Adaptive Sketching for Fast and Convergent Canonical Polyadic Decomposition: Supplementary Material

Kareem S. Aggour[*1] **Alex Gittens**[*2] **Bülent Yener**[2]

1. Overview

This supplement provides proofs of the theorems in the main body of the paper, discusses regret guarantees for CPD-MWU, and expands upon the experimental evaluation of the CPD-MWU algorithm.

Section 2 formulates two assumptions on the sketching matrices under which the objective decreases at each iteration of sketched CPD-ALS: namely, that the sketching matrix gives a regularized subspace embedding and satisfies an approximate matrix multiplication property. These assumptions are discussed and sample complexities under which they hold for several popular sketching ensembles are given in Table 1.

Section 3 provides a proof that these assumptions ensure that the CPD objective will decrease at each iteration of sketched ALS. Section 4 uses this sufficient decrease result to establish sublinear convergence to an approximate critical point; in this result the sketching rate and/or regularization parameters may vary on each iteration. In particular, this result provides a guarantee on the performance of the CPD-MWU algorithm if all the sketching rates used are sufficient conservative enough to ensure sufficient decrease.

Section 5 provides regret bound guarantees for the guarded variant of CPD-MWU, and shows that this implies the performance of guarded CPD-MWU is at least asymptotically as good as that of regularized, non-sketched CPD-ALS if one of the available sketching rates is 1.

Finally, Section 6 provides additional experimental evaluation of CPD-MWU.

2. Structural Conditions for Proximally Regularized Sketched CPD-ALS

The ALS updates can be interpreted as iterative refinements of the factor matrices to improve the quality of the tensor approximation: if the residual tensor at time t is given by

$$\boldsymbol{\mathcal{R}}_t = \boldsymbol{\mathcal{X}} - [\![\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t]\!],$$

then the ALS update for $\mathbf{A}_t$ learns $\boldsymbol{\Delta}_{t+1}$ to minimize

$$\begin{aligned} F(\mathbf{A}_t + \boldsymbol{\Delta}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) &= \|\boldsymbol{\mathcal{X}} - [\![\mathbf{A}_t + \boldsymbol{\Delta}_t, \mathbf{B}_t, \mathbf{C}_t\|_F^2 = \|\boldsymbol{\mathcal{R}}_t - [\![\boldsymbol{\Delta}_{t+1}, \mathbf{B}_t, \mathbf{C}_t]\!]\|_F^2 \\ &= \|\mathbf{R}_{(1),t} - \boldsymbol{\Delta}_{t+1}(\mathbf{C}_t \odot \mathbf{B}_t)^T\|_F^2 \\ &= \|\mathbf{R}_{(1),t}\mathbf{P}_t - \boldsymbol{\Delta}_{t+1}\mathbf{M}_t\|_F^2 + \|\mathbf{R}_{(1),t}(\mathbf{I} - \mathbf{P}_t)\|_F^2, \end{aligned}$$

where $\mathbf{P}_t$ is the orthonormal projector onto the row span of the matrix $\mathbf{M}_t = (\mathbf{C}_t \odot \mathbf{B}_t)^T$. In order to get an accurate estimate of $\mathbf{A}_{t+1} = \mathbf{A}_t + \boldsymbol{\Delta}_{t+1}$ using sketching, it seems evident that the sketching matrix $\mathbf{S}_{t+1}$ must preserve the geometry of this problem. In particular, it must essentially preserve the relationships between $\mathbf{R}_{(1),t}$, the row space of $\mathbf{M}_t$, and the orthogonal complement to that row space.

[*] Equal contribution [1] GE Global Research, Niskayuna, New York, USA [2] Rensselaer Polytechnic Institute, Troy, New York, USA. Correspondence to: Alex Gittens <gittea@rpi.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

The following structural conditions suffice to ensure that these relationships are preserved. The dependences on t are suppressed to avoid excessive notation. Thus $\mathbf{S} = \mathbf{S}_{t+1}$, $\mathbf{P} = \mathbf{P}_t$, $\mathbf{M} = \mathbf{M}_t$, $\mathbf{R} = \mathbf{R}_{(1),t}$ and so on.

Assumption 1 (Covariance preservation)**.** *There is an ε in $[0, 1)$ for which $\mathbf{S}$ satisfies*

$$(1-\varepsilon)\mathbf{P} \preceq \mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T)^\dagger\mathbf{M} \preceq (1+\varepsilon)\mathbf{P}, \quad \textit{and} \tag{1}$$

$$(1-\varepsilon)\mathbf{P} \preceq \mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T + \lambda\mathbf{I})^{-1}\mathbf{M} \preceq (1+\varepsilon)\mathbf{P}. \tag{2}$$

Assumption 2 (Approximate Matrix Multiplication)**.** *There is a η in $(0, 1)$ for which $\mathbf{S}$ satisfies*

$$\|\mathbf{R}(\mathbf{I} - \mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 \leq \eta\|\mathbf{R}\mathbf{P}\|_F^2. \tag{3}$$

To understand Assumption 1 note that when $\lambda = 0$, it reduces to the requirement that (1) hold, which is a standard subspace embedding property[1] that simply requires that we sample enough of the system matrix $\mathbf{M}$ that we capture its row space (Wang et al., 2017). This condition is achievable by many forms of sketching: e.g. by sampling rows uniformly when $\mathbf{M}$ is incoherent, or by sampling rows proportionally to their squared norms when $\mathbf{M}$ has low condition number. See (Woodruff, 2014; Wang et al., 2017) for more details.

The additional requirement (2) becomes relevant when λ is nonzero because of the biasing introduced by the proximal regularization; one trivial way to satisfy this condition is to take λ to be smaller than the minimum singular value of $\mathbf{M}\mathbf{M}^T$ and use a standard sketching method that satisfies (1). Practically, one can assume that $\mathbf{M}$ is slowly changing between iterations, and choose λ on the scale of the bottom eigenvalue of $\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T$ from the previous iteration, or simply fix λ as a small constant.

Assumption 2 is a version of the matrix multiplication property (Wang et al., 2017): it requires that we sketch conservatively enough of that we can accurately compute the projection of the residual onto the system matrix. This condition can be achieved using sketching if any of the residual is in the row space of $\mathbf{M}$. This is because $\mathbf{S}$ can be constructed using standard sketching techniques so that a approximate matrix multiplication property holds

$$\|\mathbf{R}(\mathbf{I} - \mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 \leq \eta'\|\mathbf{R}\|_F^2 \tag{4}$$

holds (Wang et al., 2017) for a small η' in $(0, 1)$, then since a portion of the residual is in the row space of $\mathbf{M}$, we have that

$$\|\mathbf{R}\mathbf{P}\|_F^2 \geq \nu\|\mathbf{R}\|_F^2,$$

for some ν in $(0, 1]$, which implies that

$$\|\mathbf{R}(\mathbf{I} - \mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 \leq \frac{\eta'}{\nu}\|\mathbf{R}\mathbf{P}\|_F^2,$$

so (3) holds with $\eta = \frac{\eta'}{\nu}$. Intuitively, the less of the residual that is in the row space of $\mathbf{M}$ (the smaller ν is), the higher the sketching rate must been to ensure that (3) holds. In the extreme, when $\mathbf{R}$ is perpendicular to the row space of $\mathbf{M}$, sketching will not ensure that (3) holds, as the geometry of the problem is too delicate. In this case, the only acceptable sketching rate is $s = 1$. This is a natural requirement: in this case $\mathbf{A}_t$ is already optimal so the only value of $\mathbf{\Delta}$ that will not increase the objective is $\mathbf{0}$, but any sketched system is likely to have a non-zero solution.

Assumption 2 therefore implies the important requirement that we increase the sketching rate as we approach convergence, because at the optimal factor matrices, the residual is orthogonal to the system matrix.

Per this discussion, sufficient conditions for both assumptions to hold are $\lambda = O(\sigma_{\min}(\mathbf{M}))$, $\|\mathbf{R}\mathbf{P}\|_F^2 \geq \nu\|\mathbf{R}\|_F^2$ for some ν in $(0, 1]$, and that $\mathbf{S}$ satisfy the standard approximate matrix multiplication property (4) with constant $\eta' = \eta\nu$ and the standard covariance preservation property (1). Table 1 gives the sketch sizes sJK required for several standard sketching methods to satisfy Assumptions 1 and 2. These complexities were adapted from Table 5 of (Wang et al., 2017); see (Wang et al., 2017) for details of these sketching modalities.

[1] Some algebraic manipulations show the stated semidefinite estimation of the projection of $\mathbf{M}$ is equivalent to the standard statement of the subspace embedding property on the matrix of right singular vectors of $\mathbf{M}$: the inequality $\|\mathbf{V}^T\mathbf{S}\mathbf{S}^T\mathbf{V} - \mathbf{I}\|_2 \leq \varepsilon$.

Sketching Method	Covariance Estimation	Matrix Multiplication
Leverage Row Sampling	$\ell = O\left(\frac{r}{\varepsilon^2}\log\frac{r}{\delta}\right)$	$\ell = O\left(\frac{r}{\nu\eta\delta}\right)$
Uniform Row Sampling	$\ell = O\left(\frac{\mu r}{\varepsilon^2}\log\frac{r}{\delta}\right)$	$\ell = O\left(\frac{\mu r}{\nu\eta\delta}\right)$
Shrinked Leverage	$\ell = O\left(\frac{r}{\varepsilon^2}\log\frac{r}{\delta}\right)$	$\ell = O\left(\frac{r}{\nu\eta\delta}\right)$
SRHT	$\ell = O\left(\frac{r+\log(JK)}{\varepsilon^2}\right)$	$\ell = O\left(\frac{r+\log(JK)}{\nu\eta\delta}\right)$
Gaussian Projection	$\ell = O\left(\frac{r+\log\delta^{-1}}{\varepsilon^2}\right)$	$\ell = O\left(\frac{r}{\nu\eta\delta}\right)$
CountSketch	$\ell = O\left(\frac{r^2}{\delta\varepsilon^2}\right)$	$\ell = O\left(\frac{r}{\nu\eta\delta}\right)$

Table 1: The two middle columns provide an upper bound on the sketch size $\ell = sJK$ necessary to satisfy the covariance estimation and matrix multiplication properties, (1) and (3) respectively, using the indicated sketching ensembles. These properties hold with (separate) failure probabilities δ. Here, r denotes the rank of $\mathbf{M}$, μ denotes the row coherence (the largest leverage score) of $\mathbf{M}$, and $\eta = \|\mathbf{RP}\|_F^2/\|\mathbf{R}\|_F^2$ denotes the fraction of the mass of the residual that is in the span of $\mathbf{M}$.

3. Monotonicity of Sketched CPD-ALS

As before we suppress the dependence on t to avoid excessive notation when it causes no ambiguities, so $\mathbf{S} = \mathbf{S}_{t+1}, \mathbf{B} = \mathbf{B}_t$, $\mathbf{C} = \mathbf{C}_t$, and so on; also, $\mathbf{X} = \mathbf{X}_{(1)}$.

Theorem 1. *Let $\mathbf{S}$ be a sketching matrix that satisfies Assumptions 1 and 2. Determine $\mathbf{A}_{t+1}$ by minimizing the proximally regularized sketched objective*

$$\mathbf{A}_{t+1} = \operatorname{argmin}_{\mathbf{A}} \left\|\left(\mathbf{X} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T\right)\mathbf{S}\right\|_F^2 + \lambda\|\mathbf{A} - \mathbf{A}_t\|_F^2. \tag{5}$$

The approximation error of $\mathbf{A}_{t+1}$ satisfies

$$F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) \geq \left(2(1-\varepsilon)(1-\eta) - (1+\varepsilon)^2(1+\eta)^2 - \varepsilon\eta\right)\|\mathbf{RP}\|_F^2.$$

Some remarks are in order.

First, the quantity $\|\mathbf{RP}\|_F^2$— the projection of the current residual onto the span of the Khatri-Rao product—measures how much of the current error in approximating the tensor can be reduced by updating $\mathbf{A}$, so is the largest possible decrease in the CPD objective that is achievable by updating only $\mathbf{A}$. Thus this theorem gives a guarantee on how much of the largest possible decrease in the objective is achieved when proximally regularized sketched CPD-ALS is employed.

Second, this bound is indeed non-vacuous: e.g. when $\varepsilon = 0.1$, and $\eta = 0.1$, we see that the sketched regularized ALS step achieves more than fourteen percent of the decrease that would have been achieved by a non-sketched, non-regularized ALS step: $F(\mathbf{A}_t) - F(\mathbf{A}_{t+1}) > 0.14\|\mathbf{RP}\|_F^2$.

The takeaway from Theorem 1 is that sketched regularized ALS will decrease the objective almost as well as non-sketched ALS when:

- the sketching rate is sufficient to capture the row space of $\mathbf{M}$, and the regularization does not bias the solution too much (Assumption 1), and
- the projection of the residual onto the system matrix is sufficiently large and the sketching rate is sufficient to capture this projection (Assumption 2).

In principle one could attempt to determine appropriate λ and sketching rate s separately at each iteration of ALS, but the obvious naïve approaches for doing so are costly. This consideration was one of the motivations behind the CPD-MWU algorithm.

The following guarantee for the performance of sketched CPD-ALS when the sketching is done using uniform row sampling is a consequence of Theorem 1 and the results in Table 1.

Lemma 1. *Let the sketching matrix $\mathbf{S} \in \mathbb{R}^{JK \times \ell}$ correspond to uniform row sampling and μ be the row coherence of $\mathbf{M}$, both as described in (Wang et al., 2017), and $\nu = \frac{\|\mathbf{RP}\|_F^2}{\|\mathbf{R}\|_F^2}$ denote the fraction of the residual that is in the span of $\mathbf{M}$. If the sketching rate s is such that $\ell = sIJK = \Omega\left(\frac{\mu R}{\nu \varepsilon^2 \delta} \log \frac{R}{2\delta}\right)$, then with probability at least $1 - \delta$,*

$$F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) \geq (1 - \varepsilon)\|\mathbf{RP}\|_F^2.$$

Proof. Note that $\mathrm{rank}(\mathbf{M}) \leq R$, so by Table 1, the given value of ℓ ensures that both Assumptions 1 and 2 are satisfied (with $\eta = \varepsilon$), simultaneously, with probability at least $1 - \delta$. Applying Theorem 1 then gives

$$F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) \geq \left(2(1-\varepsilon)^2 - (1+\varepsilon)^4 - \varepsilon^2\right) \|\mathbf{RP}\|_F^2.$$

Expansion of the polynomial in ε on the right-hand side allows us to conclude that the quantity on the right-hand side is larger than $1 - 20\varepsilon$ for all $\varepsilon \in (0, 1]$. We obtain the final statement in the theorem by absorbing the constant 20 into the asymptotic complexity of ℓ. □

To establish Theorem 1, we will use the following generalized polarization identity.

Lemma 2. *If the positive-semidefinite matrices $\mathbf{A}$ and $\mathbf{Z}$ satisfy $c\mathbf{A} \preceq \mathbf{Z} \preceq C\mathbf{A}$ where $0 < c < C$, then for all conformal matrices $\mathbf{X}$ and $\mathbf{Y}$,*

$$\mathrm{Tr}[\mathbf{X}^T\mathbf{Z}\mathbf{Y}] \geq \frac{c}{4}\mathrm{Tr}\left[(\mathbf{X}+\mathbf{Y})^T\mathbf{A}(\mathbf{X}+\mathbf{Y})\right] - \frac{C}{4}\mathrm{Tr}\left[(\mathbf{Y}-\mathbf{X})^T\mathbf{A}(\mathbf{Y}-\mathbf{X})\right].$$

Proof. Start with the generalized polarization identity,

$$\mathrm{Tr}[\mathbf{X}^T\mathbf{Z}\mathbf{Y}] = \frac{1}{4}\mathrm{Tr}\left[(\mathbf{X}+\mathbf{Y})^T\mathbf{Z}(\mathbf{X}+\mathbf{Y})\right] - \frac{1}{4}\mathrm{Tr}\left[(\mathbf{Y}-\mathbf{X})^T\mathbf{Z}(\mathbf{Y}-\mathbf{X})\right],$$

which can be verified by expanding and collecting the terms on the right side, then use the semidefinite inequalities relating $\mathbf{A}$ and $\mathbf{Z}$ to reach the claimed result. □

Proof of Theorem 1. We find it convenient to work with the refinement of the $\mathbf{A}$ factor, $\boldsymbol{\Delta} = \mathbf{A}_{t+1} - \mathbf{A}_t$. By changing variables in the sketched objective (5), we see that and the updated factor matrix obtained using proximal regularized sketching is given by

$$\begin{aligned}\boldsymbol{\Delta} &= \mathrm{argmin}_{\boldsymbol{\Xi}} \|(\mathbf{R} - \boldsymbol{\Xi}\mathbf{M})\mathbf{S}\|_F^2 + \lambda\|\boldsymbol{\Xi}\|_F^2 \\ &= \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}\mathbf{M}^T + \lambda\mathbf{I})^{-1}.\end{aligned}$$

Consider the objective function evaluated at $F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C})$:

$$\begin{aligned}F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) &= \|\mathbf{R} - \boldsymbol{\Delta}\mathbf{M}\|_F^2 \\ &= \|\mathbf{R}\|_F^2 - 2\langle \mathbf{R}, \boldsymbol{\Delta}\mathbf{M}\rangle + \|\boldsymbol{\Delta}\mathbf{M}\|_F^2 \\ &= F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - 2\langle \mathbf{R}, \boldsymbol{\Delta}\mathbf{M}\rangle + \|\boldsymbol{\Delta}\mathbf{M}\|_F^2,\end{aligned}$$

where $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}\mathbf{B}^T)$ denotes the trace inner-product. Therefore the decrease we want to lower bound can be written as

$$F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) = 2\langle \mathbf{R}, \boldsymbol{\Delta}\mathbf{M}\rangle - \|\boldsymbol{\Delta}\mathbf{M}\|_F^2 := T_1 - T_2. \tag{6}$$

Observe that

$$\begin{aligned}T_1 &= 2\mathrm{Tr}\left[\mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}\mathbf{M}^T + \lambda\mathbf{I})^{-1}\mathbf{M}\mathbf{R}^T\right] \\ T_2 &= \mathrm{Tr}\left[\mathbf{R}\mathbf{S}\mathbf{S}^T\left(\mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T + \lambda\mathbf{I})^{-1}\mathbf{M}\right)^2\mathbf{S}\mathbf{S}^T\mathbf{R}^T\right].\end{aligned}$$

We will show that $T_1 > T_2$ so that the decrease is lower bounded as claimed in the statement of the theorem.

First we upper bound T_2. Because (2) of Assumption 1 holds, we see that

$$\mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T + \lambda\mathbf{I})^{-1}\mathbf{M} \preceq (1+\varepsilon)\mathbf{P}.$$

Now observe that if a matrix $\mathbf{A} \preceq \mathbf{P}$, then by conjugating both sides of this semidefinite inequality by $\mathbf{A}^{1/2}$, it follows that $\mathbf{A}^2 \preceq \mathbf{A}$, and transitively that $\mathbf{A}^2 \preceq \mathbf{P}$; in particular,

$$\left(\mathbf{M}^T(\mathbf{M}\mathbf{S}\mathbf{S}^T\mathbf{M}^T + \lambda\mathbf{I})^{-1}\mathbf{M}\right)^2 \preceq (1+\varepsilon)^2\mathbf{P}.$$

This gives the estimate

$$\begin{aligned} T_2 &\leq (1+\varepsilon)^2 \mathrm{Tr}\left[\mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\mathbf{S}\mathbf{S}^T\mathbf{R}\right] = (1+\varepsilon)^2\|\mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\|_F^2 \\ &\leq (1+\epsilon)^2(\|\mathbf{R}\mathbf{P}\|_F + \|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F)^2, \end{aligned}$$

and using the fact that the approximate matrix multiplication property (3) of Assumption 2 holds gives the final estimate

$$T_2 \leq (1+\varepsilon)^2(1+\eta)^2\|\mathbf{R}\mathbf{P}\|_F^2. \tag{7}$$

Next we lower bound T_1. Begin by applying Lemma 2 and (2) of Assumption 1 to obtain

$$\begin{aligned} T_1 &\geq \left(\tfrac{1-\varepsilon}{2}\right)\mathrm{Tr}[\mathbf{R}(\mathbf{S}\mathbf{S}^T+\mathbf{I})\mathbf{P}(\mathbf{S}\mathbf{S}^T+\mathbf{I})\mathbf{R}^T] - \left(\tfrac{1+\varepsilon}{2}\right)\mathrm{Tr}[\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{R}^T] \\ &= \left(\tfrac{1-\varepsilon}{2}\right)\|\mathbf{R}(\mathbf{I}+\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 - \left(\tfrac{1+\varepsilon}{2}\right)\|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 \\ &= \left(\tfrac{1-\varepsilon}{2}\right)\left(\|\mathbf{R}(\mathbf{I}+\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 - \|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2\right) - \varepsilon\|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2. \end{aligned}$$

Apply the polarization identity $\|\mathbf{A}+\mathbf{B}\|_F^2 - \|\mathbf{A}-\mathbf{B}\|_F^2 = 4\langle\mathbf{A},\mathbf{B}\rangle$ to continue the estimation:

$$\begin{aligned} T_1 &\geq 2(1-\varepsilon)\langle\mathbf{R}\mathbf{P}, \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\rangle - \varepsilon\|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F^2 \\ &\geq 2(1-\varepsilon)\langle\mathbf{R}\mathbf{P}, \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\rangle - \varepsilon\eta\|\mathbf{R}\mathbf{P}\|_F^2, \end{aligned}$$

where the second inequality follows from an application of the approximate matrix multiplication property (3) of Assumption 2.

Finally, note that, again because of the same approximate matrix multiplication property,

$$\begin{aligned} \langle\mathbf{R}\mathbf{P}, \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\rangle &= \langle\mathbf{R}\mathbf{P}, \mathbf{R}\mathbf{P}\rangle - \langle\mathbf{R}\mathbf{P}, \mathbf{R}\mathbf{P} - \mathbf{R}\mathbf{S}\mathbf{S}^T\mathbf{P}\rangle \\ &\geq \|\mathbf{R}\mathbf{P}\|_F^2 - \|\mathbf{R}\mathbf{P}\|_F\|\mathbf{R}(\mathbf{I}-\mathbf{S}\mathbf{S}^T)\mathbf{P}\|_F \\ &\geq \|\mathbf{R}\mathbf{P}\|_F^2 - \eta\|\mathbf{R}\mathbf{P}\|_F^2 \\ &= (1-\eta)\|\mathbf{R}\mathbf{P}\|_F^2, \end{aligned}$$

so in fact

$$T_1 \geq (2(1-\varepsilon)(1-\eta) - \varepsilon\eta)\,\|\mathbf{R}\mathbf{P}\|_F^2. \tag{8}$$

The estimates (7) and (8) together imply the desired result, that

$$F(\mathbf{A}_t) - F(\mathbf{A}_{t+1}) = T_1 - T_2 \geq \left(2(1-\varepsilon)(1-\eta) - (1+\varepsilon)^2(1+\eta)^2 - \varepsilon\eta\right)\|\mathbf{R}\mathbf{P}\|_F^2.$$

□

4. Convergence to an approximate critical point

Theorem 1 implies that sufficient decrease is achieved at each iteration of sketched ALS. That is, the decrease in the objective is proportional to the square of the gradient.

Lemma 3. *Under the conditions of Theorem 1,*

$$F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) \leq F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - c\|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}, \mathbf{C})\|_F^2.$$

where c is at least $(2(1-\varepsilon)(1-\eta) - (1+\varepsilon)^2(1+\eta)^2 - \varepsilon\eta)\|\mathbf{M}\|_2^{-2}$.

Proof. Theorem 1 states that the objective decreases by an amount proportional to

$$\|\mathbf{RP}\|_F^2 = \|\mathbf{RM}^T(\mathbf{M}^T)^\dagger\|_F^2 \geq \sigma_{\min}(\mathbf{M}^\dagger)^2\|\mathbf{RM}^T\|_F^2 = \|\mathbf{M}\|_2^{-2}\|\nabla_{\mathbf{A}} F(\mathbf{A})\|_F^2,$$

where $\sigma_{\min}$ denotes the smallest *nonzero* singular value of a matrix. Using Theorem 1 followed by this estimate, we see that

$$\begin{aligned} F(\mathbf{A}_{t+1}, \mathbf{B}, \mathbf{C}) &\leq F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - (2(1-\varepsilon)(1-\eta) - (1+\varepsilon)^2(1+\eta)^2 - \varepsilon\eta)\|\nabla_{\mathbf{A}_t} F(\mathbf{A}_t, \mathbf{B}, \mathbf{C})\|_F^2 \\ &\leq F(\mathbf{A}_t, \mathbf{B}, \mathbf{C}) - (2(1-\varepsilon)(1-\eta) - (1+\varepsilon)^2(1+\eta)^2 - \varepsilon\eta)\|\mathbf{M}\|_2^{-2}\|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}, \mathbf{C})\|_F^2. \end{aligned}$$

□

The next result leverages this guarantee of sufficient decrease to establish that proximally regularized sketched CPD-ALS algorithms converge to an approximate critical point of the objective function at a sublinear rate. First, we make an additional assumption that is standard in the tensor decomposition literature.

Assumption 3 (Bounded Iterates)**.** *The gradients of F with respect to $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are L-Lipschitz along the solution path, e.g.*

$$\|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) - \nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F \leq L\|\mathbf{A}_{t+1} - \mathbf{A}_t\|_F$$

and similarly for the gradients with respect to $\mathbf{B}$ and $\mathbf{C}$.

Note that the alternative assumption that the factor matrices are uniformly bounded in some norm implies that this property holds, because then the gradients of F with respect to each factor matrix are smooth functions and the solution path lies in a compact set.

Theorem 2. *Assume that the bounded iterates Assumption 3 holds, and that in each iteration of sketched regularized ALS the sketching rate and regularization parameter are chosen so that Assumptions 1 and 2 are satisfied with probability at least $1-\delta$, and so that the regularization parameters λ_t and the scale factors c_t in Lemma 3 satisfy $c_t^{-1} + \lambda_t^{-1} \leq R$ for some $R > 0$.*

The sketched regularized ALS algorithm achieves a $O(T^{-1/2})$-approximate critical point in T iterations with probability at least $(1-\delta)^T$:

$$\min_{1 \leq t \leq T} \|\nabla F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F \leq \sqrt{\frac{C' F(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0))}{T}},$$

where C' is at most $12RL^2$.

This result is analogous to standard results on the convergence of gradient methods for nonconvex optimization under sufficient decrease conditions (Beck, 2017, Theorem 10.15). The difference is that this result applies to sketched ALS, which is a sketched Gauss-Siedel method. The proximal regularization is what allows the handling of the Gauss-Siedel steps: sufficient decrease bounds the size of the gradient, while proximal regularization bounds the size of the changes from iteration to iteration of the parameter variables.

Proof. Using the Lipschitz continuity of the gradient,

$$\begin{aligned} \|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F &\leq \|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) - \nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F + \\ &\quad \|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t) - \nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)\|_F + \\ &\quad \|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) - \nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F + \\ &\quad \|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F \\ &\leq L\|\mathbf{B}_{t+1} - \mathbf{B}_t\|_F + L\|\mathbf{C}_{t+1} - \mathbf{C}_t\|_F + \\ &\quad L\|\mathbf{A}_{t+1} - \mathbf{A}_t\|_F + \|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F. \end{aligned}$$

It follows that (assuming WLOG that $L \geq 1$) with C=4,

$$\|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+}1)\|_F^2 \leq CL^2 \left(\|\mathbf{B}_{t+1} - \mathbf{B}_t\|_F^2 + \|\mathbf{C}_{t+1} - \mathbf{C}_t\|_F^2 + \right. \\ \left. \|\mathbf{A}_{t+1} - \mathbf{A}_t\|_F^2 + \|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F^2 \right)$$

Conditioning on the occurrence of sufficient decrease at each solve,

$$\|\nabla_{\mathbf{A}} F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F^2 \leq \frac{F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)}{c_t},$$

and because each factor matrix was obtained using proximal steps,

$$\begin{aligned} \|\mathbf{A}_{t+1} - \mathbf{A}_t\|_F^2 &\leq \frac{F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)}{\lambda_t} \\ \|\mathbf{B}_{t+1} - \mathbf{B}_t\|_F^2 &\leq \frac{F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)}{\lambda_t} \\ \|\mathbf{C}_{t+1} - \mathbf{C}_t\|_F^2 &\leq \frac{F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})}{\lambda_t}, \end{aligned}$$

and thus

$$\|\nabla_{\mathbf{A}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F^2 \leq \frac{CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right).$$

Likewise,

$$\begin{aligned} \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F \leq\ & \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) - \nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F + \\ & \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t) - \nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)\|_F + \\ & \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)\|_F \\ \leq\ & L\|\mathbf{C}_{t+1} - \mathbf{C}_t\|_F + L\|\mathbf{B}_{t+1} - \mathbf{B}_t\|_F + \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)\|_F, \end{aligned}$$

and using the sufficient decrease condition

$$\|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t)\|_F^2 \leq \frac{F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)}{\lambda_t}$$

and the proximal condition on $\mathbf{B}$ and $\mathbf{C}$ given above, we see that

$$\begin{aligned} \|\nabla_{\mathbf{B}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F^2 &\leq \frac{CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_{t+1}, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right) \\ &\leq \frac{CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right). \end{aligned}$$

Similarly,

$$\begin{aligned} \|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F \leq\ & \|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) - \nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F + \\ & \|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F \\ \leq\ & L\|\mathbf{C}_{t+1} - \mathbf{C}_t\|_F + \|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F. \end{aligned}$$

Using the sufficient decrease condition

$$\|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t)\|_F^2 \leq \frac{F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})}{\lambda_t}$$

and the proximal condition on $\mathbf{C}$ given above, we see that

$$\begin{aligned} \|\nabla_{\mathbf{C}} F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F^2 &\leq \frac{CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right) \\ &\leq \frac{CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right). \end{aligned}$$

The total gradient therefore satisfies

$$\|\nabla F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1})\|_F^2 \leq \frac{3CL^2(c_t + \lambda_t)}{c_t \lambda_t} \left(F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t) - F(\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}) \right).$$

Assuming the constant is uniformly bounded by an absolute constant C' and summing over the first T iterates, we conclude that

$$\begin{aligned} \min_{1 \leq t \leq T} \|\nabla F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F &\leq \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\nabla F(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)\|_F^2} \\ &\leq \sqrt{\frac{C'(F(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0) - F(\mathbf{A}_T, \mathbf{B}_T, \mathbf{C}_T))}{T}} \\ &\leq \sqrt{\frac{C' F(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)}{T}}. \end{aligned}$$

□

5. Regret Guarantee for guarded CPD-MWU

As mentioned above, for sketched CPD-ALS algorithms to ensure that the tensor reconstruction error decreases, the sketching rates should be selected so that the geometry of the ALS problem is preserved. As the geometry of the problem changes from iteration to iteration, sketching rates that ensure decrease at one iteration may not ensure decrease at the next. In particular, as the reconstruction error decreases, the sketching used must become more conservative in order to capture the geometry of the problem. The CPD-MWU algorithm implictly addresses this concern by tracking which sketching rates are effective.

Theorem 2 allows the sketching rates to change, as long as they guarantee sufficient decrease, so if the sketching rates are all chosen conservatively, then CPD-MWU will converge to a critical point with high probability. It is natural to consider the regret of CPD-MWU: how does the performance of the sequence of sketching rates chosen by this algorithm compare to the performance of the single best sketching rate chosen in hindsight?

We consider a guarded version of the CPD-MWU algorithm, where a sketching rate is used to update the weights only if using it to calculate $\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}$ results in a lower approximation error: so within each iteration of CPD-MWU, sketching rates are sampled and used until one of them succeeds in lowering the approximation error[2]. Our regret guarantees for the guarded CPD-MWU algorithm follow directly from the bounded regret guarantees for the label efficient forecaster (Cesa-Bianchi & Gábor, 2006). The regret bound in Theorem 6.1 of (Cesa-Bianchi & Gábor, 2006) assumes that the loss function is bounded in $[0, 1]$, but as stated in the CPD-MWU algorithm, the loss function (of a sketching rate)

$$\ell(s_i) = \frac{\|\boldsymbol{\mathcal{X}} - [\![\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}]\!]\|_F - \|\boldsymbol{\mathcal{X}} - [\![\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t]\!]\|_F}{\text{runtime}(t)\|\boldsymbol{\mathcal{X}}\|_F}$$

may be negative and unbounded, because the error at the $t+1$ iterate may be arbitrarily large. The guarded version of the algorithm ensures the loss function is bounded: since a sketching rate is used to update the weights only when it decreases the approximation error, the loss function satisfies

$$\ell(s_i) \leq \frac{\|\boldsymbol{\mathcal{X}} - [\![\mathbf{A}_{t+1}, \mathbf{B}_{t+1}, \mathbf{C}_{t+1}]\!]\|_F}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F} \leq \frac{\sqrt{F(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F},$$

where τ_{min} is the computation time corresponding to solving a linear system using the most aggressive of the available sketching rates[3]. Similarly,

$$\ell(s_i) \geq \frac{-\sqrt{F(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F},$$

[2] There is no efficient way to check this deterministically, but a randomized error approximator can be employed as in (Battaglino et al., 2018).

[3] We assume the ridge regression problems are solved with direct methods, so their solution times are insensitive to the properties of the matrices in each iteration.

so with guarding the loss function falls in the range $[-C, C]$ for $C = \frac{\sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F}$.

The regret bound for the label efficient forecaster now implies, by shifting the loss function, that

Theorem 3 (Corollary of (Cesa-Bianchi & Gábor, 2006, Theorem 6.1))**.** *If the guarded CPD-MWU algorithm is run for T iterations with parameters $\varepsilon \in (0,1)$ and $\eta = \sqrt{\frac{2\varepsilon \ln N}{T}}$, then*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i_t,t}\right] - \min_{i=1,\ldots,N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,t}\right] \leq 2\frac{\sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau_{min}\|\boldsymbol{\mathcal{X}}\|_F}\sqrt{2\varepsilon^{-1}T\ln N},$$

where i_t denotes the sketching rate used at the t^{th} iteration of CPD-MWU.

Note that the expectation on the left in Theorem 3 is taken with respect to both the randomness in the sketching and the randomness in the updating of the weights, whereas the expectation on the right is taken only with respect to the randomness in the sketching. Thus Theorem 3 bounds the regret of the guarded CPD-MWU algorithm with respect to the best possible selection of a single sketching rate to use across all iterations that could have been made in hind-sight.

In particular, if one of the sketching rates corresponds to always fully solving the linear system, then the loss for that sketching rate is deterministic and can be computed, yielding

$$\min_{i=1,\ldots,N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,t}\right] \leq \frac{1}{\tau\|\boldsymbol{\mathcal{X}}\|_F}\sum_{t=1}^{T}(\|\boldsymbol{\mathcal{X}} - \llbracket \mathbf{A}_t^\star, \mathbf{B}_t^\star, \mathbf{C}_t^\star \rrbracket_F - \|\boldsymbol{\mathcal{X}} - \llbracket \mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{C}_{t-1}\|_F)$$

where $\mathbf{A}_t^\star, \mathbf{B}_t^\star$, and $\mathbf{C}_t^\star$ are the factors yielded by running regularized CPD-ALS from the initial guesses $\mathbf{A}_0, \mathbf{B}_0$, and $\mathbf{C}_0$, and τ is the time it takes to complete one round of regularized non-sketched CPD-ALS. Thus the guarded CPD-MWU algorithm satisfies

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell_{i_t,t}\right] &\leq \frac{1}{\tau\|\boldsymbol{\mathcal{X}}\|_F}\sum_{t=1}^{T}(\|\boldsymbol{\mathcal{X}} - \llbracket \mathbf{A}_t^\star, \mathbf{B}_t^\star, \mathbf{C}_t^\star \rrbracket_F \\ &\qquad -\|\boldsymbol{\mathcal{X}} - \llbracket \mathbf{A}_{t-1}, \mathbf{B}_{t-1}, \mathbf{C}_{t-1}\|_F) \\ &\qquad + 2\frac{\sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F}\sqrt{2\varepsilon^{-1}T\ln N} \\ &= \frac{1}{\tau\|\boldsymbol{\mathcal{X}}\|_F}\left(\sqrt{F(\mathbf{A}_T^\star, \mathbf{B}_T^\star, \mathbf{C}_T^\star)} - \sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}\right) \\ &\qquad + 2\frac{\sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F}\sqrt{2\varepsilon^{-1}T\ln N}.\end{aligned}$$

Dividing both sides by the number of iterates gives

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_{i_t,t}\right] \leq \frac{\sqrt{F(\mathbf{A}_T^\star, \mathbf{B}_T^\star, \mathbf{C}_T^\star)} - \sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau T} + 2\frac{\sqrt{F(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0)}}{\tau_{\text{min}}\|\boldsymbol{\mathcal{X}}\|_F}\sqrt{\frac{2\ln N}{\varepsilon T}}.$$

Recall that $\ell_{i_t,t}$ is the ratio between the decrease of the objective in the t-th iterate and the time it took to solve the ridge regression problem. Thus, the expectation on the left-hand side is the *average time rate of decrease in the objective error, per iteration*, and we see that as T increases, this average rate is asymptotically at least as fast as that of proximally regularized CPD-ALS without sketching. This bound is pessimistic, as in practice we see that CPD-MWU outperforms regularized CPD-ALS.

6. Additional Experimental Results

6.1. Setup

CPD-MWU is implemented in Apache Spark 1.6.0 using Python 2.7.13. Numpy 1.13.1 is used for matrix operations, with BLAS 3.71 and ATLAS 3.10.2 used by numpy. Experiments were performed on at most 500 executors across 16 nodes of an HP DL380 cluster running the Cloudera Distribution including Apache Hadoop version 5.12.0. Each node features 16 dual core 2.93GHz CPU's, 300GB of RAM, and 6TB of disk.

6.2. Synthetic Dataset Creation

The synthetic tensor dimensions were chosen such that each tensor is $\approx$ 100GB, small enough to rapidly execute thousands of Spark jobs, but large enough to be prohibitive for traditional single server CPD implementations on commodity hardware. We further generated and analyzed a 1TB synthetic tensor to demonstrate the ability to decompose Big Data tensors, and to explore the impact of the sketching rate range on the algorithms being compared.

6.2.1. Noisy Tensors

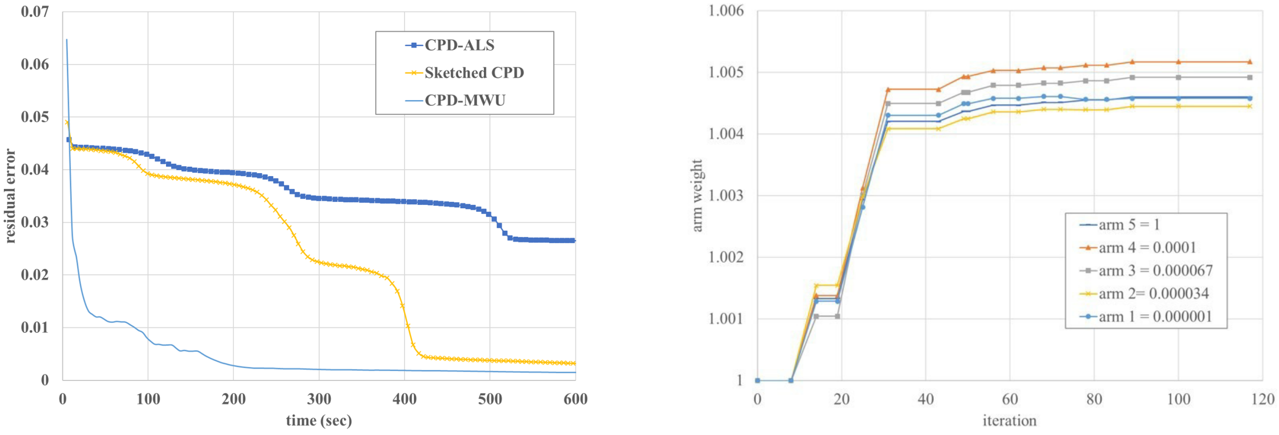
As in (Tomasi & Bro, 2006; Acar et al., 2011), we generated synthetic tensors with varying types and degrees of noise (homoscedastic, heteroskedastic, and both). Homoscedastic noise is independent of the amplitude of the values in the tensor, and thus the applied noise is sampled from the same distribution across the entire tensor. Heteroskedastic noise is proportional to the values in the tensor, so larger absolute values in the tensor may have larger levels of noise applied. While we analyzed a range of noisy tensors, the presented analysis focuses on tensors with the maximum combination of both homo- and heteroskedastic noise used by (Tomasi & Bro, 2006; Acar et al., 2011).

6.2.2. Ill-Conditioned Tensors

We further synthesized ill-conditioned tensors (tensors with collinear factor matrices) following the methodology described in (Kiers et al., 1999). The presented analysis focuses on extremely ill-conditioned tensors, in which the factor matrices have column collinearity of 0.9.

6.3. Results

Figure 1a shows the residual error over time for a representative ill-conditioned tensor. Figure 1b shows the five MWU sketching rate weights over time for the representative ill-conditioned tensor run. We observe that early in the decomposition CPD-MWU learns the second most aggressive sketching rate, $3.4 * 10^{-5}$ is the best for the given loss function. It then evolves and learns the second least aggressive sketching rate, 10^{-4} is the best. Thus, the sketching rate preference moves from more aggressive to less aggressive over time.



(a) Residual error decomposing an ill-conditioned tensor

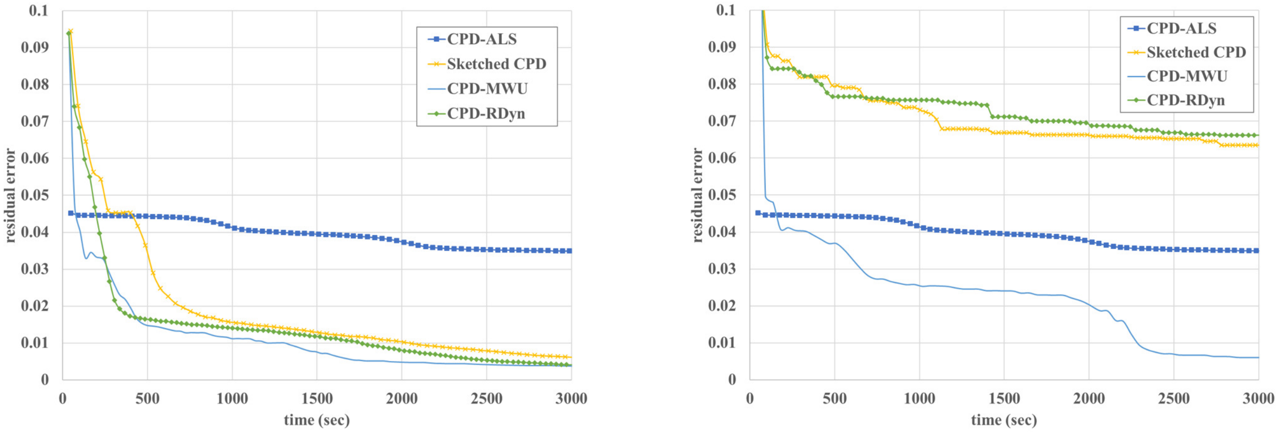
(b) Sketching rate weights over time during decomposition

Figure 1: (a) Residual error over time when decomposing an extremely ill-conditioned tensor (factor matrices each have column collinearity of 0.9). CPD-MWU converges to a significantly lower final residual error, significantly faster, as compared to traditional CPD-ALS. (b) Sketching rate weights over time for the same ill-conditioned tensor run. Early on CPD-MWU learns the second most aggressive sketching rate, $3.4{*}10^{-5}$ is the best for the given loss function. It later learns the second least aggressive sketching rate, 10^{-4} is the best, evolving the rate preference from more aggressive to less aggressive over time.

6.3.1. Impact of Sketching Rate Range on a 1TB Tensor

Here we demonstrate how the sketching rate range can impact CPD-MWU performance. We synthesized a 1TB ill-conditioned tensor in $\mathbb{R}^{366\times366\times1MM}$ to simultaneously demonstrate the scalability of the algorithm implementation. For this experiment we also compare CPD-MWU to CPD-RDyn, which is short for CPD Random Dynamic selection of the sketching arms. We explored using the same sketching arms but without the MWU-based selection algorithm and its commensurate overhead to determine if randomly selecting the arms was more efficient than intelligent arm selection.

As observed in Figure 2a, when the sketching rate range is conservative CPD-MWU slightly outperforms the other algorithms, all of which outperform traditional CPD-ALS. Here CPD-RDyn's performance is competitive with CPD-MWU. However, when we provide an overly aggressive sketching rate range, as shown in Figure 2b, we see that CPD-MWU significantly outperforms the other algorithms, as it alone is able to quickly learn which sketching rates to use and which to avoid. We also observe that CPD-MWU is superior to the fixed sketching Sketched CPD algorithm, which actually performs worse than traditional CPD-ALS when an overly aggressive static sketching rate is used. This points to the importance of algorithms like CPD-MWU that can intelligently select the sketching rate rather than requiring hand-crafted hyperparameter selection.



(a) Residual error decomposing 1TB ill-conditioned tensor with conservative sketching rate range

(b) Residual error decomposing 1TB ill-conditioned tensor with aggressive sketching rate range

Figure 2: (a) Residual error over time when decomposing an extremely ill-conditioned 1TB tensor (factor matrices each have column collinearity of 0.9) with a conservative sketching rate range $[10^{-6}, 10^{-4}]$. CPD-MWU modestly outperforms other algorithms, with CPD-RDyn competitive with CPD-MWU. (b) Residual error over time when decomposing the same ill-conditioned 1TB tensor with an aggressive sketching rate range $[10^{-9}, 10^{-6}]$. CPD-MWU significantly outperforms the other algorithms since it is able to learn which rates to use and which to avoid.

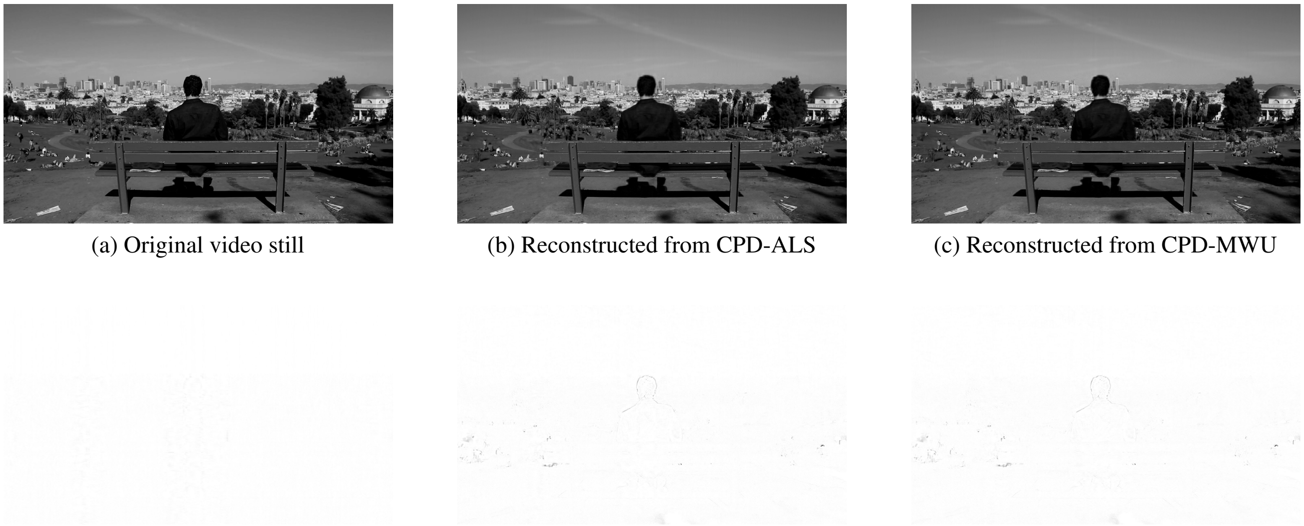
6.3.2. Downstream Application - Video Analysis

In video analysis scenarios in which the goal is to model the static or background portion of the video, residual error and mean squared error do not do an adequate job of representing the quality of the decomposition. Therefore, we instead visualize and compare frames from the video as well as the corresponding reconstructed frames derived from the factor matrices using both CPD-ALS and CPD-MWU to visually inspect the quality.

In Figure 3 we see a still frame from the park bench video, as well as images reconstructed from the CPD-ALS and CPD-MWU decompositions. We used the reconstructed tensors to subtract the background from the original video frame, and show inversed images from those subtractions in Figure 3. Figure 3d shows the difference in the background subtraction inversed. (The inverse of the difference and subtraction images is used to make it easier to visually discern the differences.)

References

Acar, E., Dunlavy, D. M., and Kolda, T. G. A scalable optimization approach for fitting canonical tensor decompositions. *J. Chemometrics*, 25(2):67–86, 2011.



(a) Original video still

(b) Reconstructed from CPD-ALS

(c) Reconstructed from CPD-MWU

(d) Difference between CPD-ALS and CPD-MWU images

(e) Background subtraction using CPD-ALS reconstructed image

(f) Background subtraction using CPD-MWU reconstructed image

Figure 3: (a) Original video frame image and those reconstructed using (b) traditional CPD-ALS and (c) CPD-MWU algorithms. The CPD-MWU decomposition took 30% less time of the CPD-ALS decomposition. The fixed portions of the image are perfectly reconstructed, and portions with movement have minimal blurring, which can be observed in the background subtraction images (e) and (f). (d) shows the difference between the CPD-ALS and CPD-MWU reconstructed images.

Battaglino, C., Ballard, G., and Kolda, T. G. A practical randomized CP tensor decomposition. *SIAM J. Matrix Anal. Appl.*, 39(2):876–901, 2018. doi: 10.1137/17m1112303.

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadephia, PA, 2017.

Cesa-Bianchi, N. and Gábor, L. *Prediction, Learning, and Games*. Cambridge Univ. Press, Cambridge, UK, 2006.

Kiers, H. A. L., ten Berge, J. M. F., and Bro, R. PARAFAC2 - Part I. a direct fitting algorithm for the PARAFAC2 model. *J. Chemometrics*, 13(3-4):275—-294, 1999.

Tomasi, G. and Bro, R. A comparison of algorithms for fitting the PARAFAC model. *Comput. Statist. Data Anal.*, 50(7): 1700–1734, April 2006.

Wang, S., Gittens, A., and Mahoney, M. W. Sketched ridge regression: optimization perspective, statistical perspective, and model averaging. *The Journal of Machine Learning Research*, 18(1):8039–8088, 2017.

Woodruff, D. P. *Sketching As a Tool for Numerical Linear Algebra*, volume 10. Now, 2014.