FIBER-SAMPLED STOCHASTIC MIRROR DESCENT FOR TENSOR DECOMPOSITION WITH β-DIVERGENCE

Wenqiang Pu[†]*, Shahana Ibrahim*[‡]*, Xiao Fu*[‡]*, Mingyi Hong*[§]

[†]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[‡]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, USA
[§]Department of Electrical and Computer Engineering, University of Minnesota, MN, USA

ABSTRACT

Canonical polyadic decomposition (CPD) has been a workhorse for multimodal data analytics. This work puts forth a stochastic algorithmic framework for CPD under β-divergence, which is well-motivated in statistical learning—where the Euclidean distance is typically not preferred. Despite the existence of a series of prior works addressing this topic, pressing computational and theoretical challenges, e.g., scalability and convergence issues, still remain. In this paper, a unified stochastic mirror descent framework is developed for large-scale β-divergence CPD. Our key contribution is the integrated design of a tensor fiber sampling strategy and a flexible stochastic Bregman divergence-based mirror descent iterative procedure, which significantly reduces the computation and memory cost per iteration for various β. Leveraging the fiber sampling scheme and the multilinear algebraic structure of low-rank tensors, the proposed lightweight algorithm also ensures global convergence to a stationary point under mild conditions. Numerical results on synthetic and real data show that our framework attains significant computational saving compared with state-of-the-art methods.

Index Terms— Tensor decomposition, β-divergence, stochastic optimization, mirror descent method

1. INTRODUCTION

Canonical polyadic decomposition (CPD) [1, 2] is arguably one of the most important tensor decomposition models that has enabled many core tasks in signal processing and machine learning [3]. A plethora of classic CPD algorithms were developed under the Euclidean distance-based fitting criterion (i.e., the least squares loss); see, e.g., the overviews in [3, 4]. Nonetheless, instead of using least squares as the CPD model fitting criterion, the β-divergence has been found useful in many applications where the 'distance' between points are typically not measured in the Euclidean space (e.g., statistical learning and integer data analysis). Some examples of the β-divergence-based data analytics tasks include neuroscience [5], gene expression analysis [6], musical analysis [7], image decomposition [8], text mining [9], recommender systems [10], just to mention a few. Both the Eulidean distance and the KL divergence are special cases of the β divergence, with $\beta = 2$ and 1, respectively.

Designing CPD algorithms under the least squares loss has been a central task in the tensor community, for which effective first-order, (quasi-)second-order, and stochastic optimization algorithms were all developed; see, e.g., [11–14]. More discussions on algorithms can be found in recent review paper [15]. However, algorithms developed under the Euclidean distance are often not easily extendable to handle the β-divergence. CPD under statistical divergence-based loss functions has also drawn increasing attention. As a special case of the β-divergence, the KL-divergence was considered in [9], where a block majorization-minimization (MM) algorithm is developed. For the more general β-divergence cases, MM and its variants for non-negative matrix decomposition were studied in [16] and a second-order based algorithm was developed recently in [17].

Batch algorithms such as those in [16, 17] are effective to a certain extent. However, in the era of *big data*, handling large-scale β-divergence CPD problems may benefit from *stochastic* optimization for reducing per-iteration computational and memory burdens. Recently, a stochastic gradient descent (SGD) based algorithm [18] was proposed for CPD with non-Euclidean distance losses including the β-divergence. In particular, the algorithm in [18] utilizes SGD to handle the problem of interest at very large scales. However, the algorithm is developed based on randomly sampling the tensor entries. Such a strategy makes it hard to exploit some interesting algebraic properties of low-rank tensors for designing more efficient updates tailored for CPD model. In addition, convergence properties of the SGD algorithm in [18] are unclear.

In this paper, by exploiting the multilinear algebraic structure of low-rank tensors, a unified stochastic mirror descent (MD) algorithmic framework is developed for large-scale β-divergence CPD. Our idea is to integrate a recently proposed tensor fiber sampling strategy [14, 19] with the MD algorithm. The fiber sampling strategy gives rise to nicely structured (non-)convex subproblems with respect to the latent factors under the β-divergence, which admits simple and efficient mirror descent-based updates. Leveraging the recently proposed notion of *Lipschitz-like convexity* [20] from the optimization literature, the proposed algorithmic framework allows flexible choices of the local surrogate functions under the MD framework to adapt to different β. Such flexibility also helps offer lightweight updates when the latent factors are under a variety of constraints that are of interest in data analytics. In addition, the β-divergence CPD loss function in general has no Lipschitz-continuous gradient. This poses challenges on analyzing the convergence behavior—especially under stochastic settings with constraints. In this work, theoretical convergence is established under the integrated fiber sampling and MD framework. To our best knowledge, this is the first stochastic MD algorithm framework with guaranteed convergence for tensor decomposition.

2. CPD UNDER β-DIVERGENCE

Consider a data tensor X with a size of $I_1 \times I_2 \times \ldots \times I_N$. Assume that X can be approximated by a low-rank tensor M,

X. Fu is supported in part by NSF ECCS 1808159, IIS-1910118 and ARO award W911NF-19-1-0247. M.Hong is supported in part by NSF Award CIF-1910385, and ARO awardW911NF-19-1-0247.

$$X \approx M = \sum_{r=1}^{R} A_1(:,r) \circ A_2(:,r) \circ \ldots \circ A_N(:,r), \quad (1)$$

where "$\circ$" denotes the outer product of vectors, $A_n \in \mathbb{R}^{I_n \times R}$ is the mode-n latent factor, and R is the smallest positive integer such that (1) holds. On the entry level, the model in (1) can be expressed as $M_{\boldsymbol{i}} \triangleq M(i_1, i_2, \ldots, i_N) = \sum_{r=1}^{R} \prod_{n=1}^{N} A_n(i_n, r)$, where

$$\boldsymbol{i} \in \mathcal{I} \triangleq \{(i_1, i_2, \ldots, i_N) |\, i_n = 1, 2, \ldots, I_n, \forall n\}$$

denotes the entry index coordinates—an N-dimensional vector. Using β-divergence as the distance measure, approximating X by M can be formulated as the following minimization problem,

$$\begin{aligned} \min_{A_1, A_2, \ldots, A_N} \ & \frac{1}{|\mathcal{I}|} \sum_{\boldsymbol{i}} d_\beta \left(X_{\boldsymbol{i}}, M_{\boldsymbol{i}} \right) \\ \text{s.t.} \quad & M_{\boldsymbol{i}} = \sum_{r=1}^{R} \prod_{n=1}^{N} A_n(i_n, r), \ \ \forall \boldsymbol{i} \in \mathcal{I} \\ & A_n \in \mathcal{A}_n, \ \forall n, \end{aligned} \quad (2)$$

where $d_\beta(\cdot, \cdot)$ denotes the β-divergence, i.e.,

$$d_\beta(x, y) = \begin{cases} \dfrac{x}{y} - \log(\dfrac{x}{y}) - 1, & \beta = 0, \\ x \log \dfrac{x}{y} + y - x, & \beta = 1, \\ \dfrac{\left(x^\beta + (\beta - 1) y^\beta - \beta x y^{\beta - 1}\right)}{\beta(\beta - 1)}, & \beta \in \mathbb{R} / \{0, 1\}, \end{cases}$$

$\mathcal{A}_n$ is a constraint set which captures the prior information about the structure of latent factors A_n, e.g., non-negativity, sparsity and smoontheness. Since the 2010s, this problem has drawn increasing attention; see early developments in, e.g., [9]. Recent works [17, 18] considered more advanced stochastic and Gauss-Newton approaches for (2). The SGD approach in [18] is particularly suitable for massive data analytics. However, by directly applying the conventional SGD to the CPD problem, the problem geometry and the tensor algebraic properties are not fully exploited. Additionally, no convergence understanding was offered.

3. STOCHASTIC MIRROR DESCENT

For very large-scale tensors (e.g., those arise in social networks, gene networks, and computer vision), designing batch algorithms faces both computation and memory challenges (e.g., a $5000 \times 5000 \times 5000$ tensor costs more than 900GB to store if the double precision is used). Instead, stochastic optimization schemes which work with 'partial data' per iteration can significantly reduce computational and memory load. Next, we first introduce a fiber sampling strategy and then present the proposed stochastic MD algorithm.

3.1. Fiber Sampling and Block Structure

In recent year, a *fiber sampling* [14, 19, 21] strategy was used in Euclidean loss based tensor decomposition and completion to reduce the complexity and memory burdens. In [14, 19], fiber sampling-based stochastic CPD algorithms select samples $\boldsymbol{i}$ that are related to a single latent factor A_n and updates A_n based on the selected fiber samples per iteration. In the context of β-divergence-based CPD, the sampling strategy admits a couple of notable advantages:

• **Incorporating Prior on A_n:** Randomly sampling some indexes $\boldsymbol{i}$ [22] or selecting a subtensor [13] faces an issue that samples may relate to only parts of A_n. Useful prior information about the entire latent factor (e.g., column norm constraints) cannot be imposed; see [14, 15] for more discussion. Fiber sampling does not have this challenge. This advantage is shared by the least squares loss and β-divergence based CPD.

• **Block-wise Convex Approximation:** Fiber sampling also provides a way to further exploit the block-wise structure under the β-divergence. With the notion of *Lipschitz-like convexity* [20], the β-divergence with respect to each block A_n can often be upper bounded by a strongly convex function—despite the fact the block subproblems have no Lipschitz-continuous gradient. Exploiting such block structure in algorithm design can simplify the procedures for solving the subproblems (see Table 2). The existence of such block structure also helps analyze the theoretical convergence behavior of the algorithm (see Section 3.3).

The fiber sampling scheme is based on *matrix unfolding* operation, which rearranges a tensor into a matrix. The mode-n matrix unfolding of X is a $J_n \times I_n$ matrix, denoted as X_n, and the correspondence is $X_{\boldsymbol{i}} = X_n(j, i_n)$, where $j = 1 + \sum_{k=1, k \neq n}^{N} (i_k - 1) J_k$ and $J_k = \prod_{m=1, m \neq n}^{k-1} I_m$. For low rank tensor M in (1), its mode-n matrix unfolding is $M_n = H_n A_n^T$, where $H_n = A_N \odot A_{N-1} \ldots \odot A_{n+1} \odot A_{n-1} \odot \ldots \odot A_1$ and $\odot$ denotes the Khatri-Rao product.

Based on the mode-n matrix unfolding for X and M, approximating X by M with $\{A_m\}_{m \neq n}$ fixed can be regarded as a linear system approximation problem under the β-divergence, i.e., $X_n \approx H_n A_n^T$. To find this approximation, the mode-n fiber sampling uses part of rows of X_n as well as the corresponding rows of H_n. Denote $\mathcal{F}_n \subset \{1, 2, \ldots, J_n\}$ as the sampled fiber index set. Then, the sampled version of problem in (2) w.r.t. A_n becomes

$$\min_{A_n \in \mathcal{A}_n} \frac{1}{|\mathcal{F}_n| I_n} \sum_{j=1}^{|\mathcal{F}_n|} \sum_{i=1}^{I_n} d_\beta(\hat{X}_n(j, i), \hat{H}_n(j, :) A_n(i, :)^T), \quad (3)$$

where $\hat{X}_n = X_n(\mathcal{F}_n, :)$ and $\hat{H}_n = H_n(\mathcal{F}_n, :)$.

3.2. Mirror Descent

Problem in (3) boils down to a constrained least squares problem when $\beta = 2$, and a stochastic proximal gradient algorithm is developed in [14]. However, for the general β-divergence, the problem is more challenging, since it has no Lipschitz-continuous gradient and $d_\beta(x, y)$ may be nonconvex in y. Since Problem (3) is often not solvable, the key question is how to construct a good approximation for it? We notice that $d_\beta(x, y)$ is convex in y for $1 \leq \beta \leq 2$ or otherwise can be decomposed into a convex part $\check{d}_\beta(x, y)$ plus a concave part $\hat{d}_\beta(x, y)$, i.e., $d_\beta(x, y) = \check{d}(x, y) + \hat{d}(x, y)$. By making use of this property, the objective function in (3) can be upper bounded by a strongly convex function.

To get an insight on its geometric property, let us take a close look at each term in (3). Specifically, consider $d_\beta(x, h^T a)$, where $x = \hat{X}_n(j, i)$, $h^T = \hat{H}_n(j, :)$, and $a = A_n(i, :)^T$. Let us consider $\ell(x, \boldsymbol{h}^T \boldsymbol{a})$ at $\bar{\boldsymbol{a}}$ with $h_r, \bar{a}_r > 0, \forall r$, then we show the following lemma:

Lemma 1. *Let $\phi(\cdot)$ be a σ-strongly convex function satisfying the following condition,*

$$\exists L > 0, \textit{such that } L\phi(a_r) - \check{d}_\beta \left(x, \frac{h_r}{\lambda_r} a_r \right) \textit{ is convex,} \quad (4)$$

where $\lambda_r = \frac{h_r \bar{a}_r}{h^T \bar{a}} > 0, r = 1, 2, \ldots, R$, and $\sum_{r=1}^{R} \lambda_r = 1$. Then,

$$d_\beta(x, h^T a) \leq d_\beta(x, h^T \bar{a}) + \langle \nabla d_\beta(x, h^T \bar{a}), a - \bar{a} \rangle + L \sum_{r=1}^{R} D_\phi(a_r, \bar{a}_r),$$

and equality holds if and only if $a = \bar{a}$ and $D_\phi(a_r, \bar{a}_r)$ is Bregman divergence by referring function $\phi(\cdot)$, defined as

$$D_\phi(a_r, \bar{a}_r) = \phi(a_r) - \phi(\bar{a}_r) - \langle \nabla \phi(\bar{a}_r), a_r - \bar{a}_r \rangle.$$

Table 1. Examples of function $\phi(y)$ with respect to different β.

β	$\phi(y)$	Convexity
$\beta < 1$	$y^{\beta-1}$	nonconvex
$\beta > 2$	y^{β}	
$\beta = 1$	$y \log y, -\log y$	convex
$1 < \beta \leq 2$	$y^{\beta},\ y^2$	

The lemma says that $d_\beta(x, h^T a)$ at $\bar{a}$ can be upper bounded by a strongly convex function based on the Bregman divergence. Applying the Jensen's inequality for convex part together with the notion of Lipschitz-like convexity [20] and combining the linearization for the concave part (if exists) lead to the conclusion of Lemma 1. Note that condition in (4) is easy to satisfy, e.g., by choosing $\phi(a) = \check{d}_\beta(x, a)$, and it is referred as Lipschitz-like convexity condition for function pair $\big(\phi(\cdot), \check{d}_\beta(x, \cdot)\big)$. Lemma 1 indicates that properly choosing $\phi(\cdot)$ to 'fit' the geometry of the convex part of β-divergence can construct a strongly convex upper bound function using the Bregman divergence. This can be straightforwardly extended to multiple i, j case. Also note that MM schemes in [9, 16] are special cases of the upper bound function indicated by Lemma 1. Consequently, by properly choosing $\phi(\cdot)$ to adapt the function geometry, Problem (3) is approximately solved via the following update:

$$A_n^{t+1} = \arg\min_{A \in \mathcal{A}_n} \langle \hat{G}^t, A - A_n^t \rangle + \frac{1}{\eta_t} D_\Phi(A, A_n^t), \qquad (5)$$

where t is the iteration index and $\eta_t > 0$ is a proper step size. $\hat{G}^t$ is the gradient at A_n^t, given as

$$\hat{G}_n^t = \frac{1}{|\mathcal{F}_n| I_n} \left[(\hat{H}_n A_n^t)^{\cdot \beta - 2} \circledast (\hat{H}_n A_n^t - \hat{X}_n) \right]^T \hat{H}_n, \qquad (6)$$

where $A^{\cdot\beta}$ denotes the entry-wise power operation and $\circledast$ is the Hadamard product. Here, $D_\Phi(A, A_n^t)$ is the Bregman divergence which measures the 'distance' between A and A_n^t by a reference function $\Phi(A) = \sum_{j,i} \phi_{ji}(A(j, i))$ and $\phi_{j,i}(\cdot)$ is a proper σ-strongly convex function.

Combining the fiber sampling together with randomized block selection, the proposed algorithm is summarized in Algorithm 1. The subproblem in (5) is also known as stochastic MD method in the optimization literature. In addition, note that some existing block coordinate descent based algorithms [9, 14, 16, 23, 24] for matrix or tensor decomposition can be regarded as special cases of the proposed algorithm, i.e., properly choosing ϕ and the step size η_t, or using full samples to update A_n. A couple of additional remarks are as follows:

Remark 1 (Choice of ϕ)**.** *Some choices of $\phi(\cdot)$ w.r.t. various β's are given in Table 1. Note that $d_\beta(x, y)$ may be nonconvex in y, under mild conditions (see Assumption 1), any strongly convex function $\phi(\cdot)$ can be used with guaranteed convergence. Detailed convergence analysis is presented in Section 3.3.*

Remark 2 (Solving problem (5))**.** *The Bregman divergence $D_\Phi(A, A')$ is defined in a entry-wise form, which makes the update in* (5) *has closed-form solutions for many kinds of pair $\phi(\cdot)$ and $\mathcal{A}_n$. Some examples are given in Table 2.*

Table 2. Examples of $\phi(\cdot)$ and $\mathcal{A}_n$ which have closed-form solution.

$\phi(\cdot)$	$\mathcal{A}_n$
y^β $(\beta \notin [0, 1]), -\log y$	non-negative
$y \log y$	non-negative, probability simplex
y^2	many forms, see Table I in [15]

Algorithm 1 Stochastic Mirror Descent (MD) Algorithm

Require: $X, A_1^0, A_2^0, \ldots, A_N^0, \phi, \{\eta_t\}_{t=0,1,\ldots}$
1: **for** $t = 0, 1, \ldots,$ until meet some convergence criteria **do**
2: Uniformly sample $n \in \{1, 2, \ldots, N\}$;
3: Uniformly sample fibers $\mathcal{F}_n \subset \{1, 2, \ldots, J_n\}$
4: Compute the sampled gradient $\hat{G}_n^t$ by (6)
5: $A_n^{t+1} = \arg\min_{A_n \in \mathcal{A}_n} \langle \hat{G}_n^t, A - A_n^t \rangle + \frac{1}{\eta_t} D_\Phi(A, A_n^t)$
6: $A_i^{t+1} = A_i^t, \forall i \neq n$
7: **end for**

3.3. Convergence Analysis

Classic convergence analysis for stochastic MD usually requires the Lipschitz-continuous gradient assumption [25], which does not hold for β-divergence. The recently proposed notion of Lipschitz-like convexity [20] provides a way for dealing with our problem without Lipschitz-continuous gradient. Based on this notion, several recent works studied stochastic MD type algorithms for problem in convex [26, 27] and nonconvex [28–30] settings, but none of them covers the proposed algorithm, where each block subproblem in (3) may be nonconvex and inexactly solved one step of stochastic MD algorithm. In this work, we offer a tailored analysis for the CPD problem of interest.

Denote the objective in (2) as $F(\boldsymbol{A}) = F(A_1, A_2, \ldots, A_N)$. Then, Problem (2) can be re-expressed as follows:

$$\min_{A_1, A_2, \ldots, A_N} F(\boldsymbol{A}) + h(\boldsymbol{A}) \qquad (7)$$

where $h(\boldsymbol{A}) = \sum_{n=1}^N h_n(A_n)$ and $h_n(A)$ is the indicator function of set $\mathcal{A}_n$, i.e., $h_n(A) = 0$ if $A \in \mathcal{A}_n$ and otherwise $h_n(A) = \infty$.

Our convergence analysis starts by using the following "reference" function in each iteration:

$$\mathcal{L}(\boldsymbol{A}; \boldsymbol{A}^t) := F(\boldsymbol{A}) + h_n(\boldsymbol{A}) + \frac{1}{2\lambda} D_\Phi(\boldsymbol{A}, \boldsymbol{A}^t)$$

where $\lambda > 0$ is a constant such that $\mathcal{L}(\boldsymbol{A}; \boldsymbol{A}^t)$ is strongly convex and $D_\Phi(\boldsymbol{A}, \boldsymbol{A}^t) = \frac{1}{N} \sum_{n=1}^N D_\Phi(A_n, A_n^t)$. Denoting $\hat{\boldsymbol{A}}^t = \arg\min_{\boldsymbol{A}} \mathcal{L}(\boldsymbol{A}; \boldsymbol{A}^t)$, the following lemma shows that $D_\Phi(\hat{\boldsymbol{A}}^t, \boldsymbol{A}^t)$ can be used as a stationarity measure for Problem (7).

Lemma 2. *$\boldsymbol{A}$ is a stationary point of Problem* (7)*, i.e., $\mathbf{0} \in \nabla F(\boldsymbol{A}) + \partial h(\boldsymbol{A})$, where $\partial h(\boldsymbol{A})$ denotes the subgradient, if and only if $D_\Phi(\hat{\boldsymbol{A}}, \boldsymbol{A}) = 0$.*

Then, we make the following mild assumption and show the convergence in Theorem 1.

Assumption 1. *The set $\mathcal{A}_n$ is a compact set, $\forall n$.*

Theorem 1. *Suppose $\mathbb{E}\left[\|\hat{G}_n^t\|^2\right]$ is bounded and $\phi(\cdot)$ satisfies condition in Table 1. Then, for diminishing step size η_t, Algorithm 1 converge to a stationary point of problem* (7) *in expectation,* $\liminf_{t \to \infty} \mathbb{E}\left[D_\phi(\hat{\boldsymbol{A}}^t, \boldsymbol{A}^t)\right] = 0$.

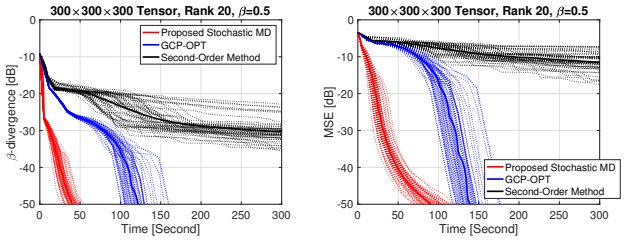

Fig. 1. β-divergence (left) and MSE metric (right) over 50 trials.

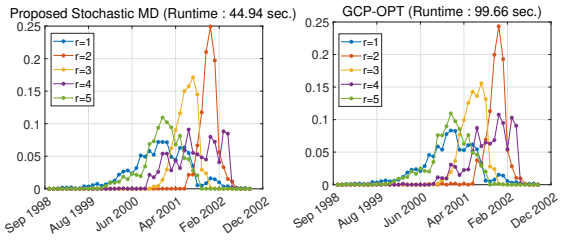

Fig. 2. Different components of mode-3 latent factor of the Enron tensor data ($200 \times 200 \times 47$ tensor with rank $R = 5$). The plot represents the (normalized) amount of email exchanges between the employees of the company during each month.

The key step for proving Theorem 1 is shown in Lemma 3. In the expectation sense, if the step size η_t is properly chosen, then the Bregman Moreau envelop [31] of Problem (7), denoted as $\mathcal{M}(\boldsymbol{A}^t)$, decreases after every iteration. As the optimization process continues in a Markovian manner [32], Lemma 3 together with the tower rule and telescope trick helps complete the proof of Theorem 1.

Lemma 3. *Suppose $\phi(\cdot)$ satisfies conditions in Table 1 and denote $\mathcal{M}(\boldsymbol{A}^t) = \min_{\boldsymbol{A}} \mathcal{L}(\boldsymbol{A}; \boldsymbol{A}^t)$. Let $\boldsymbol{A}^{t+1}$ be generated by Algorithm 1 at iteration t. Then, we have*

$$\mathbb{E}\left[\mathcal{M}(\boldsymbol{A}^{t+1})\right] \leq \mathcal{M}(\boldsymbol{A}^t) - \frac{c\eta_t}{4\lambda^2 N} D_\Phi(\hat{\boldsymbol{A}}^t, \boldsymbol{A}^t) + \frac{\eta_t^2}{4\lambda\sigma}\mathbb{E}\left[\|\hat{G}_n^t\|^2\right],$$

where c is a positive constant.

4. SIMULATIONS

4.1. Synthetic Data

A third-order tensor of size $300 \times 300 \times 300$ and rank 20 is considered. The latent matrices A_1, A_2, and A_3 are drawn from i.i.d. uniform distribution between 0 and 1 and $\mathcal{A}_n$ for all n are set to be the nonnegative orthant. Two algorithms are selected as baselines. The first one is the generalized CPD optimization (GCP-OPT) algorithm proposed in [18] and the second one is the second-order method developed in [17]. GCP-OPT is implemented by `gcp_opt` provided in Tensor Toolbox [33] and 'adam' is selected as the optimization solver with 40×300 entry samples per iteration. The second-order method [17] is implemented by `nlsb_gndl` shared by the authors of Tensorlab [34] and the 'preconditioner' is set as 'block-Jacobi'. For the proposed stochastic MD algorithm, function $\phi_{ji}(y)$ is chosen as $c_{ji}y^{0.5}$ as suggested in Table 1, where $c_{ji} \geq 0$ is specified according to Lemma 1. For the proposed algorithm, 40 fibers (40×300 entry samples) are used per iteration and the step size is set as $\eta_t = \frac{1}{t^{0.02}}$. All the algorithms are tested under β-divergence based CPD setting β to 0.5.

The objective value and the averaged mean squared error (MSE) of the estimated latent factors (see definition in [14]) over 50 independent trials are shown in Fig. 1, where the solid lines represent the average convergence curves and dash lines correspond to the individual trials. Clearly, the two stochastic algorithms, i.e., the proposed stochastic MD and GCP-OPT have much faster convergence behavior than the second order method since both of them enjoy low computation cost, i.e., only about 0.05% entries are used per iteration. Further, the proposed stochastic MD is much faster than GCP-OPT, especially in the beginning.

4.2. Real Data

The algorithms are also evaluated using the Enron email dataset that contains the emails between the employees of the Enron Corporation during the period of the infamous Enron scandal (September 1998–July 2002). A subset of the Enron email data is used. The subset counts the number of words exchaged through emails among 200 employees over 47 months, giving rise to a $200 \times 200 \times 47$ tensor. This tensor contains 24282 nonzero entries. The tensor data is preprocessed following idea in [35] to reduce bias from prolific senders or receivers. Since the actual rank of the tensor R is unknown, $R = 5$ is chosen following existing works, e.g., [36,37]. For the proposed stochastic MD algorithm, $\beta = 1$ and $\phi(y) = y \log y$ are used, and thus the objective function in (2) is equivalent to minimizing the KL divergence between the observed tensor and the low rank CPD model. The step size schedule is the same as before and the number of fibers sampled per iteration is also 40. The proposed algorithm is benchmarked by GCP-OPT using the same β divergence. The algorithms are run until the relative change of the objective value is less than 10^{-3}.

Fig. 2 presents different components in the learned mode-3 latent factor from the Enron tensor data. The columns of the latent factor are normalized with respect to the ℓ_1-norm for better visual representations. The plot represents the amount of the words exchanged between the employees of the company through emails over the 47 months. One can see that the temporal profiles extracted by both algorithms are similar to those discovered in [36, 37]. These temporal profiles maybe interpreted as email communication loads associated with 5 clusters of employees (see previous study in [36, 37]). Rediscovering these temporal profiles helps 'cross-validate' the proposed and the baseline algorithms' effectiveness in extracting information from this particular dataset. According to the temporal profiles, a major surge in email communications is identified during Apr. 2001 and Aug. 2001, when the change of CEO and the collapse of Enron shares happened. Another major peak is during Sep. 2001 and Dec. 2001, when the company was undergoing the legal trials and the bankruptcy. The highest peak profile was connected to the cluster of executives in [36], which makes sense. Besides producing meaningful data mining results, a more important observation is that the proposed algorithm is much faster than GCP-OPT for attaining essentially the same results.

5. CONCLUSION

A unified stochastic MD algorithm framework with guaranteed convergence is developed for CPD under the β-divergence. The proposed algorithm enjoys low computational and memory cost. Significant computational saving relative to state-of-the-art methods is observed on both synthetic and real data sets. The proposed algorithm framework promises future extensions for dealing a large variety of loss functions [18] that are of great interests in real-world tensor decomposition applications.

6. REFERENCES

[1] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[2] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[3] 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 Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[4] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[5] M. Mørup, L. K. Hansen, J. Parnas, and S. M. Arnfred, "Decomposing the time-frequency representation of eeg using non-negative matrix and multi-way factorization," *Technical University of Denmark Technical Report*, pp. 1–28, 2006.

[6] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[8] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1255–1261, 2001.

[9] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1272–1299, 2012.

[10] X. Luo, Y. Yuan, M. Zhou, Z. Liu, and M. Shang, "Non-negative latent factor model based on β-divergence for recommender systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.

[11] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[12] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5052–5065, 2016.

[13] 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.

[14] X. Fu, S. Ibrahim, H.-T. Wai, C. Gao, and K. Huang, "Block-randomized stochastic proximal gradient for low-rank tensor factorization," *IEEE Trans. Signal Process.*, vol. 68, pp. 2170–2185, 2020.

[15] X. Fu, N. Vervliet, L. De Lathauwer, K. Huang, and N. Gillis, "Computing large-scale matrix and tensor decomposition with structured factors: A unified nonconvex optimization perspective," *IEEE Signal Processing Magazine*, vol. 37, no. 5, pp. 78–94, 2020.

[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[17] M. Vandecappelle, N. Vervliet, and L. De Lathauwer, "A second-order method for fitting the canonical polyadic decomposition with non-least-squares cost," *IEEE Trans. Signal Process.*, vol. 68, pp. 4454–4465, 2020.

[18] D. Hong, T. G. Kolda, and J. A. Duersch, "Generalized canonical polyadic tensor decomposition," *SIAM Review*, vol. 62, no. 1, pp. 133–163, 2020.

[19] 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.

[20] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[21] M. Sørensen and L. De Lathauwer, "Fiber sampling approach to canonical polyadic decomposition and application to tensor completion," *SIAM Journal on Matrix Analysis and Applications*, vol. 40, no. 3, pp. 888–917, 2019.

[22] A. Beutel, P. P. Talukdar, A. Kumar, C. Faloutsos, E. E. Papalexakis, and E. P. Xing, "Flexifact: Scalable flexible factorization of coupled tensors on hadoop," in *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 2014, pp. 109–117.

[23] A.-H. Phan, P. Tichavskỳ, and A. Cichocki, "Fast alternating ls algorithms for high order candecomp/parafac tensor factorizations," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4834–4846, 2013.

[24] L. T. K. Hien and N. Gillis, "Algorithms for nonnegative matrix factorization with the kullback-leibler divergence," *arXiv preprint arXiv:2010.01935*, 2020.

[25] C. D. Dang and G. Lan, "Stochastic block mirror descent methods for nonsmooth and stochastic optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 856–881, 2015.

[26] F. Hanzely and P. Richtárik, "Fastest rates for stochastic mirror descent methods," *arXiv preprint arXiv:1803.07374*, 2018.

[27] H. Lu, ""relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent," *INFORMS Journal on Optimization*, vol. 1, no. 4, pp. 288–303, 2019.

[28] D. Davis, D. Drusvyatskiy, and K. J. MacPhee, "Stochastic model-based minimization under high-order growth," *arXiv preprint arXiv:1807.00255*, 2018.

[29] S. Zhang and N. He, "On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization," *arXiv preprint arXiv:1806.04781*, 2018.

[30] T. Gao, S. Lu, J. Liu, and C. Chu, "Randomized bregman coordinate descent methods for non-lipschitz optimization," *arXiv preprint arXiv:2001.05202*, 2020.

[31] H. H. Bauschke, M. N. Dao, and S. B. Lindstrom, "Regularizing with bregman–moreau envelopes," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3208–3228, 2018.

[32] 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.

[33] B. W. Bader, T. G. Kolda, et al., "Matlab tensor toolbox version 3.0-dev," *Available online, Oct*, 2017.

[34] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab 3.0," *available online, URL: www. tensorlab. net*, 2016.

[35] B. W. Bader, R. A. Harshman, and T. G. Kolda, "Temporal analysis of social networks using three-way DEDICOM.," Tech. Rep., jun 2006.

[36] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro, "From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 493–506, 2012.

[37] X. Fu, K. Huang, W.-K. Ma, N. Sidiropoulos, and R. Bro, "Joint tensor factorization and outlying slab suppression with applications," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6315–6328, 2015.