Appendix for TenIPS: Inverse Propensity Sampling for Tensor Completion

This appendix is organized as follows. Section A upper bounds the tensor completion error in the general case. Section B proves the upper bounds for both the general and the special cases. Section C computes the gradients in NonconvexPE for propensity estimation. Section D numerically studies the sensitivity of propensity estimation algorithms (ConvexPE and NonconvexPE) to their respective hyperparameters.

A Error in tensor completion (Algorithm 1 and 3): 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 $\|\mathcal{A}\|_\infty \leq \alpha$, $\|\mathcal{B}\|_\infty = \psi$. Denote the spikiness parameter $\alpha_{\text{sp}} := \psi\sqrt{I_{[N]}}/\|\mathcal{B}\|_{\text{F}}$. Then under the conditions of Lemma 2, with probability at least $1 - \dfrac{C_1}{I_\Box + I_{\Box^C}} - \sum_{n=1}^{N}[I_n + I_{(-n)}] \exp\Big[-\dfrac{\epsilon^2\|\mathcal{B}\|_{\text{F}}^2\sigma(-\alpha)/2}{I_{(-n)}\psi^2 + \epsilon\psi\|\mathcal{B}\|_{\text{F}}/3}\Big]$, 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 ConvexPE and TenIPS (Algorithms 1 and 3) with thresholds $\tau \geq \theta$ and $\gamma \geq \alpha$ satisfies*

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

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 $1 - \frac{C_1}{I_\Box + I_{\Box^C}}$,*

$$
\|\bar{\mathcal{X}}(\widehat{\mathcal{P}}) - \bar{\mathcal{X}}\|_{\text{F}} \leq \frac{\alpha_{\text{sp}}\|\mathcal{B}\|_{\text{F}}}{\sigma(-\gamma)\sigma(-\alpha)} \sqrt{4eL_\gamma\tau\Big(\frac{1}{\sqrt{I_\Box}} + \frac{1}{\sqrt{I_{\Box^C}}}\Big)}. \tag{2}
$$

On the right-hand side of Equation 1, 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.

Now we state Theorem 4 from the main paper, a corollary of the above Theorem 5 in the special case that the tensor is cubical and every unfolding has the same rank.

Theorem 4. *(Restated) 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 $\mathcal{B}_{i_1,\ldots,i_N}$ of $\mathcal{B}$ is observed with probability $\mathcal{P}_{i_1,\ldots,i_N}$ from the corresponding entry of $\mathcal{P}$. Assume $I \geq rN \log I$, and there exist constants $\psi, \alpha \in (0, \infty)$ such that $\|\mathcal{A}\|_\infty \leq \alpha$, $\|\mathcal{B}\|_\infty = \psi$. Further assume that 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 $(r, r, \ldots, r)$ approximation $\widehat{\mathcal{X}}(\widehat{\mathcal{P}})$ computed from* ConvexPE *and* TenIPS *(Algorithms 1 and 3) with thresholds $\tau \geq \theta$ and $\gamma \geq \alpha$ satisfies*

$$\frac{\|\widehat{\mathcal{X}}(\widehat{\mathcal{P}}) - \mathcal{B}\|_{\mathrm{F}}}{\|\mathcal{B}\|_{\mathrm{F}}} \leq CN\sqrt{\frac{r \log I}{I}}, \tag{3}$$

in which C depends on κ.

B Proof for Theorem 4 and 5

B.1 Proof for Theorem 5, the general case

We first show the proof for Theorem 5, the general case. This is the full version of the proof sketch in Section 5.2 of the main paper. We start with Lemma 6 on how the error in propensity estimates propagates 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 $\|\mathcal{B}\|_\infty = \psi$. 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,

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

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 $\sin(\Theta)$ Theorem [1].

Lemma 7 (matrix Bernstein for real matrices [2, Theorem 1.6.2])**.** *Let $S_1, \ldots, S_k$ be independent, centered random matrices with common dimension $m \times n$, and assume that each one is uniformly bounded*

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

Introduce the sum

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

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

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

Then

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

Lemma 8 (Variant of the Davis-Kahan $\sin(\Theta)$ Theorem [3], [4, 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 is 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

$$\mathcal{S}_{i_1i_2\cdots i_N} := \begin{cases} \left(\dfrac{1}{\mathcal{P}_{i_1i_2\cdots i_N}} - 1\right)\mathcal{B}\odot\mathcal{E}(i_1, i_2, \ldots, i_N), & \text{with probability } \mathcal{P}_{i_1i_2\cdots i_N} \\ -\mathcal{B}\odot\mathcal{E}(i_1, i_2, \ldots, i_N), & \text{with probability } 1-\mathcal{P}_{i_1i_2\cdots i_N}. \end{cases}$$

With the assumptions in Theorem 5, $\mathbb{E}\,\mathcal{S}_{i_1i_2\cdots i_N} = 0$ and $\|\mathcal{S}^{(n)}_{i_1i_2\cdots i_N}\| \leq \frac{\psi}{\sigma(-\alpha)}$. Also, the per-mode second moment is bounded as

$$\begin{aligned} v_n(\mathcal{X}) &= \max\{\|\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathbb{E}[\mathcal{S}^{(n)}_{i_1i_2\cdots i_N}(\mathcal{S}^{(n)}_{i_1i_2\cdots i_N})^\top]\|, \|\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_n=1}^{I_n}\mathbb{E}[(\mathcal{S}^{(n)}_{i_1i_2\cdots i_N})^\top\mathcal{S}^{(n)}_{i_1i_2\cdots i_N}]\|\} \\ &\leq \frac{\psi^2\cdot I_{(-n)}}{\sigma(-\alpha)}. \end{aligned}$$

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 $\mathcal{B}$

$$\begin{aligned} \bar{\mathcal{X}}(\mathcal{P}) - \mathcal{B} &= \sum_{(i_1,i_2,\ldots,i_N)\in\Omega}\frac{1}{\mathcal{P}_{i_1i_2\cdots i_N}}\mathcal{B}_{\mathrm{obs}}\odot\mathcal{E}(i_1, i_2, \ldots, i_N) - \mathcal{B} \\ &= \sum_{(i_1,i_2,\ldots,i_N)\in\Omega}\left(\frac{1}{\mathcal{P}_{i_1i_2\cdots i_N}} - 1\right)\mathcal{B}\odot\mathcal{E}(i_1, i_2, \ldots, i_N) - \sum_{(i_1,i_2,\ldots,i_N)\notin\Omega}\mathcal{B}\odot\mathcal{E}(i_1, i_2, \ldots, i_N) \end{aligned}$$

is an instance of $\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_N=1}^{I_N}\mathcal{S}_{i_1 i_2\cdots i_N}$ 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 TenIPS (Algorithm 3): In TenIPS,

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

This projects each unfolding of $\bar{\mathcal{X}}(\widehat{\mathcal{P}})$ onto the space of its top 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

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

First, we show that the cross term ③ is zero, since it is the product of two terms in 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

$$\begin{aligned}③&=2\sum_{n=1}^N\langle\mathcal{Y}_n-\mathcal{Y}_{n-1},(\bar{\mathcal{X}}(\widehat{\mathcal{P}})-\mathcal{B})\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top\rangle\\&=2\langle(Q_nQ_n^\top-I)\mathcal{Y}_{n-1}^{(n)},Q_n\mathcal{C}_n^{(n)}\rangle\\&=2\mathrm{tr}(\mathcal{Y}_{n-1}^{(n)}(Q_nQ_n^\top-I)Q_n\mathcal{C}_n^{(n)})=0.\end{aligned}$$

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

$$\mathcal{Y}_n=\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_n Q_nQ_n^\top.$$

Thus $\mathcal{B}\times_1 Q_1Q_1^\top\times_2\cdots\times_N Q_NQ_N^\top-\mathcal{B}=\mathcal{Y}_N-\mathcal{Y}_0=\sum_{n=1}^N(\mathcal{Y}_n-\mathcal{Y}_{n-1})$. Each $n\in[N]$ in the sum satisfies

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

This allows us to analyze each mode individually.

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

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

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

As for Term ②, it can be bounded using a technique similar to [5, Lemma B.1]. For each $n \in [N]$,

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

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

In the general case:

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

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

  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

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

  in which the first inequality comes from $\mathrm{tr}(AB) \le \lambda_1(A)\mathrm{tr}(B)$ 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.

B.2 Proof for Theorem 4, the special case

Recall the high-probability upper bound of Theorem 5, Equation 1 is

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

We denote $f(n) \sim g(n)$ if there exist universal constants C_1, C_2 and N_0 such that $C_1 g(n) \le f(n) \le C_2 g(n)$ 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

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

- 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 1 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 $r_n^{\text{true}} \leq r_n$, since $(U_n)_\perp = 0$. Together with $\frac{\sigma_1(\mathcal{B}^{(n)})}{\sigma_r(\mathcal{B}^{(n)})} \leq \kappa$ for every $n \in [N]$, the second term in the upper bound of Equation 1

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

- 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

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

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

- 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

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

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

C Gradient computation for NonconvexPE (Algorithm 2)

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 NonconvexPE, we use the gradient descent algorithm to minimize

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

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 7. *(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 8 for the chain rule of gradients of real-valued functions over matrices.

Lemma 8. *(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 [6, 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 $\dfrac{\partial A_{kl}}{\partial Y_{ij}} = X_{ki}Z_{jl}$. 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 8 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 9.

Theorem 9. *(gradients of the objective function in Algorithm 2) 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 $n \in [N]$ [7]. Thus we can apply each case of Lemma 8 to the corresponding case here, with A to be $\widehat{\mathcal{A}}^{(n)}$. □

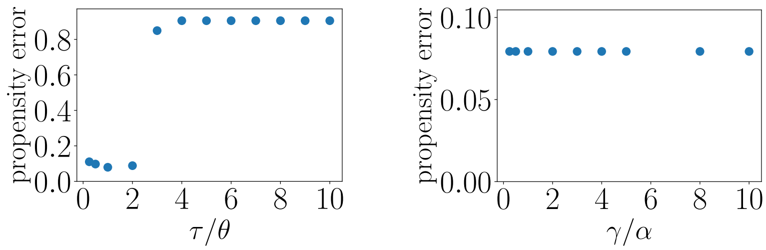
With Lemma 7, we have $\partial f/\partial \widehat{\mathcal{A}} = \sigma(\widehat{\mathcal{A}}) - \Omega$ for the logistic link function σ. This can be inserted into Theorem 9 for the gradients $\{\partial f/\partial U_n\}_{n\in[N]}$ and $\partial f/\partial \mathcal{G}$, but note that Theorem 9 does not rely on this result.

D Sensitivity of propensity estimation algorithms to hyperparameters

We study the sensitivities of ConvexPE (Algorithm 1) and NononvexPE (Algorithm 2) to their respective hyperparameters.

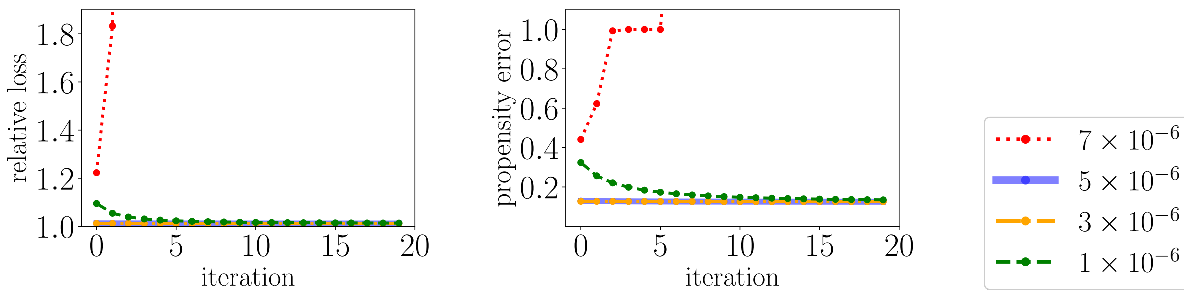
The most important hyperparameters in ConvexPE are τ and γ. Ideally, we want to set $\tau = \theta$ and $\gamma = \alpha$; this is not possible in practice, though, since we do not know the θ and α of the true parameter tensor $\mathcal{A}$. In the setting of the third experiment in Section 6.1 of the 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 NonconvexPE is the step size t. We show both the convergence and the change of propensity relative errors at various step sizes in Figure 7. We can see that the relative errors



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

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



(a) change of relative loss (b) change of propensity estimation error

Figure 7: Hyperparameter sensitivity of NononvexPE (Algorithm 2) 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.

of propensity estimates steadily decrease at all step sizes at which the gradient descent converges. Also, the respective rankings of relative losses and 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 NonconvexPE converges; it is 5×10^{-6} in our practice. This is much easier than the selection of τ in ConvexPE.

References

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

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

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

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

[5] 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.

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

[7] 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.