A Matricization

A tensor $\mathcal{A}$ can be unfolded into a matrix A in various ways. We focus here on 2-way unfoldings that specify a proper partition of $\{1,\dots,K\} := [K] = \{m_1,\dots,m_p\} \cup \{n_1,\dots,n_q\}$, an integer-valued row index function r, and a column index function c, such that $\mathcal{A}_{i_1,\dots,i_K} \mapsto A_{\mathsf{r}(i_{m_1},\dots,i_{m_p}),\ \mathsf{c}(i_{n_1},\dots,i_{n_q})}$. This is simply a rearrangement of the entries in $\mathcal{A}$ hence preserves its Frobenius norm. We use the notation $\mathcal{A}_{(\mathsf{r},\mathsf{c})}$ for the 2-way unfolding under the index functions r and c.

The above tensor unfolding interacts conveniently with the mode-k multiplication, once we define a suitable matrix Kronecker product. Fix two row index functions r and $\hat{\mathsf{r}}$. The Kronecker product of p matrices $U \in \mathbb{R}^{\hat{d}_{m_1} \times d_{m_1}}, \dots, W \in \mathbb{R}^{\hat{d}_{m_p} \times d_{m_p}}$ is a matrix that has size $\prod_{k=1}^p \hat{d}_{m_k} \times \prod_{k=1}^p d_{m_k}$ and satisfies

$$(U \boxtimes \cdots \boxtimes W)_{\hat{\imath},i} = U_{\hat{\imath}_{m_1},i_{m_1}} \cdots W_{\hat{\imath}_{m_p},i_{m_p}}, \tag{16}$$

where $\hat{\imath} = \hat{\mathsf{r}}(\hat{\imath}_{m_1},\dots,\hat{\imath}_{m_p})$ and $i = \mathsf{r}(i_{m_1},\dots,i_{m_p})$.

Similar definitions can be made using two column index functions c and $\hat{\mathsf{c}}$. It is just algebra to verify that

$$(\mathcal{A} \times_1 U_1 \cdots \times_K U_K)_{(\hat{\mathsf{r}},\hat{\mathsf{c}})} = (U_{m_1} \boxtimes \cdots \boxtimes U_{m_p})\ \mathcal{A}_{(\mathsf{r},\mathsf{c})}\ (U_{n_1} \boxtimes \cdots \boxtimes U_{n_q})^\top, \tag{17}$$

where the first and second group of Kronecker product use $(\hat{\mathsf{r}},\mathsf{r})$ and $(\hat{\mathsf{c}},\mathsf{c})$ respectively.

Example 1 (Mode-k unfolding $\mathcal{A}_{(k)}$)**.** *To illustrate the above definition, let us consider the partition* $[K] = \{k\} \cup \{1,\dots,k-1,k+1,\dots K\}$ *and the index functions*

$$\mathsf{c}(i_1,\dots,i_{k-1},i_{k+1},\dots,i_K) = 1 + \sum_{j\neq k}(i_j - 1) \prod_{m>j,m\neq k} d_m,$$

and $\mathsf{r}(i_k) = i_k$*. This is called the mode-k unfolding, together with the notation* $\mathcal{A}_{(k)}$*. Here* $\boxtimes$ *reduces to the usual matrix Kronecker product, and* $(\mathcal{A} \times_k U)_{(k)} = U\mathcal{A}_{(k)}$*. For matrices, we simply have* $A_{(1)} = A$*,* $A_{(2)} = A^\top$*, and* $A \times_1 U \times_2 V = UAV^\top$*.*

Example 2 (Balanced mode-k unfolding $\mathcal{A}_{[k]}$)**.** *The mode-k unfolding above yields an extremely unbalanced matrix with size* $d_k \times \prod_{j\neq k} d_j$*. A more balanced unfolding is proposed in [31], consisting of the partition* $[K] = \{1,\dots,k\} \cup \{k+1,\dots K\}$ *and the index functions*

$$\mathsf{r}(i_1,\dots,i_k) = 1 + \sum_{j=1}^{k}(i_j - 1) \prod_{k\geq m\geq j+1} d_m$$
$$\mathsf{c}(i_{k+1},\dots,i_K) = 1 + \sum_{j=k+1}^{K}(i_j - 1) \prod_{m\geq j+1} d_m,$$

with $\hat{\mathsf{r}},\hat{\mathsf{c}}$ *similarly defined. The resulting matrix, denoted as* $\mathcal{A}_{[k]}$*, has size* $\prod_{j=1}^k d_j \times \prod_{j=k+1}^K d_j$*. For* $k = \lfloor K/2 \rfloor$*, the unfolding is more like a square matrix, which can be beneficial in completion tasks [31].*

B Approximating Tensor Spectral Norm

Let B_2^d be the Euclidean norm ball in a d dimensional space (d is a superscript in B_2^d). We can approximate B_2^d with a *polytope*, based on a celebrated result from convex geometry [24]: For any $d, n \geq 2$ we can find in polynomial time (at most) d^n points $\{\mathbf{p}_i\}$, such that their convex hull P^d satisfies

$$\mathsf{P}^d \subseteq \mathsf{B}_2^d \subseteq \tfrac{1}{c}\sqrt{\tfrac{d}{n \log d}}\mathsf{P}^d, \tag{18}$$

where c is some universal constant. For $n = 1$, we can simply take the unit ball of the 1-norm, denoted as B_1^d, and get a similar result

$$\mathsf{B}_1^d \subseteq \mathsf{B}_2^d \subseteq \sqrt{d}\mathsf{B}_1^d. \tag{19}$$

By counting the volume, it can be proved that the factors in (18) and (19) are the best possible respectively [25].

Specializing to the tensor spectral norm, we simply replace each Euclidean ball constraint with its polytopal approximation as suggested in (18) or (19), and evaluate the inner product at each of the vertices of the polytope. Since the matrix trace norm is tractable, we need only execute the polytopal approximation for the last $K-2$ balls, i.e., solving the approximation

$$\max_{\mathbf{u}_1 \in \mathsf{B}_2^{d_1}, \mathbf{u}_2 \in \mathsf{B}_2^{d_2}, \mathbf{u}_3 \in \mathsf{P}^{d_3}, \ldots, \mathbf{u}_K \in \mathsf{P}^{d_K}} \langle \mathcal{A}, \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_K \rangle . \tag{20}$$

So for each vertex $\mathbf{p}^3 \otimes \cdots \otimes \mathbf{p}^K$ (with $\prod_{k=3}^K (d_k)^n$ of them), we evaluate the matrix spectral norm $\|\mathcal{A} \times_3 \mathbf{p}^3 \cdots \times_K \mathbf{p}^K\|_2$ and pick the maximum. This yields the optimal solution for (20), which immediately translates to an $\alpha = O(\prod_{k=3}^K \sqrt{n d_k^{-1} \log d_k})$ approximate solution for the tensor spectral norm (6). If we set $n = 1$ and use (19), then $\alpha = \prod_{k=3}^K \sqrt{1/d_k}$ and only $\prod_{k=3}^K d_k$ matrix spectral norms need to be checked. The overall computational cost is $O(\prod_{k=1}^K d_k)$. It is also easy to reduce each factor d_k to the smaller constant $\mathrm{rank}(\mathcal{A}_{(k)})$, or simply $\mathrm{rank}(\mathcal{A})$. For $n \geq 2$ we get a $\log d_k$ factor improvement in the approximation guarantee, at the expense of a more complicated and costly implementation.

C Proofs omitted in Section 4

Theorem 1. *Let $\ell \geq 0$ be convex, smooth, and have bounded sublevel sets. Denote $f(\mathcal{W}) = \ell(\mathcal{W}) + \lambda \cdot \kappa(\mathcal{W})$. Suppose in each iteration t, we find $\mathcal{Z}_t$ that satisfies*

$$\kappa(\mathcal{Z}_t) \leq 1, \quad \langle \mathcal{Z}_t, \nabla \ell(\mathcal{W}_t) \rangle \geq \alpha \cdot \max_{\kappa(\mathcal{Z}) \leq 1} \langle \mathcal{Z}, \nabla \ell(\mathcal{W}_t) \rangle . \tag{21}$$

Then for all $\mathcal{W}$ and for all $t \geq 1$, running GCG *with $\eta_t = \frac{2}{t+2}$ leads to $f(\mathcal{W}_t) - \frac{f(\mathcal{W})}{\alpha} \leq \frac{2C}{t+3}$, where C is some constant that does not depend on t or α.*

Recall that the function ℓ is smooth if its gradient is Lipschitz continuous with respect to some norm $\|\cdot\|$, namely that for all $\mathcal{W}$ and $\mathcal{Z}$,

$$\ell(\mathcal{Z}) \leq \ell(\mathcal{W}) + \langle \mathcal{Z} - \mathcal{W}, \nabla \ell(\mathcal{W}) \rangle + \frac{L}{2} \|\mathcal{Z} - \mathcal{W}\|^2, \tag{22}$$

for some constant $L := L_{\|\cdot\|} \geq 0$. The least squares loss in fact satisfies (22) with equality and $L = 2$.

Proof. Let $\mathcal{W}$ be arbitrary and $s = \kappa(\mathcal{W})$. Let $\tilde{\mathcal{W}}_{t+1}$ be the output of GCG at iteration $t+1$ and $\mathcal{W}_{t+1}$ be the

improved iterate after local search. The following chain of inequalities can be easily verified:

$$
\begin{aligned}
f(\mathcal{W}_{t+1}) &\le f(\tilde{\mathcal{W}}_{t+1}) \\
&= \ell(\tilde{\mathcal{W}}_{t+1}) + \lambda \cdot \kappa(\tilde{\mathcal{W}}_{t+1}) \\
&= \ell\left((1-\eta_t)\mathcal{W}_t + \eta_t\beta_t\mathcal{Z}_t\right) + \lambda \cdot \kappa((1-\eta_t)\mathcal{W}_t + \eta_t\beta_t\mathcal{Z}_t) && \text{(definition of } \tilde{\mathcal{W}}_{t+1}) \\
&\le \ell\left((1-\eta_t)\mathcal{W}_t + \eta_t\beta_t\mathcal{Z}_t\right) + \lambda(1-\eta_t)\kappa(\mathcal{W}_t) + \lambda\eta_t\beta_t\kappa(\mathcal{Z}_t) && \text{(sublinearity of } \kappa) \\
&\le \ell\left((1-\eta_t)\mathcal{W}_t + \eta_t\beta_t\mathcal{Z}_t\right) + \lambda(1-\eta_t)\kappa(\mathcal{W}_t) + \lambda\eta_t\beta_t && \text{(definition of } \mathcal{Z}_t) \\
&\le \ell\left((1-\eta_t)\mathcal{W}_t + \eta_t\frac{s}{\alpha}\mathcal{Z}_t\right) + \lambda(1-\eta_t)\kappa(\mathcal{W}_t) + \lambda\eta_t\frac{s}{\alpha} && \text{(definition of } \beta_t) \\
&\le f(\mathcal{W}_t) + \eta_t\left\langle \frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t, \nabla\ell(\mathcal{W}_t)\right\rangle + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 - \lambda\eta_t\kappa(\mathcal{W}_t) + \lambda\eta_t\frac{s}{\alpha} && \text{(inequality (22))} \\
&\le \min_{\mathcal{Z}:\kappa(\mathcal{Z})\le 1} f(\mathcal{W}_t) + \eta_t\left\langle s\mathcal{Z} - \mathcal{W}_t, \nabla\ell(\mathcal{W}_t)\right\rangle + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 - \lambda\eta_t\kappa(\mathcal{W}_t) + \lambda\eta_t\frac{s}{\alpha} && \text{(definition of } \mathcal{Z}_t) \\
&= \min_{\mathcal{Z}:\kappa(\mathcal{Z})\le s} f(\mathcal{W}_t) + \eta_t\left\langle \mathcal{Z} - \mathcal{W}_t, \nabla\ell(\mathcal{W}_t)\right\rangle + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 - \lambda\eta_t\kappa(\mathcal{W}_t) + \lambda\eta_t\frac{s}{\alpha} && \text{(homogeneity of } \kappa) \\
&\le \min_{\mathcal{Z}:\kappa(\mathcal{Z})\le s} f(\mathcal{W}_t) + \eta_t(\ell(\mathcal{Z}) - \ell(\mathcal{W}_t)) + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 - \lambda\eta_t\cdot\kappa(\mathcal{W}_t) + \lambda\eta_t\frac{s}{\alpha} && \text{(convexity of } \ell) \\
&= (1-\eta_t)f(\mathcal{W}_t) + \eta_t\min_{\mathcal{Z}:\kappa(\mathcal{Z})\le s}(\ell(\mathcal{Z}) + \lambda\cdot\frac{s}{\alpha}) + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 \\
&\le (1-\eta_t)f(\mathcal{W}_t) + \eta_t\frac{f(\mathcal{W})}{\alpha} + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2 && (\ell \ge 0 \text{ and } \alpha \in (0,1]).
\end{aligned}
$$

Therefore,

$$
f(\mathcal{W}_{t+1}) - \frac{f(\mathcal{W})}{\alpha} \le (1-\eta_t)\left(f(\mathcal{W}_t) - \frac{f(\mathcal{W})}{\alpha}\right) + \frac{L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2}{2}\eta_t^2.
$$

Recall that $\eta_t = \frac{2}{t+2}$. An easy induction argument establishes that

$$
f(\mathcal{W}_{t+1}) - \frac{f(\mathcal{W})}{\alpha} \le \frac{2C}{t+3},
$$

where $C := \sup_t L\left\|\frac{s}{\alpha}\mathcal{Z}_t - \mathcal{W}_t\right\|^2 \le 2L(\kappa^2(\mathcal{W})/\alpha^2 + D^2) < \infty$, since $\mathcal{W}_t$ is in the sublevel set of $\{\mathcal{Z} : f(\mathcal{Z}) \le f(\mathcal{W}_1)\}$, whose radius is assumed to be bounded by D. □

Theorem 2. *Fix $\mathcal{A} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ and $t \ge \prod_{k=1}^K d_k$, then*

$$
\begin{aligned}
\|\mathcal{A}\|_{\mathrm{tr}} &= \min\left\{\Sigma_{i=1}^t \|\mathbf{u}_i\|_2\cdots\|\mathbf{z}_i\|_2\right\} && (23)\\
&= \min\left\{\tfrac{1}{K}\Sigma_{i=1}^t \|\mathbf{u}_i\|_2^K + \cdots + \|\mathbf{z}_i\|_2^K\right\}, && (24)
\end{aligned}
$$

where the minimum is taken w.r.t. all factorizations $\mathcal{A} = \sum_{i=1}^t \mathbf{u}_i\otimes\cdots\otimes\mathbf{z}_i$, $\mathbf{u}_i \in \mathbb{R}^{d_1}, \ldots, \mathbf{z}_i \in \mathbb{R}^{d_K}$.

Proof. We first note that the atomic set A in (4) is compact, so is its convex hull $\mathrm{conv}(\mathsf{A})$. Moreover, A is connected. Recall that the trace norm is defined in (5) via the gauge function κ. Since $\mathrm{conv}(\mathsf{A})$ is compact with $\mathbf{0}$ in its interior, we know the infimum in (5) is attained. Thus there exist $\rho \ge 0$ and $\mathcal{C} \in \mathrm{conv}(\mathsf{A})$ so that $\mathcal{A} = \rho\mathcal{C}$ and $\|\mathcal{A}\|_{\mathrm{tr}} = \kappa(\mathcal{A}) = \rho$. Applying Caratheodory's theorem we know $\mathcal{C} = \sum_{i=1}^t \sigma_i\hat{\mathbf{u}}_i\otimes\cdots\otimes\hat{\mathbf{z}}_i$ for some $\sigma_i \ge 0$, $\sum_i\sigma_i = 1$, $\hat{\mathbf{u}}_i\otimes\cdots\otimes\hat{\mathbf{z}}_i \in \mathsf{A}$ and $t \le \prod_{k=1}^K d_k$. Let $\mathbf{u}_i = \sqrt[K]{\rho\sigma_i}\hat{\mathbf{u}}_i, \cdots, \mathbf{z}_i = \sqrt[K]{\rho\sigma_i}\hat{\mathbf{z}}_i$ we know $\|\mathcal{A}\|_{\mathrm{tr}}$ is at least the right-hand side of (24).

On the other hand, for any $\mathcal{A} = \sum_{i=1}^t \mathbf{u}_i\otimes\cdots\otimes\mathbf{z}_i$, denoting $\sigma_i = \|\mathbf{u}_i\|_2\cdots\|\mathbf{z}_i\|_2$ ($\neq 0$ w.l.o.g.), we have $\mathcal{A} = (\sum_i\sigma_i)\cdot\sum_{i=1}^t \frac{\sigma_i}{\sum_{j=1}^t\sigma_j}\mathcal{A}_i$, where $\mathcal{A}_i := \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2}\otimes\cdots\otimes\frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} \in \mathsf{A}$. Thus, appealing to the definition (5) we know $\|\mathcal{A}\|_{\mathrm{tr}} \le \sum_i\sigma_i$, i.e. (23) holds with $\le$.

To complete the proof, we apply the arithmetic-geometric mean inequality:

$$\sum\nolimits_i \|\mathbf{u}_i\|_2 \cdots \|\mathbf{z}_i\|_2 \leq \tfrac{1}{K} \sum\nolimits_i \|\mathbf{u}_i\|_2^K + \cdots + \|\mathbf{z}_i\|_2^K.$$

(Of course, other elementary symmetric functions can be similarly used in Theorem 2.) □

D Comparison with alternative variational forms

We compare the variational forms in Theorem 2 to some existing ones in the tensor literature. Firstly, the regularization function

$$\sum\nolimits_i \|\mathbf{u}_i\|_2^2 + \cdots + \|\mathbf{z}_i\|_2^2, \tag{25}$$

is extensively used in finding CP decompositions, since otherwise the factors could blow up in a way that still maintains their sum, the so-called degeneracy problem [9]. A second reason for employing (25) is that it adds strict convexity w.r.t. each factor hence guarantees the convergence of block coordinate ascent. However, both reasons to promote (25), albeit valid, are weak; there are certainly other, perhaps even better, candidate regularizations. For instance, (24) enjoys both properties, with the additional equivalence to the trace norm, which potentially could lead to a low rank solution. The second variational form, appeared in [13], is

$$\|\mathcal{C}\|_{\mathsf{F}}^2 + \|U\|_{\mathsf{F}}^2 + \cdots + \|Z\|_{\mathsf{F}}^2, \tag{26}$$

where $\mathcal{A} = \mathcal{C} \times_1 U \cdots \times_K Z$ is the Tucker decomposition. [13] used (26) to avoid the scaling ambiguity—a weak motivation for the particular form (26) indeed. Let us show that *neither* (25) *nor* (26) is equivalent to the trace norm. From this regard, the variational forms in Theorem 2 are advantageous and perhaps should be favored more often in practice.

Example 3. *We first prove* (25) *is not equivalent to the trace norm. Let* $\mathcal{A} = \sigma \mathbf{u} \otimes \cdots \otimes \mathbf{z}$ *be a rank-1 tensor with* $\sigma > 0$ *and* $\|\mathbf{u}\|_2 = \cdots = \|\mathbf{z}\|_2 = 1$. *It is easy to see that* $\|\mathcal{A}\|_{\mathrm{tr}} = \sigma$. *Consider the function:*

$$f(\mathcal{A}) := \min \left\{ \sum\nolimits_i \|\mathbf{u}_i\|_2^2 + \cdots + \|\mathbf{z}_i\|_2^2 \right\},$$

where the minimum is taken w.r.t. to all factorizations $\mathcal{A} = \sum_i \mathbf{u}_i \otimes \cdots \otimes \mathbf{z}_i$. *Clearly,*

$$f(\mathcal{A}) \leq K \cdot \sigma^{2/K}.$$

Choose an appropriate $\sigma > 1$ *we thus have* $f(\mathcal{A}) < \|\mathcal{A}\|_{\mathrm{tr}}$. *Of course, for any positive constant* c, *we can choose appropriate* σ *such that* $c \cdot f(\mathcal{A}) < \|\mathcal{A}\|_{\mathrm{tr}}$. *Thus* (25) *is not proportional to the trace norm.*

For the function (26), *we similarly define*

$$g(\mathcal{A}) := \inf \left\{ \|\mathcal{C}\|_{\mathsf{F}}^2 + \|U\|_{\mathsf{F}}^2 + \cdots + \|Z\|_{\mathsf{F}}^2 \right\},$$

where the infimum is taken w.r.t. all Tucker decompositions $\mathcal{A} = \mathcal{C} \times_1 U \cdots \times_K Z$. *This removes the dependence of* (26) *on a particular Tucker decomposition (which may not be unique). Consider the same rank-1 tensor* $\mathcal{A}$ *as above, we have* $g(\mathcal{A}) \leq (K+1)\sigma^{2/(K+1)}$ *while* $\|\mathcal{A}\|_{\mathrm{tr}} = \sigma$. *Again, choosing* σ *large we have* $c \cdot g(\mathcal{A}) < \|\mathcal{A}\|_{\mathrm{tr}}$ *for any positive constant* c. *Note that even for* $K = 2$, $g(\mathcal{A})$ *is not proportional to the trace norm.*

E Efficient gradient computation

An efficient implementation for the surrogate problem (14) using state-of-the-art solvers (e.g. L-BFGS) relies on the efficient computation of the gradient of ℓ. For simplicity, consider $t = 1$ hence $\mathcal{W} = \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_K$. The idea generalizes straightforwardly to all t. Let $\mathcal{G} = \nabla \ell(\mathcal{W})$. Using the chain rule,

$$\tfrac{\partial \ell}{\partial \mathbf{u}_k}(\mathcal{W}) = \mathcal{G} \times_1 \mathbf{u}_1^\top \ldots \times_{k-1} \mathbf{u}_{k-1}^\top \times_{k+1} \mathbf{u}_{k+1}^\top \ldots \times_K \mathbf{u}_K^\top.$$

The chain of product on the right-hand side costs $O(\prod_k d_k)$, thus a naive implementation for all factors would cost $O(K \prod_k d_k)$. Fortunately, using the identity in (17) we can reduce the cost by a factor of K. Define the forward and backward accumulators

$$\mathcal{F}_k := \mathcal{G} \times_1 \mathbf{u}_1^\top \ldots \times_{k-1} \mathbf{u}_{k-1}^\top,\ \mathcal{B}_k := \mathbf{u}_{k+1} \otimes \ldots \otimes \mathbf{u}_K,$$

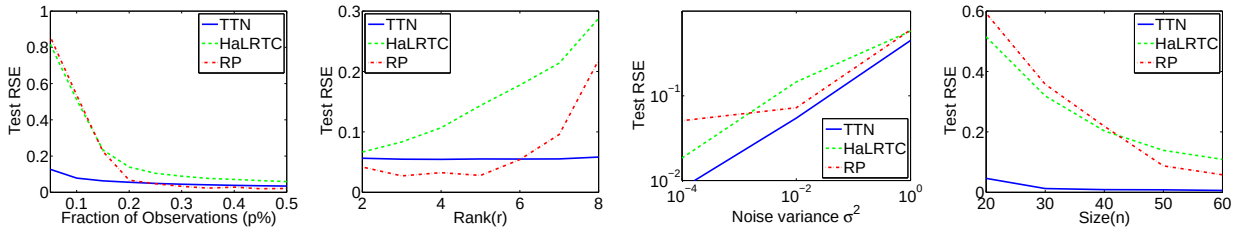
with $\mathcal{F}_1 := \mathcal{G}$ and $\mathcal{B}_K := 1$. Then we have

$$\frac{\partial \ell}{\partial \mathbf{u}_k}(\mathcal{W}) = \mathcal{F}_k \times_{k+1} \mathbf{u}_{k+1}^\top \ldots \times_K \mathbf{u}_K^\top = (\mathcal{F}_k)_{(k)} \mathcal{B}_k.$$

So we need only compute $\{\mathcal{F}_k, \mathcal{B}_k\}$, costing $O(\prod_k d_k)$. Since for all k the multiplication $(\mathcal{F}_k)_{(k)}\mathcal{B}_k$ costs $O(\prod_{j \geq k} d_j)$, the overall time and space costs are both $O(\prod_k d_k)$. Clearly the computational savings are possible due to our explicit low-rank representation, which is not available in other matricization approaches.

F Comparison on completing tensors with low-rank Tucker decompositions

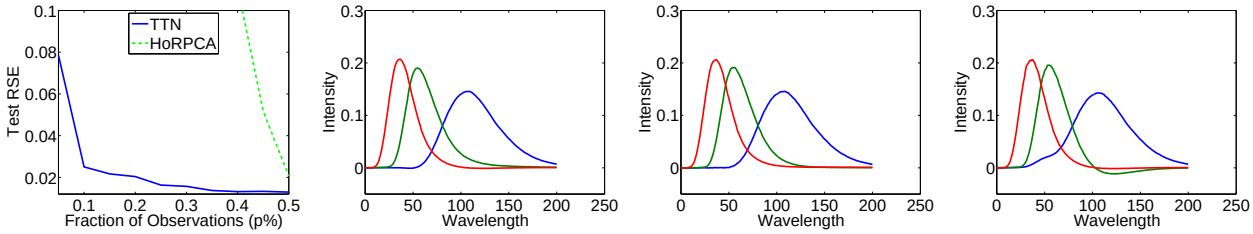
We repeated all comparisons conducted for low CP rank on low Tucker rank: $\mathcal{Z}^0 = \mathcal{S} \times_1 U_1 \times_2 U_2 \times_3 U_3$, where $\mathcal{S} \in \mathbb{R}^{r \times r \times r}$ and $U_i \in \mathbb{R}^{n \times r}$. Our setting here is exactly the same as Section 6.1. Again, all entries of $\mathcal{S}$ and U_i were drawn i.i.d. from a unit normal. We set the default $p = 20$, $\sigma = 0.1$, $r = 5$ (hence CP rank ≤ 25), and $n = 50$. Figure 7 shows that even in this case TTN still outperforms HaLRTC and RpLRTC (abbreviated as RP).



(a) RSE vs p (train%) (b) RSE vs r (rank) (c) RSE vs σ (noise std) (d) RSE vs n (tensor size)

Figure 7: Test RSE on synthetic data generated by Tucker decomposition with rank $r \times r \times r$.

G Robust PCA (RPCA)



(a) RSE vs p (b) Loading for mode 2 by ALS on noiseless data (c) Loading for mode 2 recovered by TTN (d) Loading for mode 2 recovered by HoRPCA

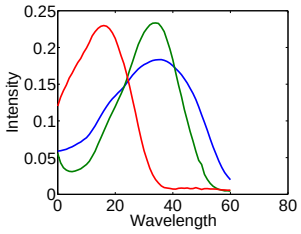
Figure 8: RPCA on amino acid data. $p = 40$ in c-d.

Higher-order RPCA (HoRPCA), introduced in [16], solves $\|\mathcal{W} - \mathcal{Z}\|_1 + \lambda\Omega(\mathcal{W})$, where $\Omega(\mathcal{W})$ is, say the sum of matrix trace norms of all mode unfoldings. The ℓ_1 loss used here aims to instill robustness to (large) outliers. To apply GCG to our proposed TTN as Ω, we smoothed the ℓ_1 loss using a quadratic prox-function as in [39].

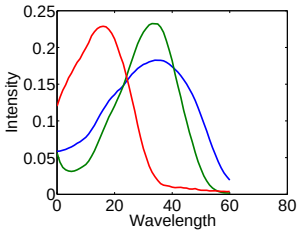
Amino acid fluorescence data. This dataset of 3-mode tensor records the fluorescence intensity of 5 solutions (mode 1), with 201 intervals of emission wavelength (mode 2) and 61 intervals of excitation wavelength (mode 3) [40]. Since the solutions contain three amino acids, the tensor has approximately low rank of 3. To mimic scatter outliers in chmometrics like [16], we randomly selected 10% entries on which i.i.d. corruptions uniformly distributed in $[-100, 100]$ were added. Like above, we observe only a fraction, p%, of the entries.

Figure 8(a) shows the RSE on the remaining $(90 - p)$% entries, measured on the tensors recovered by using TTN and HoRPCA. Over a wide range of p, TTN achieves significantly lower RSE. For a more intuitive illustration,

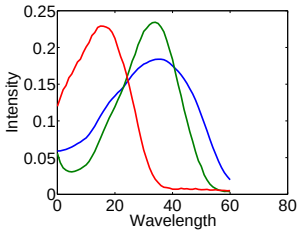
we plot in Figure 8(b) the original loading (CP factor) for mode 2, as well as the loadings recovered by TTN and HoRPCA in Figures 8(c) and 8(d), respectively. Clearly TTN discovers a more faithful reconstruction of the CP factors in the presence of noise. Loadings for mode 3 are provided in Figure 9.



(a) Original loading for mode 3 by ALS on noiseless complete data



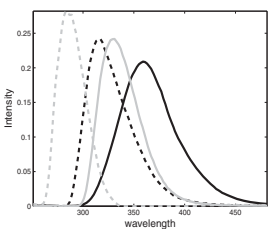
(b) Loading for mode 3 recovered by TTN

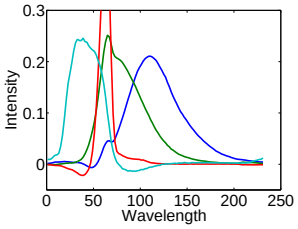


(c) Loading for mode 3 recovered by HoRPCA

Figure 9: Loadings for mode 3 (excitation) with RPCA on amino acid data. $p = 40$ in b-c.

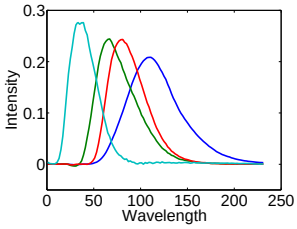
We also experimented on another fluorescence dataset: Dorrit. It is a $23 \times 116 \times 18$ tensor whose modes have the same meanings as in the amino acid data. However, it already contains scattering noise hence we no longer corrupted it. In Figure 10(c) and 10(d), we plot the loading for mode 2 recovered by TTN and HoRPCA, with comparison to that obtained from the noisy raw tensor (Figure 10(b)). All methods used the full tensor with no subsampling. Referring to the pure component loadings copied to Figure 10(a) from [41, Figure 2], one can observe that TTN attains a higher level of faithfulness than HoRPCA.



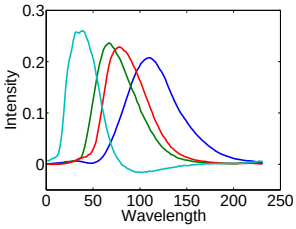


(a) Loading for mode 2 with pure component

(b) Loading for mode 2 using ALS on noisy raw input

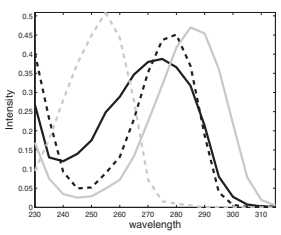


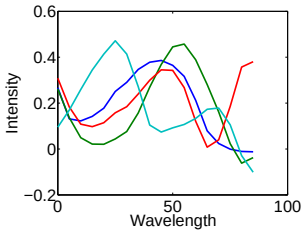
(c) Loading for mode 2 recovered by TTN



(d) Loading for mode 2 recovered by HoRPCA

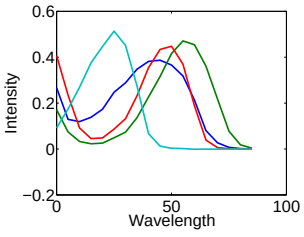
Figure 10: Loadings for mode 2 (emission) with RPCA on Dorrit data. $p = 100$ in c-d.



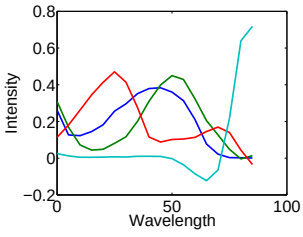


(a) Loading for mode 3 with pure component

(b) Loading for mode 3 using ALS on noisy raw input



(c) Loading for mode 3 recovered by TTN



(d) Loading for mode 3 recovered by HoRPCA

Figure 11: Loadings for mode 3 (excitation) with RPCA on Dorrit data. $p = 100$ in c-d.