TenIPS: Inverse Propensity Sampling for Tensor Completion

Chengrun Yang CY438@CORNELL.EDU
Lijun Ding LD446@CORNELL.EDU
Ziyang Wu ZW287@CORNELL.EDU
Madeleine Udell UDELL@CORNELL.EDU
Cornell University, USA

Abstract

Tensors are widely used to represent multiway arrays of data. The recovery of missing entries in a tensor has been extensively studied, generally under the assumption that entries are missing completely at random (MCAR). However, in most practical settings, observations are missing not at random (MNAR): the probability that a given entry is observed (also called the propensity) may depend on other entries in the tensor or even on the value of the missing entry. In this paper, we study the problem of completing a partially observed tensor with MNAR observations, without prior information about the propensities. To complete the tensor, we assume that both the original tensor and the tensor of propensities have low multilinear rank. The algorithm first estimates the propensities using a convex relaxation and then predicts missing values using a higher-order SVD approach, reweighting the observed tensor by the inverse propensities. We provide finite-sample error bounds on the resulting complete tensor. Numerical experiments demonstrate the effectiveness of our approach.

1. Problem setting

In this paper, we study the following problem: given a partially observed tensor $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with MNAR entries, how can we recover its missing values?

Throughout the paper, we denote the true order-N tensor we want to complete as $\mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$. For each $n \in [N]$, we suppose $I_n \leq I_{(-n)} := \prod_{m \in [N], m \neq n} I_m$ for cleanliness. We assume there exists a propensity tensor $\mathcal{P} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, such that $\mathcal{B}_{i_1 \cdots i_N}$ is observed with probability $\mathcal{P}_{i_1 \cdots i_N}$. We observe the entries without noise: with the binary mask tensor $\Omega \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, in which each entry with value 1 indicates the corresponding entry in $\mathcal{B}$ is observed and 0 otherwise, we observe $\mathcal{B}_{\text{obs}} = \mathcal{B} \odot \Omega$.

A tensor $\mathcal{B}$ has *multilinear rank* $(r_1^{\text{true}}, \ldots, r_N^{\text{true}})$ if r_n^{true} is the rank of $\mathcal{B}^{(n)}$. For any $n \in [N]$, $r_n^{\text{true}} \leq I_n$. We can write the *Tucker decomposition* of the tensor $\mathcal{B}$ as $\mathcal{B} = \mathcal{G}^{\text{true}} \times_1 U_1^{\text{true}} \times_2 \cdots \times_N U_N^{\text{true}}$, with *core tensor* $\mathcal{G}^{\text{true}} \in \mathbb{R}^{r_1^{\text{true}} \times \cdots \times r_N^{\text{true}}}$ and column orthonormal *factor matrices* $U_n^{\text{true}} \in \mathbb{R}^{I_n \times r_n^{\text{true}}}$ for $n \in [N]$.

We seek a *fixed-rank approximation* of $\mathcal{B}$ by a tensor with multilinear rank $(r_1, \ldots, r_N)$: we want to find a core tensor $\mathcal{W} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_N}$ and N factor matrices $Q_n \in \mathbb{R}^{I_n \times r_n}$, $n \in [N]$ with orthonormal columns, such that $\mathcal{B} \approx \mathcal{W} \times_1 Q_1 \times_2 \cdots \times_N Q_N$. We generally seek a low multilinear rank decomposition with $r_n < I_n$.

There have been many other approaches to the tensor completion problem under different settings, and a literature review can be found in Appendix A.

2. Notations

We defer the more standard notations to Appendix B, and emphasize here the ones defined specifically in this paper.

Square unfoldings Extending the notation of [20], with a matrix $A \in \mathbb{R}^{I \times J}$ and integers I', J' satisfying $IJ = I'J'$, $\texttt{reshape}(A, I', J')$ gives a matrix $A' \in \mathbb{R}^{I' \times J'}$ with entries taken columnwise from A. Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, we can partition the indices of its N modes into two sets, S and S^C, and permute the order of the N modes by a map $c_S : [N] \to [N]$ that satisfies $\{c_S^{-1}(1), c_S^{-1}(2), \ldots, c_S^{-1}(|S|)\} = S$. Then we get an order-permuted tensor ${}^{c_S}\mathcal{X} \in \mathbb{R}^{I_{c_S^{-1}(1)} \times \cdots \times I_{c_S^{-1}(N)}}$, and the *set-$S$ unfolding* $\mathcal{X}_S := \texttt{reshape}({}^{c_S}\mathcal{X}^{(1)}, \prod_{n \in S} I_n, \prod_{n \in [N] \backslash S} I_n)$. To make $\mathcal{X}_S$ as square as possible, we want to minimize $\left| \prod_{n \in S} I_n - \prod_{n \in [N] \backslash S} I_n \right|$. In this case,

$$\mathcal{X}_\square := \texttt{reshape}({}^{c_\square}\mathcal{X}^{(1)}, \prod_{n \in S_\square} I_n, \prod_{n \in [N] \backslash S_\square} I_n),$$

in which

$$S_\square = \operatorname*{arg\,min}_{S \subset [N]} \left| \prod_{n \in S} I_n - \prod_{n \in [N] \backslash S} I_n \right|.$$

We call $\mathcal{X}_\square$ the *square unfolding*, and $S_\square$ the *square set* of $\mathcal{X}$.

Dimensions of unfoldings For brevity, we denote $I_S := \prod_{n \in S} I_n$, $I_{S^C} := \prod_{n \in [N] \backslash S} I_n$, $I_\square := \prod_{n \in S_\square} I_n$, $I_{\square^C} := \prod_{n \in [N] \backslash S_\square} I_n$, $I_{(-n)} := \prod_{m \in [N], m \neq n} I_m$. Thus $I_{[N]} = \prod_{n \in [N]} I_n = I_S \cdot I_{S^C} = I_\square \cdot I_{\square^C} = I_n \cdot I_{(-n)}$.

3. Methodology

Our algorithm proceeds in two steps. First, we estimate the propensities by Algorithm 1 or 3, both use a Bernoulli maximum likelihood estimator for 1-bit matrix completion [7] to estimate the propensities from the mask tensor Ω, aiming to recover propensities that come from the low rank parameters. Algorithm 1 explicitly requires the propensities to be neither too large or too small. As a gradient-descent-based substitute, Algorithm 3 in Appendix C does not require the associated tuning parameters, but empirically returns a good solution if the true propensity tensor $\mathcal{P}$ has this property. With the estimated propensity tensor $\widehat{\mathcal{P}}$, we estimate the data tensor $\mathcal{B}$ by Algorithm 2, a procedure that only requires a Tucker decomposition on the propensity-reweighted observations.

Our propensity estimation uses the observation model of 1-bit matrix completion. Each entry of $\mathcal{P}$ comes from applying a non-decreasing and differentiable link function $\sigma : \mathbb{R} \to [0, 1]$ to the corresponding entry of a parameter tensor $\mathcal{A}$, which we are trying to solve. An instance is the logistic function $\sigma(x) = 1/(1 + e^{-x})$. We assume $\mathcal{A}$ has low multilinear rank. In Algorithm 1, $\mathcal{A}_\square$ is low-rank from Lemma 1. We also assume an upper bound on the nuclear norm of $\mathcal{A}_\square$, a convex surrogate for its low-rank property. Algorithm 1 can be implemented by the proximal-proximal-gradient method [23] (which we call $\texttt{prox-prox}$) or the proximal alternating gradient descent. In Section 4, we will show that on a square tensor, the square unfolding achieves the smallest upper bound for propensity estimation among all possible unfoldings.

Algorithm 2 completes the observed data tensor $\mathcal{B}_{\text{obs}}$ by HOSVD on its entrywise inverse propensity reweighting $\bar{\mathcal{X}}(\mathcal{P})$, as defined in Line 2. The input propensity tensor can be either true

Algorithm 1 Propensity estimation (convex and provable)

```
Input: partially observed tensor $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, thresholds $\tau$, $\gamma$, link function $\sigma$
Output: Estimated propensity tensor $\widehat{\mathcal{P}}$
$\Omega, S_\square \leftarrow$ mask tensor and square set of $\mathcal{B}_{\text{obs}}$
$\widehat{\mathcal{A}}_\square \leftarrow \underset{\Gamma\in\mathcal{S}_{\tau,\gamma}}{\operatorname{argmax}} \sum_{i=1}^{I_\square}\sum_{j=1}^{I_{\square^C}} [(\Omega_\square)_{i,j}\log\sigma(\Gamma_{i,j}) + (1-(\Omega_\square)_{i,j})\log(1-\sigma(\Gamma_{i,j}))]$,
    where $\mathcal{S}_{\tau,\gamma} = \{\Gamma \in \mathbb{R}^{I_\square\times I_{\square^C}} : \|\Gamma\|_\star \le \tau\sqrt{I_{[N]}}, \|\Gamma\|_{\max} \le \gamma\}$.
$\widehat{\mathcal{P}} \leftarrow \sigma(\widehat{\mathcal{A}})$
return $\widehat{\mathcal{P}}$
```

Algorithm 2 TenIPS: TENsor completion by Inverse Propensity Sampling

```
Input: partially observed tensor $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, propensity tensor $\mathcal{P}$, target rank
    $(r_1, r_2, \cdots, r_N)$
Output: estimated tensor $\widehat{\mathcal{X}}(\mathcal{P})$
$\Omega \leftarrow$ mask set of $\mathcal{B}_{\text{obs}}$
$\bar{\mathcal{X}}(\mathcal{P}) \leftarrow \sum_{(i_1,i_2,\ldots,i_N)\in\Omega} \frac{1}{\mathcal{P}_{i_1,i_2,\ldots,i_N}} \mathcal{B}_{\text{obs}} \odot \mathcal{E}(i_1, i_2, \ldots, i_N)$
for $n = 1, 2, \ldots, N$ do                                        ▷ Recover factors
    $Q_n(\mathcal{P}) \leftarrow$ left $r_n$ singular vectors of $\bar{\mathcal{X}}(\mathcal{P})^{(n)}$
$\mathcal{W} \leftarrow \bar{\mathcal{X}}(\mathcal{P}) \times_1 Q_1(\mathcal{P})^\top \times_2 \cdots \times_N Q_N(\mathcal{P})^\top$      ▷ Recover core
$\widehat{\mathcal{X}}(\mathcal{P}) \leftarrow \mathcal{W}(\mathcal{P}) \times_1 Q_1(\mathcal{P}) \times_2 \cdots \times_N Q_N(\mathcal{P})$
return $\widehat{\mathcal{X}}(\mathcal{P})$
```

($\mathcal{P}$) or estimated ($\widehat{\mathcal{P}}$). With estimated propensity $\widehat{\mathcal{P}}$, we get $\widehat{\mathcal{X}}(\widehat{\mathcal{P}})$ instead of $\widehat{\mathcal{X}}(\mathcal{P})$. We analyze its estimation error from $\mathcal{B}$ in Appendix F.

4. Error analysis

To bound the relative estimation error $\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_{\text{F}}/\|\mathcal{B}\|_{\text{F}}$, we first bound the error in the propensity estimates in Algorithm 1, and then consider how this error propagates into the error of our final tensor estimate in Algorithm 2. Theorem 3 shows the optimality of the square unfolding for propensity estimation; Theorem 4 presents a special case of our bound on the tensor completion error with estimated propensities, with the full version as Appendix E, Theorem 5. We defer their proofs to Appendix D and F.

4.1. Error in propensity estimates (Algorithm 1)

We first show a corollary of [20, Lemma 6 (2) and Lemma 7] that bounds the rank of an unfolding.

Lemma 1 *Suppose $\mathcal{X}$ has Tucker decomposition $\mathcal{X} = \mathcal{C} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_N U_N$, where $\mathcal{C} \in \mathbb{R}^{r_1^{\text{true}}\times r_2^{\text{true}}\times\cdots\times r_N^{\text{true}}}$ and $U_n \in \mathbb{R}^{I_n\times r_n^{\text{true}}}$ for $n \in [N]$. Given $S \subset [N]$, $\mathcal{X}_S = \underset{j\in S}{\otimes} U_j \cdot \mathcal{C}_S \cdot \left(\underset{j\in[N]\setminus S}{\otimes} U_j\right)^\top$, and thus* $\operatorname{rank}(\mathcal{X}_S) \le \min\left\{\prod_{n\in S} r_n^{\text{true}}, \prod_{n\in[N]\setminus S} r_n^{\text{true}}\right\}$.

As a corollary of [7, Lemma 1] and [18, Theorem 2], we have Lemma 2 for the Frobenius norm error of the propensity tensor estimate.

Lemma 2 *Assume that $\mathcal{P} = \sigma(\mathcal{A})$. Given a set $S \subset [N]$, together with the following assumptions:*
A1. *$\mathcal{A}_S$ has bounded nuclear norm: there exists a constant $\theta > 0$ such that $\|\mathcal{A}_S\|_\star \leq \theta\sqrt{I_{[N]}}$.*
A2. *Entries of $\mathcal{A}$ have bounded absolute value: there exists a constant $\alpha > 0$ such that $\|\mathcal{A}\|_{\max} \leq \alpha$.*
Suppose we run Algorithm 1 with thresholds satisfying $\tau \geq \theta$ and $\gamma \geq \alpha$ to obtain an estimate $\widehat{\mathcal{P}}$ of $\mathcal{P}$. With $L_\gamma := \sup_{x\in[-\gamma,\gamma]} \frac{|\sigma'(x)|}{\sigma(x)(1-\sigma(x))}$, there exists a universal constant $C > 0$ such that if $I_S + I_{S^C} \geq C$, with probability at least $1 - \frac{C}{I_S+I_{S^C}}$, the estimation error $\frac{1}{I_{[N]}}\|\widehat{\mathcal{P}} - \mathcal{P}\|_{\mathrm{F}}^2 \leq 4eL_\gamma\tau\Big(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}}\Big)$.

In the simplest case, when N is even and $I_1 = \ldots = I_N = I$, the propensity estimation upper bound above is $O(I^{N/4})$. Theorem 3 then shows that the square unfolding achieves the smallest upper bound for the propensity estimation error among all possible unfoldings sets S.

Theorem 3 *Instate the same conditions as Lemma 2, and further assume that there exists a constant $c > 0$ such that $r_n^{\text{true}} \leq cI_n$, $\forall n \in [N]$. Then $S = S_\square$ gives the smallest upper bound (RHS of the result in Lemma 2) on the propensity estimation error $\|\widehat{\mathcal{P}} - \mathcal{P}\|_{\mathrm{F}}^2$ among all unfolding sets $S \subset [N]$.*

4.2. Error in tensor completion (Algorithm 2): special case

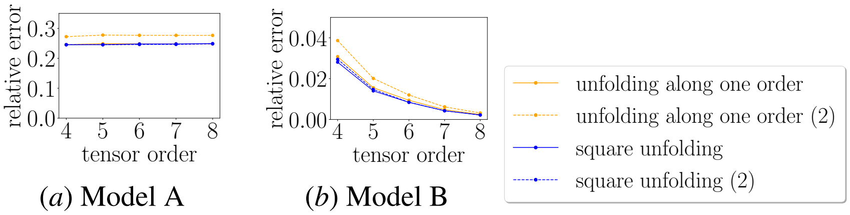
In Theorem 4, we present a special case of our bound for the recovery error on a cubical tensor with equal multilinear rank. This bound is dominated by the error from the matrix Bernstein inequality [26] on each of the N unfoldings, and asymptotically goes to 0 when the tensor size $I \to \infty$. Note that our full theorem applies to any tensor; we defer the formal statement to Appendix E, Theorem 5.

Theorem 4 *Consider an order-N cubical tensor $\mathcal{B}$ with size $I_1 = \cdots = I_N = I$ and multilinear rank $r_1^{\text{true}} = \cdots = r_N^{\text{true}} = r < I$, and two order-$N$ cubical tensors $\mathcal{P}$ and $\mathcal{A}$ with the same shape as $\mathcal{B}$. Each entry of $\mathcal{B}$ is observed with probability from the corresponding entry of $\mathcal{P}$. Assume: (1) $I \geq rN\log I$; (2) there exist constants $\psi, \alpha \in (0,\infty)$ such that $\|\mathcal{A}\|_{\max} \leq \alpha$, $\|\mathcal{B}\|_{\max} = \psi$; and (3) for each $n \in [N]$, the condition number $\frac{\sigma_1(\mathcal{B}^{(n)})}{\sigma_r(\mathcal{B}^{(n)})} \leq \kappa$ is a constant independent of tensor sizes and dimensions. Then under the conditions of Lemma 2, with probability at least $1 - I^{-1}$, the fixed multilinear rank $(r,\ldots,r)$ approximation $\widehat{\mathcal{X}}(\widehat{\mathcal{P}})$ computed from Algorithms 1 and 2 with thresholds $\tau \geq \theta$ and $\gamma \geq \alpha$ satisfies $\frac{\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B}\|_{\mathrm{F}}}{\|\mathcal{B}\|_{\mathrm{F}}} \leq CN\sqrt{\frac{r\log I}{I}}$, in which C depends on κ.*

Note how this bound compares with the bounds for similar algorithms. `HOSVD_w` has a relative error of $O(rN^2 I^{-1/2}\log I)$ [14, Theorem 3.3] for noiseless recovery with known propensities, and `SO-HOSVD` achieves a better bound of $O(\sqrt{\frac{r\log I}{I^{N/2-1}}})$ [30, Theorem 3] but assumes that the tensor is MCAR. In contrast, our bound holds for the tensor MNAR setting and does not require known propensities. It is the first bound in this setting, to our knowledge.

5. Experiments

We ran all experiments on a Linux machine with Intel® Xeon® E7-4850 v4 2.10GHz CPU and 1056GB memory, and used the logistic link function $\sigma(x) = 1/(1+e^{-x})$ throughout. We use both synthetic and semi-synthetic data for evaluation. We first compare the propensity estimation



(a) Model A (b) Model B

Figure 1: Propensity estimation performance on different orders of tensors with size 8. The small size makes it possible to evaluate high-order tensors within a reasonable time. Figure 1a follows Model A with a 40% observation ratio. Figure 1b follows Model B. At each N, we have the order-N parameter tensor $\mathcal{A} \in \mathbb{R}^{8\times\cdots\times 8} = \mathcal{G}^A \times_1 U_1^A \times \cdots \times_N U_N^A$, in which $\mathcal{G} \in \mathbb{R}^{2\times\cdots\times 2}$ has i.i.d. $\mathrm{Uniform}[-2, 2]$ entries, and each $U_n^A \in \mathbb{R}^{8\times 2}$ is the left singular matrix of an 8-by-2 random matrix with i.i.d. $\mathrm{Uniform}[-1, 1]$ entries. The results without "(2)" correspond to setting hyperparameters $\tau = \theta$ and $\gamma = \alpha$ in 1-bit matrix completion, and those with "(2)" correspond to $\tau = 2\theta$ and $\gamma = 2\alpha$.

performance on square and non-square unfoldings, and then compare the tensor recovery error under different approaches. The relative error is defined as $\|\widehat{\mathcal{T}} - \mathcal{T}\|_\mathrm{F}/\|\mathcal{T}\|_\mathrm{F}$, in which $\widehat{\mathcal{T}}$ is the predicted tensor and $\mathcal{T}$ is the true tensor.

There are four algorithms similar to TenIPS for tensor completion in our experiments: SqUnfold, which performs SVD to seek for the low-rank approximation of the square unfolding of the propensity-reweighted $\mathcal{B}_\mathrm{obs}$; RectUnfold, which applies SVD to the unfolding of the propensity-reweighted $\mathcal{B}_\mathrm{obs}$ along a specific mode; HOSVD_w [14]; SO-HOSVD [30]. The popular nuclear-norm-regularized least squares LstSq that seeks for $\hat{B} = \arg\min\limits_X \sum_{(i,j)\in\Omega}(B_{ij} - X_{ij})^2 + \lambda\|X\|_\star$ on an unfolding of $\mathcal{B}_\mathrm{obs}$ takes much longer to finish in our experiments and is thus prohibitive in tensor completion practice, so we omit most of its results.

5.1. Synthetic data

We have the following observation models for synthetic tensors:

Model A. MCAR. $\mathcal{P}$ has all equal entries.

Model B. MNAR with an approximatly low multilinear rank $\mathcal{A}$. One special case is that $\mathcal{A}$ is proportional to $\mathcal{B}$: a larger entry is more likely to be observed.

In the first experiment, we evaluate the propensity estimation performance on synthetic tensors with approximately low multilinear rank. In each of the above observation models, we generate synthetic tensors with equal sidelengths on each mode, and predict the propensity tensor by either estimating the parameter tensor on the square unfolding, or on the unfolding along a specific mode. Figure 1 compares the propensity estimation error on tensors with different orders. We can see that:

1 Propensity estimation errors on square unfoldings are always smaller than those on unfolding the tensor along one specific mode. This comes from the fact that the low-nuclear-norm constraint is a surrogate for the low rank property, and that the square unfolding has a smaller relative rank (rank divided by size) compared to the unfolding along one mode.

2 Propensity estimation errors on the square unfolding have a smaller increase when optimization hyperparameters τ and γ increase from θ and α. This suggests that the square unfolding makes the constrained optimization problem more robust to the selection of optimization hyperparameters.

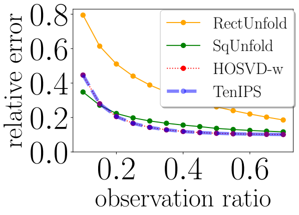


Figure 2: Error on MCAR tensors.

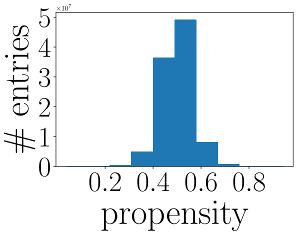


Figure 3: Propensity histogram.

Table 1: Completion performance on the order-4 MNAR synthetic cubical tensor with size 100. The "time" here is the time taken for the tensor completion step, with true or estimated propensities. $\widehat{\mathcal{P}}_1$ is from running the provable Algorithm 1 at target rank 25 for 84 seconds, and has relative error 0.08 from the true tensor $\mathcal{P}$; $\widehat{\mathcal{P}}_2$ is from running the gradient descent algorithm of Algorithm 3 with i.i.d. Uniform$[-1, 1]$ initialization and at step size 5×10^{-6} for 81 seconds, and has relative error 0.13.

Algorithm	time (s)	relative error from $\mathcal{B}$		
		with $\mathcal{P}$	with $\widehat{\mathcal{P}}_1$	with $\widehat{\mathcal{P}}_2$
TENIPS	26	**0.110**	**0.110**	**0.109**
HOSVD_W	35	0.129	0.116	0.110
SQUNFOLD	29	0.141	0.138	0.139
RECTUNFOLD	**8**	0.259	0.256	0.256
LSTSQ	>600	-	-	-
SO-HOSVD	>600	-	-	-

In the second experiment, we complete tensors with MCAR entries. With $N = 4$, $I = 100$ and $r = 5$, we generate an order-N data tensor $\mathcal{B} \in \mathbb{R}^{I\times\cdots\times I}$ in the following way: first generate $\mathcal{B}^\natural$ by Tucker decomposition $\mathcal{G}^{\text{true}} \times_1 U_1^{\text{true}} \times_2 \cdots \times_N U_N^{\text{true}}$, in which each factor matrix $U_n^{\text{true}} \in \mathbb{R}^{I\times r}$ is the left singular matrix of a different I-by-r matrix with i.i.d. Uniform$[-1, 1]$ entries, and the core tensor $\mathcal{G}^{\text{true}} \in \mathbb{R}^{r\times\cdots\times r}$ has i.i.d. $\mathcal{N}(0, 100^2)$ entries. Then we generate a noise-corrupted $\mathcal{B}$ by $\mathcal{B}^\natural + (\gamma\|\mathcal{B}^\natural\|_F/I^{N/2})\epsilon$, where the noise tensor ϵ has i.i.d. $\mathcal{N}(0, 1)$ entries.

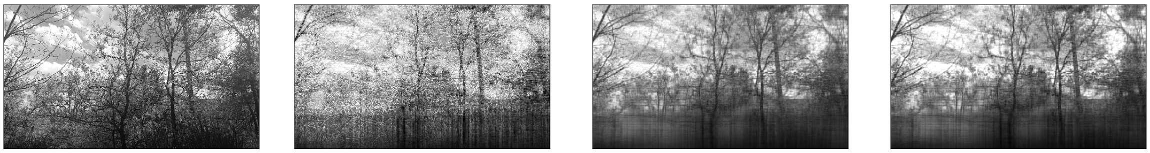
We compare the relative error of TENIPS with SQUNFOLD, RECTUNFOLD and HOSVD_W at different observation ratios in Figure 2. We can see that TENIPS and HOSVD_W achieves the lowest recovery error on average, and the results of these two methods are nearly identical.

In the third experiment, we complete tensors with MNAR entries. We use the same $\mathcal{B}$ as the second experiment, and further generate an order-4 parameter tensor $\mathcal{A} \in \mathbb{R}^{100\times\cdots\times 100}$ in the same way as $\mathcal{B}$. 99.97% of the propensities in $\mathcal{P} = \sigma(\mathcal{A})$ lie in the range of $[0.2, 0.8]$, as shown in Figure 3. In Table 1, we can see that:

1 TENIPS has the smallest error among methods that can finish within a reasonable time.

2 Tensor completion errors from estimated propensities are roughly equal to, and sometimes even smaller than those from true propensities despite a propensity estimation error. This can be attributed to the fact that propensity estimation by Algorithm 1 or Algorithm 3 denoises the parameter tensor $\mathcal{A}$ and thus gets a good estimate of $\mathcal{P}$ in terms of recovering its low multilinear rank structure.

3 On the sensitivity to hyperparameters: it is mentioned in the title of Table 1 that Algorithm 1 achieves a smaller accuracy within a similar time as Algorithm 3. However, this is because we set τ and γ correctly: $\tau = \theta$ and $\gamma = \alpha$. In real cases, θ and α are unknown, and are hard to infer from surrogate metrics within the optimization process. This may lead to large propensity estimation errors: an example is that Algorithm 1 with $\tau = 100\theta$ and $\gamma = 100\alpha$ would get a $\widehat{\mathcal{P}}$ with relative error larger than 0.7 after a few iterations. Moreover, the relative error does not always decrease



(*a*) original (*b*) TENIPS, assuming MCAR (*c*) TENIPS, assuming MNAR with true $\mathcal{P}$ (*d*) TENIPS, assuming MNAR with estimated $\widehat{\mathcal{P}}$

Figure 4: Video recovery visualization on Frame 500 of the [19] video data. The missingness patterns in 4b, 4c and 4d only refer to our assumption in tensor recovery; the partially observed data tensors we start from are the same and are MNAR.

with more iterations, despite that the decrease of objective value. On every algorithm in Table 1, this $\widehat{\mathcal{P}}$ gives a larger than 0.7 relative error for the estimation of data tensor $\mathcal{B}$. On the other hand, the initialization and step size in Algorithm 3 can be tuned more easily by monitoring the value of function f with the increase of number of iterations. More discussion can be found in Appendix I.

In the fourth experiment, we compare the above methods in both MCAR and MNAR settings when increasing target ranks, and the result is shown in Appendix H.

5.2. Semi-synthetic data

We use the video from [19] and generate synthetic propensities. The video was taken by a camera mounted at a fixed position with a person walking by. We convert it to grayscale and discard the frames with severe vibration. This gives our order-3 data tensor $\mathcal{B} \in \mathbb{R}^{2200\times1080\times1920}$ that takes 102.0GB memory in Python3. To get an MNAR tensor $\mathcal{B}_{\text{obs}}$, we generate the parameter tensor $\mathcal{A}$ by entrywise transformation $\mathcal{A} = (\mathcal{B} - 128)/64$. This centers and rescales the entries within the range of $[0, 255]$ in $\mathcal{B}$ to $[-2, 2]$ in $\mathcal{A}$ and gives propensities in $[0.12, 0.88]$ in $\mathcal{P} = \sigma(\mathcal{A})$. Finally we subsample $\mathcal{B}$ by $\mathcal{P}$ to get $\mathcal{B}_{\text{obs}}$. In Figure 4, we visualize the 500-th frame in three TENIPS experiments by fixed-rank approximation with target multilinear rank $(50, 50, 50)$: the original frame without missing pixels 4a, the frame recovered under MCAR assumption (tensor recovery error 0.42) 4b, the frame recovered by propensities under the MNAR assumption with the true propensity tensor $\mathcal{P}$(tensor recovery error 0.28) 4c, and the frame recovered by propensities under the MNAR assumption with the estimated propensity tensor $\widehat{\mathcal{P}}$ from Algorithm 1 (propensity estimation error 0.15, tensor recovery error 0.28) 4d. We can see that:

1 With MNAR pixels, the image recovered from the naive MCAR assumption in Figure 4b is more noisy than that from MNAR in Figure 4c and 4d, and misses more details.

2 There is no significant difference between the recovered video frames in 4c and 4d, in terms of both the frame image itself and the tensor recovery error.

6. Conclusion

This paper develops a provable two-step approach for MNAR tensor completion with unknown propensities. The square unfolding allows us to recover propensities with a smaller upper bound, and we then use HOSVD complete MNAR tensor with the estimated propensities. This method enjoys theoretical guarantee and fast running time in practice.

This paper is the first provable method for completing a general MNAR tensor. There are many avenues for improvement and extensions. For example, one could explore whether nonconvex matrix completion methods can be generalized to MNAR tensors, explore other observation models, and design provable algorithms that estimate the propensities even faster.

References

[1] Anastasia Aidini, Grigorios Tsagkatakis, and Panagiotis Tsakalides. 1-bit tensor completion. *Electronic Imaging*, 2018(13):261–1, 2018.

[2] Morteza Ashraphijuo and Xiaodong Wang. Fundamental conditions for low-cp-rank tensor completion. *The Journal of Machine Learning Research*, 18(1):2116–2145, 2017.

[3] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445, 2016.

[4] Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

[5] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[6] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3): 283–319, 1970.

[7] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[8] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[9] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[10] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[11] Navid Ghadermarzy, Yaniv Plan, and Ozgur Yilmaz. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2018.

[12] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[13] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163, 2020.

[14] Longxiu Huang and Deanna Needell. Hosvd-based algorithm for weighted tensor completion. *arXiv preprint arXiv:2003.08537*, 2020.

[15] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[16] Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in neural information processing systems*, pages 836–844, 2013.

[17] Allen Liu and Ankur Moitra. Tensor completion made practical. *arXiv preprint arXiv:2006.03134*, 2020.

[18] Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems*, pages 14871–14880, 2019.

[19] Osman Asif Malik and Stephen Becker. Low-rank tucker decomposition of large tensors using tensorsketch. In *Advances in Neural Information Processing Systems*, pages 10096–10106, 2018.

[20] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81, 2014.

[21] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May): 1665–1697, 2012.

[22] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5): 2295–2317, 2011.

[23] Ernest K Ryu and Wotao Yin. Proximal-proximal-gradient method. *arXiv preprint arXiv:1708.06908*, 2017.

[24] Yiming Sun, Yang Guo, Charlene Luo, Joel Tropp, and Madeleine Udell. Low-rank tucker approximation of a tensor from streaming data. *arXiv preprint arXiv:1904.10951*, 2019.

[25] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

[26] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[27] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31 (3):279–311, 1966.

[28] Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. Tensor completion by alternating minimization under the tensor train (TT) model. *arXiv preprint arXiv:1609.05587*, 2016.

[29] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[30] Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

[31] Tatsuya Yokota, Burak Erem, Seyhmus Guler, Simon K Warfield, and Hidekata Hontani. Missing slice recovery for tensors using a low-rank model in embedded space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8251–8259, 2018.

[32] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[33] Longhao Yuan, Qibin Zhao, Lihua Gui, and Jianting Cao. High-dimension tensor completion via gradient-based optimization under tensor-train format. *arXiv preprint arXiv:1804.01983*, 2018.

[34] Anru Zhang et al. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2): 936–964, 2019.

Appendix A. Related works

Tensor completion is gaining increasing popularity and is one of the major tensor-related research topics. The literature we survey here is by no means exhaustive. A straightforward method is to flatten a tensor along one of its dimensions to a matrix and then pick one of the extensively studied matrix completion algorithms [4, 5, 21]. However, this method neglects the multiway structure along all other dimensions and does not make full use of the combinatorial relationships. Instead, it is common to assume that the tensor is low rank along each mode. Tensors differ from matrices in having many incompatible notions of rank and low rank decompositions, including CANDECOMP/PARAFAC (CP) [6, 12], Tucker [27] and tensor-train [22]. Each of the decompositions exploits a slightly different definition of tensor rank, and can be used to recover tensors that are low rank in that sense, including CP [2, 3, 11, 15–17], Tucker [10, 14, 20, 30, 31, 34] and tensor-train [28, 33]. In this paper, we assume the tensor has approximately low multilinear rank, which corresponds to a tensor that can be approximated by a low rank Tucker decomposition.

Existing techniques used for tensor completion include subspace projection onto unfoldings [16], alternating minimization [15, 17, 28], gradient descent [33] and expectation-maximization [31]; different surrogates for the rank penalty have been used, including convex surrogates like nuclear norm on unfoldings [1, 10, 25] or specific flattenings [20] and the maximum norm on factors [11], and nonconvex surrogates such as the minimum number of rank-1 sign tensor components [11]. We use a higher-order SVD (HOSVD) approach that does not require rank surrogates. The two methods closest to ours are `HOSVD_w` [14], which computes a weighted HOSVD by reweighting the tensor twice by the inverse square root propensity, before and after the HOSVD, and the method in [30], which computes a HOSVD on a second-order estimator of a missing completely at random tensor, which we call `SO-HOSVD`. We compare our method with `HOSVD_w` and `SO-HOSVD` theoretically in Section 4.2 and numerically in Section 5.1.

Most previous works on tensor completion have used the assumption that entries are missing completely at random (MCAR). The missing not at random (MNAR) setting is less studied, especially for tensors. A missingness pattern is MNAR when the observation probabilities (also called propensities) of different entries are not equal and may depend on the entry values themselves. In the matrix

MNAR setting, a popular observation model is 1-bit observations [7]: each entry is observed with a probability that comes from applying a differentiable function $\sigma : \mathbb{R} \to [0, 1]$ to the corresponding entry in a parameter matrix, which is assumed to be low rank. Two popular convex surrogates for the rank have been used to estimate the parameter matrix from an entrywise binary observation pattern using a regularized likelihood approach: nuclear norm [1, 7, 18] and max-norm [4, 11]. We show that we can achieve a small propensity estimation error by solving for a parameter tensor with low multilinear rank using a (roughly) square flattening. This approach outperforms simple slicing or flattening methods.

In this paper, we study the problem of provably completing a MNAR tensor with (approximately) low multilinear rank. We use a two-step procedure to first estimate the propensities in this tensor and then predict missing values by HOSVD on the inverse propensity reweighted tensor. We give the error bound on final estimation as Theorem 4.

Appendix B. More notations

Basics We define $[N] = \{1, \ldots, N\}$ for a positive integer N. Given a set S, we denote its cardinality by $|S|$. $\subset$ denotes strict subset. We denote $f(n) = O(g(n))$ if there exists $C > 0$ and N such that $|f(n)| \leq Cg(n)$ for all $n \geq N$.

Matrices and tensors We denote *vector*, *matrix*, and *tensor* variables respectively by lowercase letters (x), capital letters (X) and Euler script letters ($\mathcal{X}$). For a matrix $X \in \mathbb{R}^{m \times n}$, $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_{\min\{m,n\}}(X)$ denote its singular values, $\|X\|$ denotes its 2-norm, $\|X\|_\star$ denotes its nuclear norm, $\mathrm{tr}(X)$ denotes its trace, and with another matrix $Y \in \mathbb{R}^{m \times n}$, $\langle X, Y \rangle := \mathrm{tr}(X^\top Y)$ denotes the matrix inner product. For a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $\|\mathcal{X}\|_{\max}$ denotes its entrywise maximum absolute value. The *order* of a tensor is the number of dimensions; matrices are order-two tensors. Each dimension is called a *mode*. To denote a part of matrix or tensor, we use a colon to denote the mode that is not fixed: given a matrix $A \in \mathbb{R}^{I \times J}$, $A_{i,:}$ and $A_{:,j}$ denote the ith row and jth column of A, respectively. A *fiber* is a one-dimensional section of a tensor $\mathcal{X}$, defined by fixing every index but one; for example, a fiber of the order-3 tensor $\mathcal{X}$ is $X_{:,j,k}$. A *slice* is an $(N-1)$-dimensional section of an order-N tensor $\mathcal{X}$: a slice of the order-3 tensor $\mathcal{X}$ is $X_{:,:,k}$. The *size* of a mode is the number of slices along that mode: the n-th mode of $\mathcal{X}$ has size I_n. A tensor is *cubical* if every mode is the same size: $\mathcal{X} \in \mathbb{R}^{I \times I \times \cdots \times I}$. The *mode-$n$ unfolding* of $\mathcal{X}$, denoted as $\mathcal{X}^{(n)}$, is a matrix whose columns are the mode-n fibers of $\mathcal{X}$. For example, given an order-3 tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, $\mathcal{X}^{(1)} \in \mathbb{R}^{I \times (J \times K)}$.

Products We denote the *n-mode product* of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $U \in \mathbb{R}^{J \times I_n}$ by $\mathcal{X} \times_n U \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$; the $(i_1, i_2, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N)$-th entry is $\Sigma_{i_n=1}^{I_n} x_{i_1 i_2 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} u_{j i_n}$. $\otimes$ denotes the Kronecker product. Given two tensors with the same shape, we use $\odot$ to denote their entrywise product.

Missingness Given a partially observed order-N tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, we denote its observation pattern by $\Omega \in \{0, 1\}^{I_1 \times \cdots \times I_N}$: the *mask tensor* of $\mathcal{X}$. It is a binary tensor that denotes whether each entry of $\mathcal{X}$ is observed or not. Ω has the same shape as $\mathcal{X}$, with entry value 1 if the corresponding entry of $\mathcal{X}$ is observed, and 0 otherwise. With an abuse of notation, we call $\Omega := \{(i_1, i_2, \ldots, i_N) | \Omega_{i_1, i_2, \ldots, i_N} = 1\}$ the *mask set* of $\mathcal{X}$. Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, we use $\mathcal{E}(i_1, i_2, \ldots, i_N)$ to denote a binary tensor with the same shape as $\mathcal{X}$, with value 1 at the $(i_1, i_2, \ldots, i_N)$-th entry and 0 elsewhere.

Algorithm 3 Propensity estimation by gradient descent

```
Input: partially observed tensor $\mathcal{B}_{\text{obs}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, link function $\sigma$, step size $t$, initializations
    $\mathcal{G}_0^{\mathcal{A}}, (U_1)_0^{\mathcal{A}}, \ldots, (U_N)_0^{\mathcal{A}}$
Output: Estimated propensity tensor $\widehat{\mathcal{P}}$
1  $\Omega \leftarrow$ mask tensor of $\mathcal{B}_{\text{obs}}$
2  $\mathcal{G}^{\mathcal{A}}, U_1^{\mathcal{A}}, \ldots, U_N^{\mathcal{A}} \leftarrow \mathcal{G}_0^{\mathcal{A}}, (U_1)_0^{\mathcal{A}}, \ldots, (U_N)_0^{\mathcal{A}}$
3  do
4      $f(\mathcal{G}^{\mathcal{A}}, \{U_n^{\mathcal{A}}\}_{n \in [N]}) \leftarrow \sum_{i_1 \cdots i_N} -\Omega_{i_1 \cdots i_N} \log \sigma(\widehat{\mathcal{A}}_{i_1 \cdots i_N}) - (1 - \Omega_{i_1 \cdots i_N}) \log(1 - \sigma(\widehat{\mathcal{A}}_{i_1 \cdots i_N})),$
       in which $\widehat{\mathcal{A}} = \mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}}$.
5      $\mathcal{G}^{\mathcal{A}}, U_1^{\mathcal{A}}, \ldots, U_N^{\mathcal{A}} \leftarrow (\mathcal{G}^{\mathcal{A}} - t\frac{\partial f}{\partial \mathcal{G}^{\mathcal{A}}}, U_1^{\mathcal{A}} - t\frac{\partial f}{\partial U_1^{\mathcal{A}}}, \ldots, U_N^{\mathcal{A}} - t\frac{\partial f}{\partial U_N^{\mathcal{A}}})$ ▷ gradient descent updates
6  while not converged
7  $\widehat{\mathcal{P}} \leftarrow \sigma(\widehat{\mathcal{A}})$
8  return $\widehat{\mathcal{P}}$
```

Appendix C. The gradient descent algorithm for propensity estimation

In practice, the square unfolding of a tensor is often a large matrix: $I^{N/2}$-by-$I^{N/2}$ for a cubical tensor with order N. Since each iteration of the `prox-prox` subroutine in Algorithm 1 requires the computation of a truncated SVD, this algorithm becomes too expensive in such case. Also, it does not make full use of the low multilinear rank property of $\mathcal{A}$. As a substitute, we propose Algorithm 3, which uses gradient descent on the core tensor $\mathcal{G}^{\mathcal{A}}$ and factor matrices $\{U_n^{\mathcal{A}}\}_{n \in [N]}$ to minimize the objective function $f(\mathcal{G}^{\mathcal{A}}, \{U_n^{\mathcal{A}}\}_{n \in [N]})$ defined in Line 4 of Algorithm 3. It achieves a feasible solution with similar quality as Algorithm 1, and does not require the tuning of thresholds τ and γ. This can be attributed to the fact that the objective function f is multi-convex with respect $(\mathcal{G}^{\mathcal{A}}, U_1^{\mathcal{A}}, \ldots, U_N^{\mathcal{A}})$. The gradient computation can be found in Appendix G.

Appendix D. Proof for Section 4.1 on propensity estimates

Lemma 1 **Proof** [20, Lemma 6 (2)] states that

$$\mathcal{X}_{[n]} = (U_n \otimes U_{n-1} \otimes \cdots \otimes U_1)\, \mathcal{C}_{[n]}\, (U_K \otimes U_{K-1} \otimes \cdots \otimes U_{n+1})^\top \tag{1}$$

for $n \in [N]$. Thus Lemma 1 holds for $\mathcal{X}$ by applying Equation 1 to an entry-reordered tensor $\widetilde{\mathcal{X}} \in \mathbb{R}^{I_{n_1} \times I_{n_2} \times \cdots \times I_{n_N}}$, such that $S = \{n_j\}_{j \in [S]}$ and $\widetilde{\mathcal{X}}_{i_{n_1} i_{n_2} \cdots i_{n_N}} = \mathcal{X}_{i_1 i_2 \cdots i_N}$. The upper bound for $\text{rank}(\mathcal{X}_S)$ follows. ■

Theorem 3 **Proof** Denote $r_S^{\text{true}} := \prod_{n \in S} r_n^{\text{true}}$ and $r_{S^C}^{\text{true}} := \prod_{n \in [N] \setminus S} r_n^{\text{true}}$. We know from Lemma 1 of the main paper that for every unfolding of $\mathcal{A}$, $\text{rank}(\mathcal{A}_S) \leq \min\left\{r_S^{\text{true}}, r_{S^C}^{\text{true}}\right\}$. Since $\|\mathcal{A}_S\|_\star \leq \sqrt{\text{rank}(\mathcal{A}_S)} \cdot \|\mathcal{A}\|_{\text{F}} \leq \alpha\sqrt{\text{rank}(\mathcal{A}_S)} \cdot I^{\frac{N}{2}}$, we need $\tau \geq \theta \geq \alpha\sqrt{\text{rank}(\mathcal{A}_S)}$ for the conditions of Lemma 2 to hold. For simplicity, suppose $\tau = \alpha\sqrt{\mathrm{rank}(\mathcal{A}_S)}$, the smallest possible value for the exact recovery of $\mathcal{A}$.

Without loss of generality, suppose $|S| \le \frac{N}{2}$. We have

$$\begin{aligned}\frac{1}{I_{[N]}}\|\widehat{\mathcal{P}} - \mathcal{P}\|_{\mathrm{F}}^2 &\le 4eL_\gamma\tau\Big(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}}\Big)\\ &= 4eL_\gamma\alpha\cdot\sqrt{\mathrm{rank}(\mathcal{A}_S)}\Big(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}}\Big)\\ &\le 4eL_\gamma\alpha\Big(\sqrt{\frac{r_S^{\mathrm{true}}}{I_S}} + \sqrt{\frac{r_{S^C}^{\mathrm{true}}}{I_{S^C}}}\Big)\\ &\le 4eL_\gamma\alpha\Big(\sqrt{c^{|S|}} + \sqrt{c^{N-|S|}}\Big).\end{aligned}$$

The final expression is the smallest when $S = S_\square$. ■

Appendix E. Error in tensor completion (Algorithm 1 and 2): general case

We first state Theorem 5, the tensor completion error in the most general case. For brevity, we denote $\widehat{\mathcal{X}}(\mathcal{P})$ and $\bar{\mathcal{X}}(\mathcal{P})$ by $\widehat{\mathcal{X}}$ and $\bar{\mathcal{X}}$, respectively, in which $\mathcal{P}$ is the true propensity tensor.

Theorem 5 *Consider an order-N tensor $\mathcal{B} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, and two order-$N$ tensors $\mathcal{P}$ and $\mathcal{A}$ with the same shape as $\mathcal{B}$. Each entry $\mathcal{B}_{i_1,\ldots,i_N}$ of $\mathcal{B}$ is observed with probability $\mathcal{P}_{i_1,\ldots,i_N}$ from the corresponding entry of $\mathcal{P}$. Assume there exist constants $\psi, \alpha \in (0,\infty)$ such that $\|\mathcal{A}\|_{\max} \le \alpha$, $\|\mathcal{B}\|_{\max} = \psi$. Denote the spikiness parameter $\alpha_{\mathrm{sp}} := \psi\sqrt{I_{[N]}}/\|\mathcal{B}\|_{\mathrm{F}}$. Then under the conditions of Lemma 2, with probability at least $1 - \frac{C_1}{I_\square + I_{\square^C}} - \sum_{n=1}^N [I_n + I_{(-n)}]\exp\Big[-\frac{\epsilon^2\|\mathcal{B}\|_{\mathrm{F}}^2\sigma(-\alpha)/2}{I_{(-n)}\psi^2 + \epsilon\psi\|\mathcal{B}\|_{\mathrm{F}}/3}\Big]$, in which $C_1 > 0$ is a universal constant, the fixed multilinear rank $(r_1, r_2, \cdots, r_N)$ approximation $\widehat{\mathcal{X}}(\widehat{\mathcal{P}})$ computed from Algorithms 1 and 2 with thresholds $\tau \ge \theta$ and $\gamma \ge \alpha$ satisfies*

$$\begin{aligned}\frac{\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_{\mathrm{F}}^2}{\|\mathcal{B}\|_{\mathrm{F}}^2} \le \min_{n\in[N]}&\left\{r_n\cdot\Big[\frac{\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}}}{\|\mathcal{B}\|_{\mathrm{F}}} + \epsilon\Big]^2\right\}\\ &+ \sum_{n=1}^N \frac{12 r_n \sigma_1(\mathcal{B}^{(n)})^2}{\|\mathcal{B}\|_{\mathrm{F}}^2}\cdot\left\{\frac{[2\sigma_1(\mathcal{B}^{(n)}) + \|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}} + \epsilon\|\mathcal{B}\|_{\mathrm{F}}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)}) + \sigma_{r_n+1}(\mathcal{B}^{(n)})]^2}\cdot\frac{[\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}} + \epsilon\|\mathcal{B}\|_{\mathrm{F}}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)}) - \sigma_{r_n+1}(\mathcal{B}^{(n)})]^2}\right\}\\ &+ \frac{1}{\|\mathcal{B}\|_{\mathrm{F}}^2}\sum_{n=1}^N(\tau_{r_n}^{(n)})^2,\end{aligned} \tag{2}$$

in which:

1. *$(\tau_{r_n}^{(n)})^2 := \sum_{i=r_n+1}^{I_n}\sigma_i^2(\mathcal{B}^{(n)})$ is the r_n-th tail energy for $\mathcal{B}^{(n)}$,*
2. *from Lemma 2, with $L_\gamma = \sup_{x\in[-\gamma,\gamma]}\frac{|\sigma'(x)|}{\sigma(x)(1-\sigma(x))}$, and with probability at least $1 - \frac{C_1}{I_\square + I_{\square^C}}$,*

$$\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}} \le \frac{\alpha_{\mathrm{sp}}\|\mathcal{B}\|_{\mathrm{F}}}{\sigma(-\gamma)\sigma(-\alpha)}\sqrt{4eL_\gamma\tau\Big(\frac{1}{\sqrt{I_\square}} + \frac{1}{\sqrt{I_{\square^C}}}\Big)}. \tag{3}$$

On the right-hand side of Equation 2, the first term comes from the error between $\bar{\mathcal{X}}(\mathcal{P})$ and $\mathcal{B}$ when projected onto the truncated column singular spaces in each mode $n \in [N]$; the second and third terms come from the projection error of $\mathcal{B}$ onto the above spaces.

Appendix F. Proof for Theorem 4 and 5

F.1. Proof for Theorem 5, the general case

We first show the proof for Theorem 5, the general case. We start with Lemma 6 on how the error in propensity estimates propagate to the error in the inverse propensity estimator $\bar{\mathcal{X}}(\widehat{\mathcal{P}})$, then bound the error between $\widehat{\mathcal{X}}(\widehat{\mathcal{P}})$ and $\mathcal{B}$.

Lemma 6 *Instate the conditions of Lemma 2 and further suppose* $\|\mathcal{B}\|_{\max} = \psi$. *Then with probability at least* $1 - \frac{C_1}{I_S + I_{S^C}}$, *in which* $C_1 > 0$ *is a universal constant,*

$$\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}}^2 \leq \frac{4eL_\gamma \tau \psi^2}{\sigma(-\gamma)^2 \sigma(-\alpha)^2} \Big(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}} \Big) I_{[N]}. \tag{4}$$

Proof Under the above conditions,

$$\begin{aligned}
\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\mathrm{F}}^2 &= \sum_{(i_1, i_2, \dots, i_N) \in \Omega} \mathcal{B}_{i_1 i_2 \cdots i_N}^2 \big(\frac{1}{\mathcal{P}_{i_1 i_2 \cdots i_N}} - \frac{1}{\widehat{\mathcal{P}}_{i_1 i_2 \cdots i_N}} \big)^2 \\
&\leq \psi^2 \sum_{(i_1, i_2, \dots, i_N) \in \Omega} \Big(\frac{\mathcal{P}_{i_1 i_2 \cdots i_N} - \widehat{\mathcal{P}}_{i_1 i_2 \cdots i_N}}{\mathcal{P}_{i_1 i_2 \cdots i_N} \widehat{\mathcal{P}}_{i_1 i_2 \cdots i_N}} \Big)^2 \\
&\leq \frac{\psi^2}{\sigma(-\gamma)^2 \sigma(-\alpha)^2} \sum_{(i_1, i_2, \dots, i_N) \in \Omega} \Big(\mathcal{P}_{i_1 i_2 \cdots i_N} - \widehat{\mathcal{P}}_{i_1 i_2 \cdots i_N} \Big)^2 \\
&\leq \frac{4eL_\gamma \tau \psi^2}{\sigma(-\gamma)^2 \sigma(-\alpha)^2} \Big(\frac{1}{\sqrt{I_S}} + \frac{1}{\sqrt{I_{S^C}}} \Big) I_{[N]}.
\end{aligned}$$

The second inequality comes from $\widehat{\mathcal{P}}_{i_1 i_2 \cdots i_N} \geq \sigma(-\gamma)$ and $\mathcal{P}_{i_1 i_2 \cdots i_N} \geq \sigma(-\alpha)$; the last inequality follows Lemma 2. ■

We then state two lemmas that we will apply to tensor unfoldings. Lemma 7 is the matrix Bernstein inequality. Lemma 8 is a variant of the Davis-Kahan sin(Θ) Theorem [8].

Lemma 7 (matrix Bernstein for real matrices [26, Theorem 1.6.2]) *Let* $S_1, \dots, S_k$ *be independent, centered random matrices with common dimension* $m \times n$, *and assume that each one is uniformly bounded*

$$\mathbb{E}\, S_i = 0 \quad \textit{and} \quad \|S_i\| \leq L \quad \textit{for each } i = 1, \dots, k.$$

Introduce the sum

$$Z = \sum_{i=1}^{k} S_i,$$

and let $v(Z)$ denote the matrix variance statistic of the sum:

$$\begin{aligned} v(Z) &= \max\left\{\|\mathbb{E}(ZZ^\top)\|,\ \|\mathbb{E}(Z^\top Z)\|\right\} \\ &= \max\left\{\|\sum_{i=1}^{k}\mathbb{E}\left(S_i S_i^\top\right)\|,\ \|\sum_{i=1}^{k}\mathbb{E}\left(S_i^\top S_i\right)\|\right\}. \end{aligned}$$

Then

$$\mathbb{P}\left\{\|Z\| \geq t\right\} \leq (m+n)\cdot\exp\left(\frac{-t^2/2}{v(Z)+Lt/3}\right) \quad \textit{for all } t \geq 0.$$

Lemma 8 (Variant of the Davis-Kahan sin(Θ) Theorem [29], [32, Theorem 4]) *Let $A, \widehat{A} \in \mathbb{R}^{p\times q}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{\min(p,q)}$ and $\widehat{\sigma}_1 \geq \ldots \geq \widehat{\sigma}_{\min(p,q)}$ respectively, and have singular vectors $\{u_i\}_{i=1}^n$, $\{v_i\}_{i=1}^n$ and $\{\widehat{u}_i\}_{i=1}^n$, $\{\widehat{v}_i\}_{i=1}^n$, respectively. Let $V = (v_1, \cdots, v_r) \in \mathbb{R}^{n\times r}$, $\widehat{V} = (\widehat{v}_1, \cdots, \widehat{v}_r) \in \mathbb{R}^{n\times r}$, $V_\perp = (v_{r+1}, \cdots, v_n) \in \mathbb{R}^{n\times(n-r)}$ and $\widehat{V}_\perp = (\widehat{v}_{r+1}, \cdots, \widehat{v}_n) \in \mathbb{R}^{n\times(n-r)}$. Assume that $\sigma_r^2 - \sigma_{r+1}^2 > 0$, then*

$$\|\widehat{V}_\perp^\top V\|_{\mathrm{F}} = \|V_\perp^\top \widehat{V}\|_{\mathrm{F}} = \|\widehat{V}\widehat{V}^\top - VV^\top\|_{\mathrm{F}} \leq \frac{2(2\sigma_1 + \|\widehat{A}-A\|)\min(r^{1/2}\|\widehat{A}-A\|, \|\widehat{A}-A\|_{\mathrm{F}})}{\sigma_r^2 - \sigma_{r+1}^2}.$$

Identical bounds also hold if V and $\widehat{V}$ are replaced with the matrices of left singular vectors U and $\widehat{U}$, where $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{R}^{p\times d}$ and $\widehat{U} = (\widehat{u}_r, \widehat{u}_{r+1}, \ldots, \widehat{u}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $A^\top u_j = \sigma_j v_j$ and $\widehat{A}^\top \hat{u}_j = \widehat{\sigma}_j \widehat{v}_j$ for $j = r, r+1, \ldots, s$.

Upper bound on $\|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|$: We decompose it into the error between $\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}})$ and $\bar{\mathcal{X}}^{(n)}(\mathcal{P})$, and the error between $\bar{\mathcal{X}}^{(n)}(\mathcal{P})$ and $\mathcal{B}$, and independently bound these two terms:

$$\begin{aligned} \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\| &\leq \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}^{(n)}\| + \|\bar{\mathcal{X}}^{(n)} - \mathcal{B}^{(n)}\| \\ &\leq \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}^{(n)}\|_{\mathrm{F}} + \|\bar{\mathcal{X}}^{(n)} - \mathcal{B}^{(n)}\|. \end{aligned} \tag{5}$$

The first RHS term bounded by Lemma 6, the error given by propensity estimation. Note that we can get a tighter bound if we can directly bound $\|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}^{(n)}\|$. The second RHS term can be bounded by Lemma 7, the matrix Bernstein inequality, as below.

For each $(i_1, \ldots, i_N)$, define the random variable

$$\mathcal{S}_{i_1 i_2\cdots i_N} := \begin{cases} \Big(\dfrac{1}{\mathcal{P}_{i_1 i_2\cdots i_N}} - 1\Big)\mathcal{B} \odot \mathcal{E}(i_1, i_2, \ldots, i_N), & \text{with probability } \mathcal{P}_{i_1 i_2\cdots i_N} \\ -\mathcal{B} \odot \mathcal{E}(i_1, i_2, \ldots, i_N), & \text{with probability } 1 - \mathcal{P}_{i_1 i_2\cdots i_N}. \end{cases}$$

With the assumptions in Theorem 5, $\mathbb{E}\,\mathcal{S}_{i_1 i_2\cdots i_N} = 0$ and $\|\mathcal{S}^{(n)}_{i_1 i_2\cdots i_N}\| \leq \frac{\psi}{\sigma(-\alpha)}$. Also, the per-mode second moment is bounded as

$$\begin{aligned} v_n(\mathcal{X}) = \max\{&\|\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathbb{E}[\mathcal{S}^{(n)}_{i_1 i_2\cdots i_N}(\mathcal{S}^{(n)}_{i_1 i_2\cdots i_N})^\top]\|, \|\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathbb{E}[(\mathcal{S}^{(n)}_{i_1 i_2\cdots i_N})^\top \mathcal{S}^{(n)}_{i_1 i_2\cdots i_N}]\|\} \\ \leq \frac{\psi^2\cdot I_{(-n)}}{\sigma(-\alpha)}&. \end{aligned}$$

With probability at least $1-[I_n+I_{(-n)}]\exp\left[-\frac{\epsilon^2\|\mathcal{B}\|_{\mathrm{F}}^2\sigma(-\alpha)/2}{I_{(-n)}\psi^2+\epsilon\psi\|\mathcal{B}\|_{\mathrm{F}}/3}\right]$, the sum of random variables is bounded as $\|\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathcal{S}_{i_1i_2\cdots i_N}\|\leq\epsilon\|\mathcal{B}\|_{\mathrm{F}}$. Notice the difference between the propensity-reweighted observed tensor $\bar{\mathcal{X}}(\mathcal{P})$ and the true tensor $\mathcal{B}$,

$$\bar{\mathcal{X}}(\mathcal{P})-\mathcal{B}=\sum_{(i_1,i_2,\ldots,i_N)\in\Omega}\frac{1}{\mathcal{P}_{i_1,i_2,\ldots,i_N}}\mathcal{B}_{\text{obs}}\odot\mathcal{E}(i_1,i_2,\ldots,i_N)$$

is an instance of $\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathcal{S}_{i_1i_2\cdots i_N}$ over the randomness of entry-wise observation, hence we can use the matrix Bernstein inequality (Lemma 7) to bound $\|\bar{\mathcal{X}}(\mathcal{P})-\mathcal{B}\|$. Together with Equations 4 and 5, we get the upper bound on $\|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}})-\mathcal{B}^{(n)}\|$.

How $\|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}})-\mathcal{B}^{(n)}\|$ propagates into the final error in Algorithm 2: In Algorithm 2,

$$\widehat{\mathcal{X}}(\widehat{\mathcal{P}})=\underbrace{\left[\bar{\mathcal{X}}(\widehat{\mathcal{P}})\times Q_1^\top\times_2\cdots\times_N Q_N^\top\right]}_{\mathcal{W}(\widehat{\mathcal{P}})}\times_1 Q_1\times_2\cdots\times_N Q_N=\bar{\mathcal{X}}(\widehat{\mathcal{P}})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top.$$

This projects each unfolding of $\bar{\mathcal{X}}(\widehat{\mathcal{P}})$ onto the space of its truncated left singular vectors. Thus by adding and subtracting $\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top$ within the Frobenius norm, we decompose the error as

$$\begin{aligned}
\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B}\|_{\mathrm{F}}^2 &= \|\bar{\mathcal{X}}(\widehat{\mathcal{P}})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}\|_{\mathrm{F}}^2\\
&= \|\bar{\mathcal{X}}(\widehat{\mathcal{P}})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top\\
&\quad+\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}\|_{\mathrm{F}}^2\\
&= \underbrace{\|(\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top\|_{\mathrm{F}}^2}_{\textcircled{1}}\\
&+\underbrace{\|\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}\|_{\mathrm{F}}^2}_{\textcircled{2}}\\
&+\underbrace{2\langle(\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top,\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}\rangle}_{\textcircled{3}}.
\end{aligned}$$

First, we show that the cross term $\textcircled{3}$ is zero, since it is the product of two terms that are projected onto mutually orthogonal subspaces. For each $n\in[N]$,

$$[(\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top]^{(n)}=Q_n\mathcal{C}_n^{(n)},$$

where $\mathcal{C}_n^{(n)}$ is the mode-n unfolding of the tensor $\mathcal{C}_n$ defined as

$$\mathcal{C}_n=[(\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B})\times_1 Q_1^\top\cdots\times_N Q_N^\top]\times_1 Q_1\cdots\times_{n-1}Q_{n-1}\times_{n+1}Q_{n+1}\cdots\times_N Q_N.$$

Thus we have

$$\begin{aligned}\textcircled{3} &= 2\sum_{n=1}^{N}\langle \mathcal{Y}_n - \mathcal{Y}_{n-1}, (\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \mathcal{B}) \times_1 Q_1 Q_1^\top \times_2 \cdots \times_N Q_N Q_N^\top \rangle \\ &= 2\langle (Q_n Q_n^\top - I)\mathcal{Y}_{n-1}^{(n)}, Q_n \mathcal{C}_n^{(n)} \rangle \\ &= 2\mathrm{tr}(\mathcal{Y}_{n-1}^{(n)}(Q_n Q_n^\top - I) Q_n \mathcal{C}_n^{(n)}) = 0.\end{aligned}$$

Next, for Terms ① and ②, we introduce more notation before we analyze the error. Define $\mathcal{Y}_0 = \mathcal{B}$, and for each $n \in [N]$ let

$$\mathcal{Y}_n = \mathcal{B} \times_1 Q_1 Q_1^\top \times_2 \cdots \times_n Q_n Q_n^\top.$$

Thus $\mathcal{B} \times_1 Q_1 Q_1^\top \times_2 \cdots \times_N Q_N Q_N^\top - \mathcal{B} = \mathcal{Y}_N - \mathcal{Y}_0 = \sum_{n=1}^N (\mathcal{Y}_n - \mathcal{Y}_{n-1})$. Each $n \in [N]$ in the sum satisfies

$$\mathcal{Y}_n - \mathcal{Y}_{n-1} = \mathcal{Y}_{n-1} \times_n (Q_n Q_n^\top - I).$$

This allows us to analyze each mode individually.

For Term ①, for any $n \in [N]$, we have

$$\begin{aligned}\textcircled{1} &\leq \min_{n\in[N]} \left\{ \|Q_n Q_n^\top (\bar{\mathcal{X}}(\widehat{\mathcal{P}})^{(n)} - \mathcal{B}^{(n)})\|_{\mathrm{F}}^2 \right\} \\ &\leq \min_{n\in[N]} \left\{ r_n \cdot \|\bar{\mathcal{X}}(\widehat{\mathcal{P}})^{(n)} - \mathcal{B}^{(n)}\|^2 \right\},\end{aligned}$$

the RHS of which can be bounded from Section F.1.

As for Term ②, it can be bounded using a technique similar to [24, Lemma B.1]. For each $n \in [N]$,

$$\begin{aligned}\|\mathcal{Y}_n - \mathcal{Y}_{n-1}\|_{\mathrm{F}}^2 &= \|\mathcal{B} \times_n (I - Q_n Q_n^\top) \times_1 Q_1 Q_1^\top \cdots \times_n Q_{n-1} Q_{n-1}^\top\|_{\mathrm{F}}^2 \\ &\leq \|\mathcal{B} \times_n (I - Q_n Q_n^\top)\|_{\mathrm{F}}^2 \\ &= \|(I - Q_n Q_n^\top)\mathcal{B}^{(n)}\|_{\mathrm{F}}^2 \\ &= \|(U_n U_n^\top - Q_n Q_n^\top)\mathcal{B}^{(n)} + (U_n)_\perp (U_n)_\perp^\top \mathcal{B}^{(n)}\|_{\mathrm{F}}^2 \\ &= \underbrace{\|(U_n U_n^\top - Q_n Q_n^\top)\mathcal{B}^{(n)}\|_{\mathrm{F}}^2}_{\textcircled{4}} + \underbrace{\|(U_n)_\perp (U_n)_\perp^\top \mathcal{B}^{(n)}\|_{\mathrm{F}}^2}_{\textcircled{5}} + \underbrace{2\mathrm{tr}\big((\mathcal{B}^{(n)})^\top Q_n Q_n^\top (U_n)_\perp (U_n)_\perp^\top \mathcal{B}^{(n)}\big)}_{\textcircled{6}},\end{aligned}$$

in which ⑤ and ⑥ vanish when $r_n^{\mathrm{true}} \leq r_n$, since $(U_n)_\perp = 0$.

In the general case:

- The error between projections of $\mathcal{B}^{(n)}$ onto U_n and Q_n is

$$\begin{aligned}\textcircled{4} &\leq \sigma_1(\mathcal{B}^{(n)})^2 \|U_n U_n^\top - Q_n Q_n^\top\|_{\mathrm{F}}^2 \\ &\leq 4\sigma_1(B^{(n)})^2 r_n \cdot \frac{[2\sigma_1(\mathcal{B}^{(n)}) + \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|]^2 \cdot \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|^2}{[\sigma_{r_n}^2(\mathcal{B}^{(n)}) - \sigma_{r_n+1}^2(\mathcal{B}^{(n)})]^2},\end{aligned}$$

  in which the last inequality comes from Lemma 8.

- The residual ⑤ $= \sum_{i=r_n+1}^{I_n} \sigma_i^2(\mathcal{B}^{(n)}) = (\tau_{r_n}^{(n)})^2$ is the r_n-th tail energy for $\mathcal{B}^{(n)}$.
- The inner product of projections is

$$\begin{aligned}
⑥ &\leq 2\|(\mathcal{B}^{(n)})^\top \mathcal{B}^{(n)}\|_2 \cdot \mathrm{tr}\Big[[Q_n^\top (U_n)_\perp]^\top Q_n^\top (U_n)_\perp\Big] \\
&\leq 2\sigma_1(\mathcal{B}^{(n)})^2 \cdot \|Q_n^\top (U_n)_\perp\|_\mathrm{F}^2 \\
&\leq 2\sigma_1(\mathcal{B}^{(n)})^2 \cdot \Big\{\frac{2[2\sigma_1(\mathcal{B}^{(n)}) + \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|] \min(r_n^{1/2}\|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|, \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|_\mathrm{F})}{\sigma_{r_n}^2(\mathcal{B}^{(n)}) - \sigma_{r_n+1}^2(\mathcal{B}^{(n)})}\Big\}^2 \\
&\leq 8\sigma_1(\mathcal{B}^{(n)})^2 r_n \cdot \frac{[2\sigma_1(\mathcal{B}^{(n)}) + \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|]^2 \cdot \|\bar{\mathcal{X}}^{(n)}(\widehat{\mathcal{P}}) - \mathcal{B}^{(n)}\|^2}{[\sigma_{r_n}^2(\mathcal{B}^{(n)}) - \sigma_{r_n+1}^2(\mathcal{B}^{(n)})]^2},
\end{aligned}$$

  in which the first inequality comes from $\mathrm{tr}(AB) \leq \lambda_1(A)\mathrm{tr}(B)$ for positive semidefinite matrices A, B, and the second from last inequality comes from Lemma 8.

Together the above conclude the proof for Theorem 5.

F.2. Proof for Theorem 4, the special case

Recall the high-probability upper bound of Theorem 5, Equation 2 is

$$\begin{aligned}
\frac{\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_\mathrm{F}^2}{\|\mathcal{B}\|_\mathrm{F}^2} \leq \min_{n\in[N]} &\left\{ r_n \cdot \Big[\frac{\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_\mathrm{F}}{\|\mathcal{B}\|_\mathrm{F}} + \epsilon\Big]^2 \right\} \\
&+ \sum_{n=1}^N \frac{12 r_n \sigma_1(\mathcal{B}^{(n)})^2}{\|\mathcal{B}\|_\mathrm{F}^2} \cdot \left\{ \frac{[2\sigma_1(\mathcal{B}^{(n)}) + \|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_\mathrm{F} + \epsilon\|\mathcal{B}\|_\mathrm{F}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)}) + \sigma_{r_n+1}(\mathcal{B}^{(n)})]^2} \cdot \frac{[\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_\mathrm{F} + \epsilon\|\mathcal{B}\|_\mathrm{F}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)}) - \sigma_{r_n+1}(\mathcal{B}^{(n)})]^2} \right\} \\
&+ \frac{1}{\|\mathcal{B}\|_\mathrm{F}^2} \sum_{n=1}^N (\tau_{r_n}^{(n)})^2.
\end{aligned}$$

We denote $f(n) \sim g(n)$ if there exist universal constants C_1, C_2 and N_0 such that $C_1 g(n) \leq f(n) \leq C_2 g(n)$ for each $n > N_0$.

For an order-N cubical tensor $\mathcal{B}$ with size $I_1 = \cdots = I_N = I$, multilinear rank $r_1^\text{true} = \cdots = r_N^\text{true} = r < I$, and target multilinear rank $(r, r, \ldots, r)$, we choose $\epsilon \sim \sqrt{\frac{N \log I}{I}}$. In this scenario:

- From Lemma 6, we have

$$\frac{\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_\mathrm{F}}{\|\mathcal{B}\|_\mathrm{F}} \leq \frac{\alpha_\text{sp}}{\sigma(-\gamma)\sigma(-\alpha)} \sqrt{4eL_\gamma \tau \Big(\frac{1}{\sqrt{I_\square}} + \frac{1}{\sqrt{I_{\square^C}}}\Big)} \sim I^{-N/8} = O(\epsilon).$$

- When $I \geq rN \log I$, $\epsilon\|\mathcal{B}^{(n)}\|_\mathrm{F} = O(\frac{1}{\sqrt{r}}\|\mathcal{B}^{(n)}\|_\mathrm{F}) = O(\sigma_1(\mathcal{B}^{(n)}))$ for every $n \in [N]$.
- For every $n \in [N]$, the tail singular values $\sigma_j(\mathcal{B}^{(n)}) = 0$ for $j = r+1, \ldots, I$.

Thus in the upper bound of Theorem 5, Equation 2 above:

- The first term

$$\min_{n\in[N]}\left\{r_n\cdot\Big[\frac{\|\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\bar{\mathcal{X}}\|_{\mathrm{F}}}{\|\mathcal{B}\|_{\mathrm{F}}}+\epsilon\Big]^2\right\}=O(4r\epsilon^2).$$

- In the proof of Theorem 5, Term ⑤ and ⑥ vanish when $r_n^{\text{true}}\le r_n$, since $(U_n)_\perp=0$. Together with $\frac{\sigma_1(\mathcal{B}^{(n)})}{\sigma_r(\mathcal{B}^{(n)})}\le\kappa$ for every $n\in[N]$, the second term in the upper bound of Equation 2

$$\begin{aligned}
&\sum_{n=1}^{N}\frac{4r_n\sigma_1(\mathcal{B}^{(n)})^2}{\|\mathcal{B}\|_{\mathrm{F}}^2}\cdot\left\{\frac{[2\sigma_1(\mathcal{B}^{(n)})+\|\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\bar{\mathcal{X}}\|_{\mathrm{F}}+\epsilon\|\mathcal{B}\|_{\mathrm{F}}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)})+\sigma_{r_n+1}(\mathcal{B}^{(n)})]^2}\cdot\frac{[\|\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\bar{\mathcal{X}}\|_{\mathrm{F}}+\epsilon\|\mathcal{B}\|_{\mathrm{F}}]^2}{[\sigma_{r_n}(\mathcal{B}^{(n)})-\sigma_{r_n+1}(\mathcal{B}^{(n)})]^2}\right\}\\
&\le\sum_{n=1}^{N}\frac{4r\sigma_1(\mathcal{B}^{(n)})^2}{\|\mathcal{B}\|_{\mathrm{F}}^2}\cdot\left\{\frac{[4\sigma_1(\mathcal{B}^{(n)})]^2}{\sigma_{r_n}^2(\mathcal{B}^{(n)})}\cdot\frac{(2\epsilon\|\mathcal{B}\|_{\mathrm{F}})^2}{\sigma_{r_n}^2(\mathcal{B}^{(n)})}\right\}\\
&\le 256Nr\kappa^4\epsilon^2.
\end{aligned}$$

- The third term $\frac{1}{\|\mathcal{B}\|_{\mathrm{F}}^2}\sum_{n=1}^{N}(\tau_{r_n}^{(n)})^2=0$.

Together we have the simplified high-probability upper bound

$$\frac{\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B}\|_{\mathrm{F}}}{\|\mathcal{B}\|_{\mathrm{F}}}\le\epsilon\sqrt{4r+256Nr\kappa^4}=O\Big(N\sqrt{\frac{r\log I}{I}}\Big).$$

As for the probability lower bound $1-\dfrac{C_1}{I_\square+I_{\square^C}}-\displaystyle\sum_{n=1}^{N}[I_n+I_{(-n)}]\exp\Big[-\dfrac{\epsilon^2\|\mathcal{B}\|_{\mathrm{F}}^2\sigma(-\alpha)/2}{I_{(-n)}\psi^2+\epsilon\psi\|\mathcal{B}\|_{\mathrm{F}}/3}\Big]$:

- With the universal constant $C_1>0$, we have $\frac{C_1}{I_\square+I_\square^C}=O(I^{-1})$.
- The sum of probabilities from the matrix Bernstein inequality

$$\begin{aligned}
\sum_{n=1}^{N}[I_n+I_{(-n)}]\exp\Big[-\frac{\epsilon^2\|\mathcal{B}\|_{\mathrm{F}}^2\sigma(-\alpha)/2}{I_{(-n)}\psi^2+\epsilon\psi\|\mathcal{B}\|_{\mathrm{F}}/3}\Big]&=O(NI^{N-1}\cdot\exp\Big[-\frac{\epsilon^2\|\mathcal{B}\|_{\mathrm{F}}^2}{I^{N-1}}\Big])\\
&=O(NI^{N-1}\cdot\exp(-2\epsilon^2 I))\\
&=O(NI^{N-1}\cdot I^{-2N})\\
&=O(I^{-1}).
\end{aligned}$$

Thus the probability is at least $1-I^{-1}$. This concludes the proof for Theorem 4.

Appendix G. Gradient computation for Algorithm 3

For any $y\in\mathbb{R}$ and $X\in\mathbb{R}^{m\times n}$, we define the scalar-to-matrix derivative $\partial y/\partial X$ as a matrix of the same size as X, with the (i,j)-th entry $[\partial y/\partial X]_{ij}=\partial y/\partial X_{ij}$ for every $i\in[m]$, $j\in[n]$.

Recall that in Algorithm 3, we are using the gradient descent algorithm to minimize

$$f(\mathcal{G}^{\mathcal{A}}, \{U_n^{\mathcal{A}}\}_{n\in[N]}) = \sum_{i_1}^{I_1}\sum_{i_2}^{I_2}\cdots\sum_{i_N}^{I_N} - \Omega_{i_1\cdots i_N}\log\sigma[(\mathcal{G}^{\mathcal{A}}\times_1 U_1^{\mathcal{A}}\times_2\cdots\times_N U_N^{\mathcal{A}})_{i_1\cdots i_N}]$$
$$- (1-\Omega_{i_1\cdots i_N})\log\{1-\sigma[(\mathcal{G}^{\mathcal{A}}\times_1 U_1^{\mathcal{A}}\times_2\cdots\times_N U_N^{\mathcal{A}})_{i_1\cdots i_N}]\}, \tag{6}$$

in which σ is the link function. Denote $\widehat{\mathcal{A}} := \mathcal{G}^{\mathcal{A}}\times_1 U_1^{\mathcal{A}}\times_2\cdots\times_N U_N^{\mathcal{A}}$. When we use the logistic link function $\sigma(x) = 1/(1+e^{-x})$, f is the sum of entry-wise logistic losses between the true binary mask tensor Ω and the observation probability tensor $\sigma(\widehat{\mathcal{A}})$.

We first show the gradient of the logistic loss, and we omit the calculations.

Lemma 9 *(gradient of the logistic loss) For the logistic loss $\ell(x,y) = -y\log\sigma(x) - (1-y)\log(1-\sigma(x))$, we have $\partial\ell/\partial x = \sigma(x) - y$.*

We then show Lemma 10 for the chain rule of gradients of real-valued functions over matrices.

Lemma 10 *(chain rule of scalar-to-matrix derivatives) Let A be a matrix of size $m\times n$, and $g:\mathbb{R}\to\mathbb{R}$ be a continuously differentiable function. Define the real-valued function $\tilde{G}:\mathbb{R}^{m\times n}\to\mathbb{R}$ as*

$$\tilde{G}(A) = \sum_{i=1}^{m}\sum_{j=1}^{n} g(A_{ij}).$$

Then:

1. *If X, Y are matrices of size $m\times p$ and $p\times n$, respectively, and $A = XY$, then*

$$\frac{\partial\tilde{G}(A)}{\partial X} = \frac{\partial\tilde{G}(A)}{\partial A}Y^\top.$$

2. *If X, Y, Z are matrices of size $m\times p$, $p\times q$ and $q\times n$, respectively, and $A = XYZ$, then*

$$\frac{\partial\tilde{G}(A)}{\partial Y} = X^\top\frac{\partial\tilde{G}(A)}{\partial A}Z^\top.$$

Proof We show our proof in a similar fashion as [13, Lemma 2]. In Case 1,

$$\frac{\partial A_{kl}}{\partial X_{ij}} = \begin{cases} Y_{jl}, & \text{if } k = i \\ 0, & \text{if } k \neq i \end{cases}$$

for every $k, i\in[m]$, $l\in[n]$, $j\in[p]$. Thus

$$\begin{aligned}\frac{\partial\tilde{G}(A)}{\partial X_{ij}} &= \sum_{k=1}^{m}\sum_{l=1}^{n}\frac{\partial\tilde{G}(A)}{\partial A_{kl}}\frac{\partial A_{kl}}{\partial X_{ij}} \\ &= \sum_{l=1}^{n}\frac{\partial\tilde{G}(A)}{\partial A_{il}}Y_{jl} = \Big(\frac{\partial\tilde{G}(A)}{\partial A}Y^\top\Big)_{ij}.\end{aligned}$$

In Case 2, since $A_{kl} = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ki}Y_{ij}Z_{jl}$, we have $\frac{\partial A_{kl}}{\partial Y_{ij}} = X_{ki}Z_{jl}$. Thus

$$\begin{aligned}
\frac{\partial \tilde{G}(A)}{\partial Y_{ij}} &= \sum_{k=1}^{m}\sum_{l=1}^{n} \frac{\partial \tilde{G}(A)}{\partial A_{kl}} \frac{\partial A_{kl}}{\partial Y_{ij}} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{n} X_{ki} \frac{\partial \tilde{G}(A)}{\partial A_{kl}} Z_{jl} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{n} (X^\top)_{ik} \frac{\partial \tilde{G}(A)}{\partial A_{kl}} (Z^\top)_{lj} \\
&= \left(X^\top \frac{\partial \tilde{G}(A)}{\partial A} Z^\top\right)_{ij}.
\end{aligned}$$

These conclude the proof for Lemma 10 based on the definition of scalar-to-matrix derivatives. ■

Finally, we show the gradients $\{\partial f/\partial U_n\}_{n\in[N]}$ and $\partial f/\partial \mathcal{G}$ in Theorem 11.

Theorem 11 *(gradients of the objective function in Algorithm 3) For each $n \in [N]$, with*

$$\begin{aligned}
f(\mathcal{G}^{\mathcal{A}}, \{U_n^{\mathcal{A}}\}_{n\in[N]}) = \sum_{i_1}^{I_1}\sum_{i_2}^{I_2}\cdots\sum_{i_N}^{I_N} &- \Omega_{i_1\cdots i_N} \log \sigma[(\mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}})_{i_1\cdots i_N}] \\
&- (1-\Omega_{i_1\cdots i_N}) \log\{1 - \sigma[(\mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}})_{i_1\cdots i_N}]\},
\end{aligned}$$

and $\widehat{\mathcal{A}} = \mathcal{G}^{\mathcal{A}} \times_1 U_1^{\mathcal{A}} \times_2 \cdots \times_N U_N^{\mathcal{A}}$, we have:

1. *The gradient with respect to the factor matrix U_n*

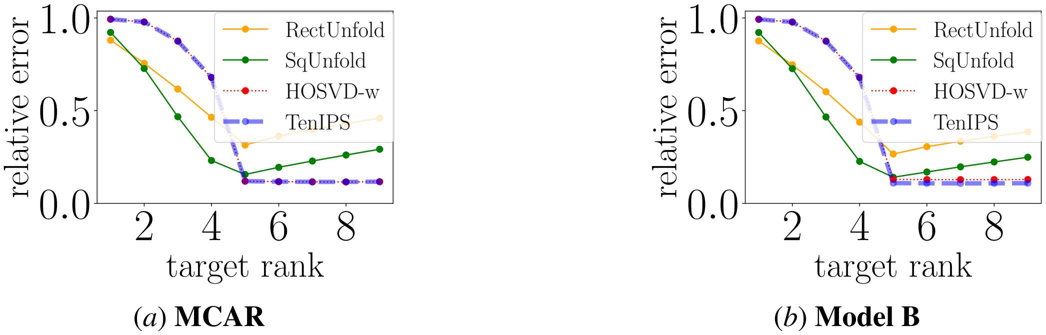
$$\frac{\partial f}{\partial U_n^{\mathcal{A}}} = \frac{\partial f}{\partial \widehat{\mathcal{A}}^{(n)}} \cdot \left(U_{n+1}^{\mathcal{A}} \otimes U_{n+2}^{\mathcal{A}} \otimes \cdots \otimes U_N^{\mathcal{A}} \otimes U_1^{\mathcal{A}} \otimes U_2^{\mathcal{A}} \otimes \cdots \otimes U_{n-1}^{\mathcal{A}}\right) \cdot [(\mathcal{G}^{\mathcal{A}})^{(n)}]^\top.$$

2. *The gradient with respect to the unfolded core tensor $(\mathcal{G}^{\mathcal{A}})^{(n)}$*

$$\frac{\partial f}{\partial (\mathcal{G}^{\mathcal{A}})^{(n)}} = (U_n^{\mathcal{A}})^\top \cdot \frac{\partial f}{\partial \widehat{\mathcal{A}}^{(n)}} \cdot \left(U_{n+1}^{\mathcal{A}} \otimes U_{n+2}^{\mathcal{A}} \otimes \cdots \otimes U_N^{\mathcal{A}} \otimes U_1^{\mathcal{A}} \otimes U_2^{\mathcal{A}} \otimes \cdots \otimes U_{n-1}^{\mathcal{A}}\right).$$

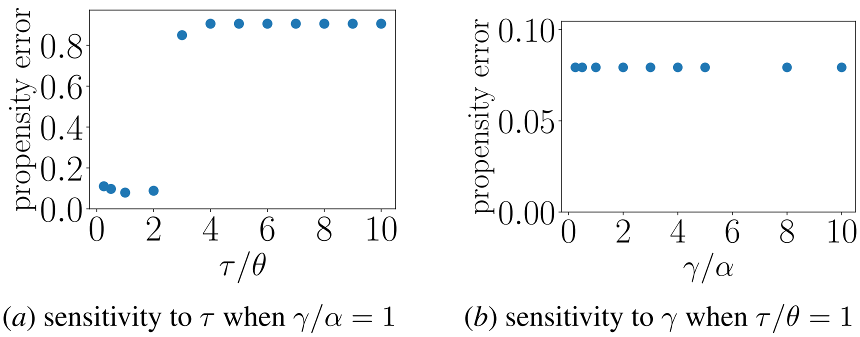
Proof With the Tucker decomposition of $\widehat{\mathcal{A}}$, we have $\widehat{\mathcal{A}}^{(n)} = U_n^{\mathcal{A}} \cdot (\mathcal{G}^{\mathcal{A}})^{(n)} \cdot \left(U_{n+1}^{\mathcal{A}} \otimes U_{n+2}^{\mathcal{A}} \otimes \cdots \otimes U_N^{\mathcal{A}} \otimes U_1^{\mathcal{A}} \otimes U_2^{\mathcal{A}} \otimes \cdots \otimes U_{n-1}^{\mathcal{A}}\right)^\top$ for the unfolding in each of the $n \in [N]$ [9]. Thus we can apply each case of Lemma 10 to the corresponding case here, with A to be $\widehat{\mathcal{A}}^{(n)}$. ■

With Lemma 9, we have $\partial f/\partial \widehat{\mathcal{A}} = \sigma(\widehat{\mathcal{A}}) - \Omega$ for the logistic link function σ. This can be inserted into Theorem 11 for the gradients $\{\partial f/\partial U_n\}_{n\in[N]}$ and $\partial f/\partial \mathcal{G}$, but note that Theorem 11 does not rely on this result.



(*a*) **MCAR** (*b*) **Model B**

Figure 5: Relative errors at different target ranks. True rank of the data tensors are 5 in every mode.



(*a*) sensitivity to τ when $\gamma/\alpha = 1$ (*b*) sensitivity to γ when $\tau/\theta = 1$

Figure 6: Hyperparameter sensitivity of Algorithm 1 to τ and γ.

Appendix H. Synthetic data: varying target ranks

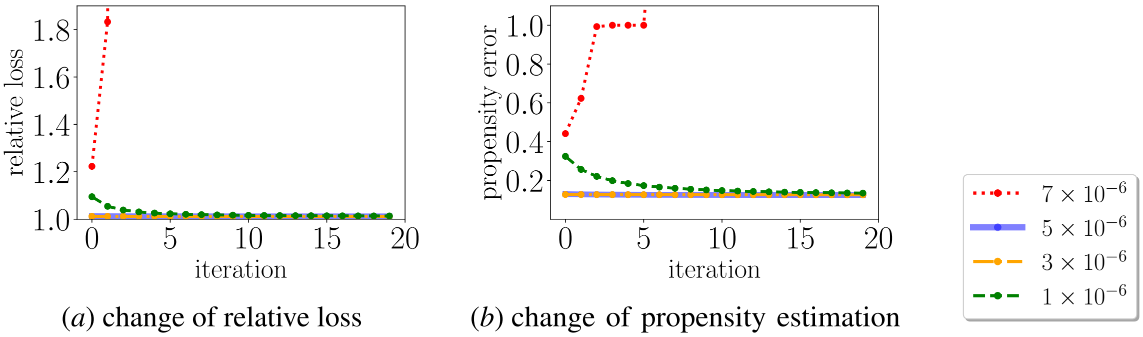
In Figure 5, we can see that both TENIPS and HOSVD_W are more stable at target ranks larger than the true rank, while RECTUNFOLD and SQUNFOLD achieve smaller errors at smaller ranks. This shows that TENIPS (and HOSVD_W) are robust to large target ranks in practice, despite the theory that $r_n = r_n^{\text{true}}$, for all $n \in [N]$.

Appendix I. Sensitivity of propensity estimation algorithms to hyperparameters

We study the sensitivities of the provable `prox-prox` Algorithm 1 and the alternative gradient descent Algorithm 3 to their respective hyperparameters.

The most important hyperparameters in Algorithm 1 are τ and γ. Ideally, we want to set $\tau = \theta$ and $\gamma = \alpha$; this is not possible in practice, though, since we do not know the θ and α of the true parameter tensor $\mathcal{A}$. In the setting of the third experiment in Section 5.1 of the main paper, we study the relationship between relative errors of propensity estimates and the ratios τ/θ and γ/α in Figure 6. We can see that the performance is much more sensitive to τ than γ, and a slight deviation of τ/θ from 1 results in a much larger propensity estimation error.

The most important hyperparameter in Algorithm 3 is the step size t. We show both the convergence and the change of propensity relative errors of Algorithm 3 at several step sizes in Figure 7. We can see that the relative errors of propensity estimates steadily decrease at all step sizes at which the gradient descent converges. Also, the respective rankings of relative losses and



(a) change of relative loss

(b) change of propensity estimation error

Figure 7: Hyperparameter sensitivity of Algorithm 3 to step size t. Since the objective function is the logistic loss between the mask tensor Ω and the parameter tensor $\mathcal{A}$, the relative loss in Figure 7a is the ratio of actual logistic loss to the best logistic loss computed from the true parameter tensor. Propensity error in Figure 7b is $\|\widehat{\mathcal{P}} - \mathcal{P}\|_{\mathrm{F}}/\|\mathcal{P}\|_{\mathrm{F}}$, the same as in the main paper.

propensity errors at different step sizes are the same across all iterations, indicating that the relative loss is a good surrogate metric for us to seek a good propensity estimate. Thus practitioners can select the largest step size at which Algorithm 3 converges; it is 5×10^{-6} in our practice. This is much easier than the selection of τ in Algorithm 1.