h_i} \left(\mathbf{A}_k^{(i)} - \frac{\alpha_k}{|\mathcal{F}_k^i|} \nabla f_k^{(i)}(\mathbf{A}) \right), \quad (11)$$

where h_i is the indicator function of set \mathbb{A}^i and

$$\begin{aligned} \nabla f_k^{(i)}(\mathbf{A}) = \mathbf{A}_k^{(i)} \mathbf{K}_k^{(i)T}(\mathcal{F}_k^i, :) \mathbf{K}_k^{(i)}(\mathcal{F}_k^i, :) \\ - \mathbf{X}^{(i)T}(\mathcal{F}_k^i, :) \mathbf{K}_k^{(i)}(\mathcal{F}_k^i, :). \end{aligned} \quad (12)$$

The other factors do not change during iteration k , that is, for $j \neq i$, $\mathbf{A}_{k+1}^{(j)} = \mathbf{A}_k^{(j)}$. Note that the computational cost of the *partial* MTTKRP appearing in (12) drops to $\mathcal{O}(|\mathcal{F}_k^i| R I_i)$ flops.

The performance of the algorithm is mainly determined by the step-sizes α_k [15]. In BrasCPD, diminishing step-sizes are employed, namely, $\alpha_k = \frac{\alpha}{k^\beta}$, for appropriate values of parameters α and β .

A method based on Adagrad [16], called AdaCPD, has also been proposed in [13]. The AdaCPD computes its step-sizes using an accumulated-gradient mechanism with parameters $\eta > 0$, $\beta > 0$, and $\epsilon > 0$, namely

$$\alpha_k^{(i)} = \frac{\eta}{\left(\beta + \sum_{l=1}^k \nabla f_l^{(i)}(\mathbf{A}) \right)^{1/2+\epsilon}}. \quad (13)$$

IV. ACCELERATED STOCHASTIC GRADIENT CPD

We propose an accelerated version of BrasCPD, which we call Accelerated Stochastic CPD (ASCPD). At iteration k , we follow the same sampling scheme as in [13], [12], and form the problem

$$\min_{\mathbf{A}^{(i)} \in \mathbb{A}^i} F_k^{(i)}(\mathbf{A}^{(i)}) = f_k^{(i)}(\mathbf{A}^{(i)}) + \frac{\lambda_k^{(i)}}{2} \left\| \mathbf{A}^{(i)} - \mathbf{A}_k^{(i)} \right\|_F^2. \quad (14)$$

The parameter $\lambda_k^{(i)}$ determines the condition number of the problem and is chosen such that the condition number remains “small.” More specifically, the Hessian of $f_k^{(i)}$ is

$$\mathbf{H}_k^{(i)} := \nabla^2 f_k^{(i)}(\mathbf{A}^{(i)}) = \mathbf{K}_k^{(i)T}(\mathcal{F}_k^i, :) \mathbf{K}_k^{(i)}(\mathcal{F}_k^i, :) \otimes \mathbf{I}_{I_i}. \quad (15)$$

Let $L_k^{(i)}$ and $\mu_k^{(i)}$ be, respectively, the largest and smallest eigenvalues of $\mathbf{H}_k^{(i)}$. We choose $\lambda_k^{(i)}$ such that the condition number of (14) becomes “almost constant,” that is

$$\frac{\bar{L}_k^{(i)}}{\bar{\mu}_k^{(i)}} := \frac{L_k^{(i)} + \lambda_k^{(i)}}{\mu_k^{(i)} + \lambda_k^{(i)}} \lesssim \mathcal{C}, \quad (16)$$

where \mathcal{C} is a given value (in our experiments, we set $\mathcal{C} = 10, 10^2, 10^3$). More specifically, we set

$$\lambda_k^{(i)} = \begin{cases} \mu_k^{(i)}, & \text{if } \frac{L_k^{(i)}}{\mu_k^{(i)}} < \mathcal{C}, \\ \frac{L_k^{(i)}}{\mathcal{C}}, & \text{otherwise.} \end{cases} \quad (17)$$

We perform a proximal step

$$\mathbf{A}_{k+1}^{(i)} = \text{prox}_{h_i} \left(\mathbf{Y}_k^{(i)} - \frac{1}{\bar{L}_k^{(i)}} \nabla F_k^{(i)}(\mathbf{Y}_k^{(i)}) \right), \quad (18)$$

followed by a momentum step

$$\mathbf{Y}_{k+1}^{(i)} = \mathbf{A}_{k+1}^{(i)} + \beta_k^{(i)} (\mathbf{A}_{k+1}^{(i)} - \mathbf{A}_k^{(i)}), \quad (19)$$

where

$$\beta_k^{(i)} := \frac{\sqrt{\bar{L}_k^{(i)}} - \sqrt{\bar{\mu}_k^{(i)}}}{\sqrt{\bar{L}_k^{(i)}} + \sqrt{\bar{\mu}_k^{(i)}}}. \quad (20)$$

The values of $\bar{L}_k^{(i)}$ and $\beta_k^{(i)}$ are the steps used by the “constant step scheme III” of the accelerated gradient algorithm [17, p. 81]. They can be considered as “locally optimal” for problem (14).

Note that we essentially compute $\mathbf{H}_k^{(i)}$ during the computation of $\nabla f_k^{(i)}$ (see (12) and (15)). Furthermore, the computation of its largest and smallest eigenvalues does not pose significant computational cost, especially in the cases of small R .

The ASCPD algorithm appears in Algorithm 1. A variation of our scheme is to use only the stochastic proximal gradient step of (18), without the acceleration step. We will test the effectiveness of this variation in our numerical experiments.

An important future research topic is the development of algorithms that fully exploit the second-order information $\mathbf{H}_k^{(i)}$. Some initial efforts have not resulted in algorithms superior to the one proposed in this paper, especially in the noisy cases.

V. NUMERICAL EXPERIMENTS

In this section, we run numerical experiments and test the performance, in terms of convergence speed and estimation accuracy, of the NALS algorithm of [7], AdaCPD, ASCPD, and BrasCPD with locally optimal step-size, using both synthetic and real-world data.

In our experiments, AdaCPD always outperformed BrasCPD, thus, we do not consider BrasCPD. For both

Algorithm 1: ASCPD

Result: $\{\mathbf{A}^{(i)}\}_{i=1}^N$
1 Input: tensor \mathcal{X} , $\mathbf{A}_0^{(i)} = \mathbf{Y}_0^{(i)}$, $i = 1, \dots, N$,
 blocksizes B^i , $i = 1, \dots, N$.
2 $k = 0$;
3 while *terminating condition is not satisfied* **do**
4 Uniformly sample i from $\{1, 2, \dots, N\}$;
5 Uniformly sample B^i mode- i fibers;
6 Compute stochastic gradient $\nabla F_k^{(i)}(\mathbf{Y}_k^{(i)})$;
7 Compute $L_k^{(i)}$, $\mu_k^{(i)}$, and $\lambda_k^{(i)}$;
8 Compute $\mathbf{A}_{k+1}^{(i)}$ using (18) ;
9 Compute $\mathbf{Y}_{k+1}^{(i)}$ using (19) ;
10 $k = k + 1$;
11 end

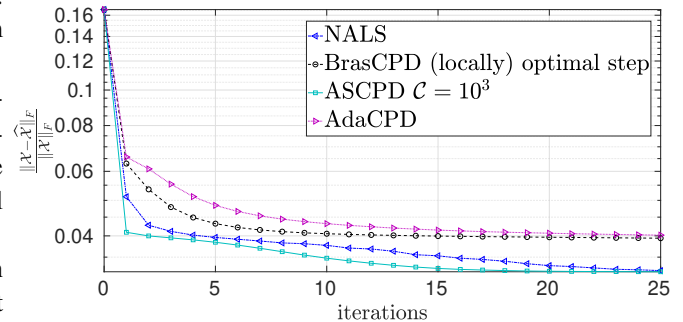
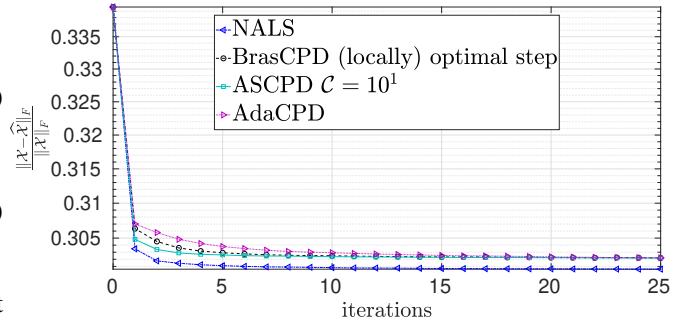


Fig. 1: Relative tensor reconstruction error for $I_1 = I_2 = I_3 = 200$, $R = 100$, $B = 500$, and SNR = 10dB (top) and 30 dB (bottom).

synthetic and real-world data, the results are extracted after averaging over 10 Monte-Carlo trials.

A. Synthetic Data

We generate a 3-rd order nonnegative tensor $\mathcal{X}^o \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$ as $\mathcal{X}^o = \llbracket \mathbf{A}^{o(1)}, \dots, \mathbf{A}^{o(N)} \rrbracket$, where the elements of each factor are independent and identically distributed (i.i.d.), uniformly distributed in $[0, 1]$. The noisy tensor is given by $\mathcal{X} = \mathcal{X}^o + \sigma_e \mathcal{E}$, where the elements of \mathcal{E} are i.i.d. $\mathcal{N}(0, 1)$.

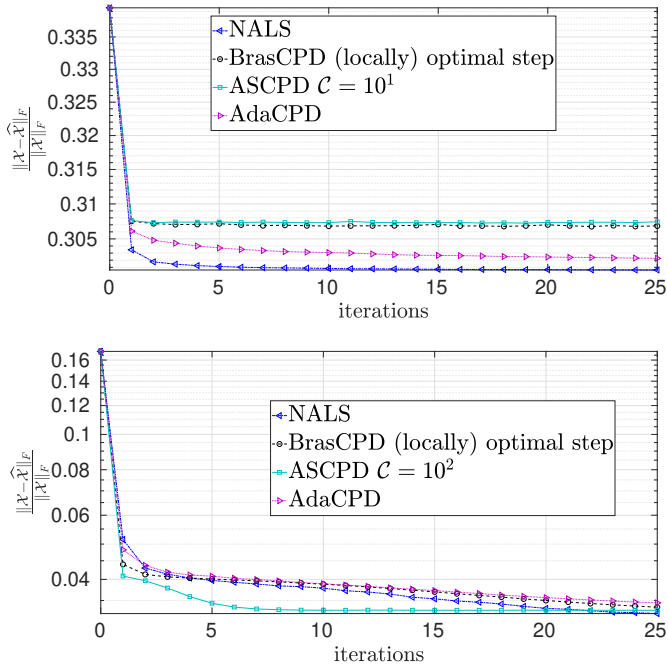


Fig. 2: Relative tensor reconstruction error for $I_1 = I_2 = I_3 = 200$, $R = 100$, $B = 100$, and SNR = 10 dB (top) and 30 dB (bottom).

We define the Signal-to-Noise Ratio (SNR)

$$\text{SNR} := \frac{\|\mathcal{X}^o\|_F^2}{\sigma_\varepsilon^2 \|\mathcal{E}\|_F^2}.$$

We adopt as performance metric the relative tensor reconstruction error at iteration k

$$m_k := \frac{\|\mathcal{X} - \hat{\mathcal{X}}_k\|_F}{\|\mathcal{X}\|_F}.$$

All stochastic gradient based algorithms use the same block-size, that is, $B^1 = B^2 = B^3 = B$. Furthermore, we note that one full iteration of the NALS algorithm of [7] requires the computation of four full MTTKRPCs (three for the factor updates and one for the acceleration step). Thus, in order to be fair, in our plots, we depict the performance metric m_k attained by each algorithm after the *same number of full iterations*, that is, we compute the performance metric for each stochastic algorithm after the required number of stochastic iterations that correspond to one full iteration of the NALS algorithm.

In Fig. 1, we plot the metric m_k versus the number of full iterations for the case where $I_1 = I_2 = I_3 = 200$, $R = 100$, $B = 500$, and SNR = 10 dB (top) and 30 dB (bottom).

In Fig. 2, we set $B = 100$ and present the corresponding plot. In all cases, we set the Adagrad parameter $\eta = 1$ (in our experiments, this was the best value), while we set the ASCPD parameter $\mathcal{C} = 10$ in the low SNR cases and $\mathcal{C} = 10^2, 10^3$ in the high SNR cases.

We observe that

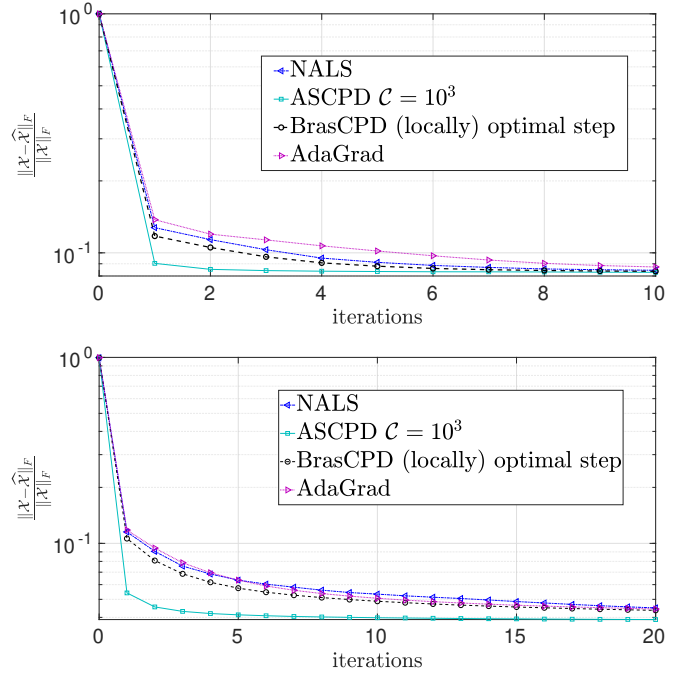


Fig. 3: Indian Pines dataset: relative tensor reconstruction error for $B = 500$, and $R = 10$ (top), $R = 100$ (bottom).

- 1) in the low SNR cases, the NALS algorithm outperforms all stochastic gradient based approaches. The relative performance of AdaCPD and ASCPD depends on the block size. For “small” block sizes, AdaCPD outperforms the ASCPD, while, for “large” block sizes, the opposite happens.

- 2) in the high SNR cases, the ASCPD outperforms all other methods. We note that the BrasCPD with locally optimal step-size in some cases outperforms AdaCPD.

B. Real-world Data

Similarly to [13], we use the Indian Pine and PaviaU datasets which are Hyperspectral Images (HSIs). HSI sensors collect data in a group of images, for different wavelength ranges. The resulting data-cube is a third order tensor. The Indian Pine dataset is of size $145 \times 145 \times 220$ and consists of data acquired via the AVIRIS sensor, on Indian Pines site in Indiana (USA). The PaviaU dataset has size $610 \times 340 \times 103$ and consists of a scene of Pavia University in Italy.¹

In Fig. 3, we plot the quantity m_k versus the number of full iterations for the Indian Pine dataset for $B = 500$ and $R = 10$ (top) and $R = 100$ (bottom).

In Fig. 4, we plot the corresponding results for the PaviaU dataset for $B = 500$ and $R = 50$ (top) and $R = 200$ (bottom). In all cases, we set $\eta = 2$ and $\mathcal{C} = 10^3$.

We observe that, in all cases, the ASCPD outperforms all other methods in terms of m_k .

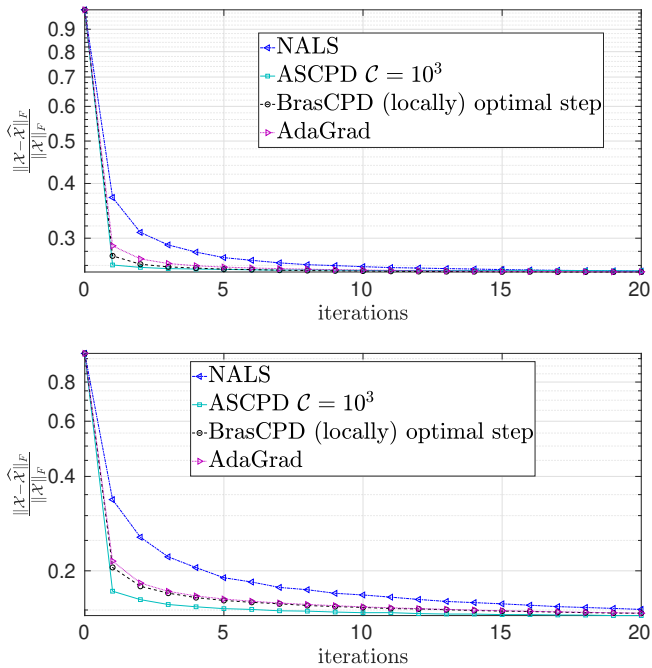


Fig. 4: PaviaU dataset: relative tensor reconstruction error for $B = 500$ and $R = 50$ (top), $R = 200$ (bottom).

VI. CONCLUSION

We adopted a stochastic gradient based framework for the solution of the structured CPD problem. We improved upon known stochastic proximal gradient schemes by incorporating Nesterov-type acceleration using parameters that are “locally optimal.” In numerical experiments, with both synthetic and real-world data, our algorithm has been proven efficient. Convergence analysis of the proposed algorithm is an interesting future topic.

REFERENCES

[1] S. Rabanser, O. Shchur, and S. Günnemann, “Introduction to tensor decompositions and their applications in machine learning,” *arXiv preprint arXiv:1711.10781*, 2017.

[2] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, “Tensor decomposition for signal processing and machine learning,” *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[3] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[4] A. Cichocki, D. P. Mandic, A. H. Phan, C. F. Caiafa, G. Zhou, Q. Zhao, and L. D. Lathauwer, “Tensor decompositions for signal processing applications from two-way to multiway component analysis,” *CoRR*, vol. abs/1403.4462, 2014. [Online]. Available: <http://arxiv.org/abs/1403.4462>

[5] G. Ballard, K. Hayashi, and R. Kannan, “Parallel nonnegative CP decomposition of dense tensors,” *CoRR*, vol. abs/1806.07985, 2018. [Online]. Available: <http://arxiv.org/abs/1806.07985>

[6] S. Smith and G. Karypis, “A medium-grained algorithm for sparse tensor factorization,” in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2016, pp. 902–911.

[7] A. P. Liavas, G. Kostoulas, G. Lourakis, K. Huang, and N. D. Sidiropoulos, “Nesterov-based alternating optimization for nonnegative tensor factorization: Algorithm and parallel implementation,” *IEEE Transactions on Signal Processing*, vol. 66, no. 4, pp. 944–953, 2018.

[8] C. I. Kanatsoulis and N. D. Sidiropoulos, “Large-scale canonical polyadic decomposition via regular tensor sampling,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[9] N. Vervliet and L. De Lathauwer, “A randomized block sampling approach to canonical polyadic decomposition of large-scale tensors,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 284–295, 2015.

[10] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, “Parcube: Sparse parallelizable tensor decompositions,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 521–536.

[11] N. D. Sidiropoulos, E. E. Papalexakis, and C. Faloutsos, “Parallel randomly compressed cubes: A scalable distributed architecture for big tensor decomposition,” *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 57–70, 2014.

[12] C. Battaglino, G. Ballard, and T. G. Kolda, “A practical randomized cp tensor decomposition,” *SIAM Journal on Matrix Analysis and Applications*, vol. 39, no. 2, pp. 876–901, 2018.

[13] X. Fu, S. Ibrahim, H.-T. Wai, C. Gao, and K. Huang, “Block-randomized stochastic proximal gradient for low-rank tensor factorization,” *IEEE Transactions on Signal Processing*, 2020.

[14] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[15] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[16] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of machine learning research*, vol. 12, no. 7, 2011.

[17] Y. Nesterov, *Introductory lectures on convex optimization*. Kluwer Academic Publishers, 2004.

¹Datasets available at http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.