Supplements for “Tensor denoising and completion based on ordinal observations”

1 Proofs

Here, we provide proofs of the theoretical results presented in Sections 4.

1.1 Estimation error for tensor denoising

Proof of Theorem 4.1. We suppress the subscript Ω in the proof, because the tensor denoising assumes complete observation $\Omega = [d_1] \times \cdots \times [d_K]$. It follows from the expression of $\mathcal{L}_{\mathcal{Y}}(\Theta)$ that

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega} &= \sum_{\ell\in[L]} \mathbb{1}_{\{y_\omega=\ell\}} \frac{\dot{g}_\ell(\theta_\omega)}{g_\ell(\theta_\omega)},\\ \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega^2} &= \sum_{\ell\in[L]} \mathbb{1}_{\{y_\omega=\ell\}} \frac{\ddot{g}_\ell(\theta_\omega)g_\ell(\theta_\omega)-\dot{g}_\ell^2(\theta_\omega)}{g_\ell^2(\theta_\omega)} \quad \text{and} \quad \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega \theta'_\omega} = 0 \text{ if } \omega \neq \omega',\end{aligned} \tag{1}$$

for all $\omega \in [d_1] \times \cdots \times [d_K]$. Define $d_{\text{total}} = \prod_k d_k$. Let $\nabla_\Theta \mathcal{L}_{\mathcal{Y}} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ denote the tensor of gradient with respect to $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$, and $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ the corresponding Hession matrix of size d_{total}-by-d_{total}. Here, $\text{Vec}(\cdot)$ denotes the operation that turns a tensor into a vector. By (1), $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ is a diagonal matrix. Recall that

$$U_\alpha = \max_{\ell\in[L],|\theta|\leq\alpha} \frac{|\dot{g}_\ell(\theta)|}{g_\ell(\theta)} > 0 \quad \text{and} \quad L_\alpha = \min_{\ell\in[L],|\theta|\leq\alpha} \frac{\dot{g}_\ell^2(\theta)-\ddot{g}_\ell(\theta)g_\ell(\theta)}{g_\ell^2(\theta)} > 0. \tag{2}$$

Therefore, the entries in $\nabla_\Theta \mathcal{L}_{\mathcal{Y}}$ are upper bounded in magnitude by $U_\alpha > 0$, and all diagonal entries in $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ are upper bounded by $-L_\alpha < 0$.

By the second-order Taylor's expansion of $\mathcal{L}_{\mathcal{Y}}(\Theta)$ around Θ^{true}, we obtain

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) + \langle \text{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y}}),\ \text{Vec}(\Theta-\Theta^{\text{true}})\rangle + \frac{1}{2}\text{Vec}(\Theta-\Theta^{\text{true}})^T \nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta})\,\text{Vec}(\Theta-\Theta^{\text{true}}), \tag{3}$$

$\check{\Theta} = \gamma\Theta^{\text{true}} + (1-\gamma)\Theta$ for some $\gamma \in [0,1]$, and $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta})$ denotes the $\prod_k d_k$-by-$\prod_k d_k$ Hession matrix evaluated at $\check{\Theta}$.

We first bound the linear term in (3). Note that, by Lemma 4,

$$|\langle \text{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}), \text{Vec}(\Theta-\Theta^{\text{true}})\rangle| \leq \|\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}})\|_\sigma \|\Theta-\Theta^{\text{true}}\|_*, \tag{4}$$

where $\|\cdot\|_\sigma$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. Define

$$s_\omega = \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega}\Big|_{\Theta=\Theta^{\text{true}}} \quad \text{for all } \ \omega \in [d_1] \times \cdots \times [d_K].$$

Based on (1) and the definition of U_α, $\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) = [\![s_\omega]\!]$ is a random tensor whose entries are independently distributed satisfying

$$\mathbb{E}(s_\omega) = 0, \quad |s_\omega| \leq U_\alpha, \quad \text{for all } \omega \in [d_1] \times \cdots \times [d_K]. \tag{5}$$

By lemma 6, with probability at least $1-\exp(-C_1\sum_k d_k)$, we have

$$\|\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}})\|_\sigma \le C_2 U_\alpha \sqrt{\sum_k d_k}, \tag{6}$$

where C_1, C_2 are two positive constants that depend only on K. Furthermore, note that $\text{rank}(\Theta) \le \boldsymbol{r}$, $\text{rank}(\Theta^{\text{true}}) \le \boldsymbol{r}$, so $\text{rank}(\Theta - \Theta^{\text{true}}) \le 2\boldsymbol{r}$. By lemma 3, $\|\Theta - \Theta^{\text{true}}\|_* \le (2r_{\max})^{\frac{K-1}{2}} \|\Theta - \Theta^{\text{true}}\|_F$. Combining (4), (5) and (6), we have that, with probability at least $1-\exp(-C_1\sum_k d_k)$,

$$|\langle \text{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}})), \text{Vec}(\Theta - \Theta^{\text{true}})\rangle| \le C_2 U_\alpha \sqrt{r_{\max}^{K-1}\sum_k d_k}\|\Theta - \Theta^{\text{true}}\|_F. \tag{7}$$

We next bound the quadratic term in (3). Note that

$$\begin{aligned}
\text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta})\, \text{Vec}(\Theta - \Theta^{\text{true}}) &= \sum_\omega \left(\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_\omega^2}\Big|_{\Theta = \check{\Theta}} \right) (\theta_\omega - \theta_{\text{true},\omega})^2 \\
&\le -L_\alpha \sum_\omega (\Theta_\omega - \Theta_{\text{true},\omega})^2 \\
&= -L_\alpha \|\Theta - \Theta^{\text{true}}\|_F^2,
\end{aligned} \tag{8}$$

where the second line comes from the fact that $\|\check{\Theta}\|_\infty \le \alpha$ and the definition of L_α.

Combining (3), (7) and (8), we have that, for all $\Theta \in \mathcal{P}$, with probability at least $1 - \exp(-C_1 \sum_k d_k)$,

$$\mathcal{L}_{\mathcal{Y}}(\Theta) \le \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) + C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\Theta - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2}\|\Theta - \Theta^{\text{true}}\|_F^2.$$

In particular, the above inequality also holds for $\hat{\Theta} \in \mathcal{P}$. Therefore,

$$\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}) \le \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) + C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2.$$

Since $\hat{\Theta} = \arg\max_{\Theta \in \mathcal{P}} \mathcal{L}_{\mathcal{Y}}(\Theta)$, $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}) - \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}) \ge 0$, which gives

$$C_2 U_\alpha \left(r_{\max}^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F - \frac{L_\alpha}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \ge 0.$$

Henceforth,

$$\frac{1}{\sqrt{\prod_k d_k}} \|\hat{\Theta} - \Theta^{\text{true}}\|_F \le \frac{2C_2 U_\alpha \sqrt{r_{\max}^{K-1} \sum_k d_k}}{L_\alpha \sqrt{\prod_k d_k}} = \frac{2C_2 U_\alpha r_{\max}^{(K-1)/2}}{L_\alpha} \sqrt{\frac{\sum_k d_k}{\prod_k d_k}}.$$

This completes the proof. □

Proof of Corollary 1. The result follows immediately from Theorem 4.1 and Lemma 8. □

Proof of Theorem 4.2. Let $d_{\text{total}} = \prod_{k\in[K]} d_k$, and $\gamma \in [0,1]$ be a constant to be specified later. Our strategy is to construct a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{P}$ satisfying the properties of (i)-(iv) in Lemma 9. By Lemma 9, such a subset of tensors exist. For any tensor $\Theta \in \mathcal{X}$, let $\mathbb{P}_\Theta$ denote the distribution of $\mathcal{Y}|\Theta$, where $\mathcal{Y}$ is the ordinal tensor. In particular, $\mathbb{P}_{\mathbf{0}}$ is the distribution of $\mathcal{Y}$ induced by the zero parameter tensor $\mathbf{0}$, i.e., the distribution of $\mathcal{Y}$ conditional on the parameter tensor $\Theta = \mathbf{0}$. Based on the Remark for Lemma 8, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \le C\|\Theta\|_F^2, \tag{9}$$

where $C = \frac{(4L-6)\dot{f}^2(0)}{A_\alpha} > 0$ is a constant independent of the tensor dimension and rank. Combining the inequality (9) with property (iii) of $\mathcal{X}$, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \le \gamma^2 R_{\max} d_{\max}. \tag{10}$$

From (10) and the property (i), we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{X}) - 1} \sum_{\Theta\in\mathcal{X}} \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \le \varepsilon \log_2 \{\mathrm{Card}(\mathcal{X}) - 1\} \tag{11}$$

holds for any $\varepsilon \ge 0$ when $\gamma \in [0,1]$ is chosen to be sufficiently small depending on ε, e.g., $\gamma \le \sqrt{\frac{\varepsilon \log 2}{8}}$. By applying Lemma 11 to (11), and in view of the property (iv), we obtain that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\text{true}}\in\mathcal{X}} \mathbb{P}\left(\|\hat{\Theta} - \Theta^{\text{true}}\|_F \ge \frac{\gamma}{8} \min\left\{ \alpha\sqrt{d_{\text{total}}}, C^{-1/2}\sqrt{R_{\max} d_{\max}} \right\} \right) \ge \frac{1}{2}\left(1 - 2\varepsilon - \sqrt{\frac{16\varepsilon}{R_{\max} d_{\max} \log 2}} \right). \tag{12}$$

Note that $\mathrm{Loss}(\hat{\Theta}, \Theta^{\text{true}}) = \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 / d_{\text{total}}$ and $\mathcal{X} \subset \mathcal{P}$. By taking $\varepsilon = 1/10$ and $\gamma = 1/11$, we conclude from (12) that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\text{true}}\in\mathcal{P}} \mathbb{P}\left(\mathrm{Loss}(\hat{\Theta}, \Theta^{\text{true}}) \ge c \min\left\{ \alpha^2, \frac{C^{-1} R_{\max} d_{\max}}{d_{\text{total}}} \right\} \right) \ge \frac{1}{2}\left(\frac{4}{5} - \sqrt{\frac{1.6}{R_{\max} d_{\max} \log 2}} \right) \ge \frac{1}{8},$$

where $c = \frac{1}{88^2}$ and the last inequality comes from the condition for $d_{\max}$. This completes the proof. $\square$

1.2 Sample complexity for tensor completion

Proof of Theorem 4.3. For notational convenience, we use $\|\Theta\|_{F,\Omega} = \sum_{\omega\in\Omega} \Theta_\omega^2$ to denote the sum of squared entries over the observed set Ω, for a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$.

Following a similar argument as in the proof of Theorem 4.1, we have

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta) = \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}}) + \langle \mathrm{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}),\ \mathrm{Vec}(\Theta - \Theta^{\text{true}}) \rangle + \frac{1}{2} \mathrm{Vec}(\Theta - \Theta^{\text{true}})^T \nabla_\Theta^2 \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta})\, \mathrm{Vec}(\Theta - \Theta^{\text{true}}), \tag{13}$$

where

1. $\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}$ is a $d_1 \times \cdots \times d_K$ tensor with $|\Omega|$ nonzero entries, and each entry is upper bounded by $U_\alpha > 0$.

2. $\nabla_\Theta^2 \mathcal{L}_{\mathcal{Y},\Omega}$ is a diagonal matrix of size d_{total}-by-d_{total} with $|\Omega|$ nonzero entries, and each entry is upper bounded by $-L_\alpha < 0$.

Similar to (4) and (8), we have

$$|\langle \mathrm{Vec}(\nabla_\Theta \mathcal{L}_{\mathcal{Y},\Omega}),\ \mathrm{Vec}(\Theta-\Theta^{\mathrm{true}})\rangle| \leq C_2 U_\alpha \sqrt{r_{\max}^{K-1}\sum_k d_k}\|\Theta-\Theta^{\mathrm{true}}\|_{F,\Omega}$$

and

$$\mathrm{Vec}(\Theta-\Theta^{\mathrm{true}})^T \nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}(\check{\Theta})\,\mathrm{Vec}(\Theta-\Theta^{\mathrm{true}}) \leq -L_\alpha \|\Theta-\Theta^{\mathrm{true}}\|^2_{F,\Omega}. \tag{14}$$

Combining (13)-(14) with the fact that $\mathcal{L}_{\mathcal{Y},\Omega}(\hat{\Theta}) \geq \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\mathrm{true}})$, we have

$$\|\hat{\Theta}-\Theta^{\mathrm{true}}\|_{F,\Omega} \leq \frac{2C_2 U_\alpha r_{\max}^{(K-1)/2}}{L_\alpha}\sqrt{\sum_k d_k}, \tag{15}$$

with probability at least $1-\exp(-C_1\sum_k d_k)$. Lastly, we invoke the result regarding the closeness of Θ to its sampled version Θ_Ω, under the entrywise bound condition. Note that $\|\hat{\Theta}-\Theta^{\mathrm{true}}\|_\infty \leq 2\alpha$ and $\mathrm{rank}(\hat{\Theta}-\Theta^{\mathrm{true}}) \leq 2\boldsymbol{r}$. By Lemma 2, $\|\hat{\Theta}-\Theta^{\mathrm{true}}\|_M \leq 2^{(3K-1)/2}\alpha\left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Therefore, the condition in Lemma 12 holds with $\beta = 2^{(3K-1)/2}\alpha\left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Applying Lemma 12 to (15) gives

$$\begin{aligned}\|\hat{\Theta}-\Theta^{\mathrm{true}}\|^2_{F,\Pi} &\leq \frac{1}{m}\|\hat{\Theta}-\Theta^{\mathrm{true}}\|^2_{F,\Omega} + c\beta\sqrt{\frac{\sum_k d_k}{|\Omega|}} \\ &\leq C_2 r_{\max}^{K-1}\frac{\sum_k d_k}{|\Omega|} + C_1\alpha r_{\max}^{3(K-1)/2}\sqrt{\frac{\sum_k d_k}{|\Omega|}},\end{aligned}$$

with probability at least $1-\exp(-\frac{\sum_k d_k}{\sum_k \log d_k})$ over the sampled set Ω. Here $C_1, C_2 > 0$ are two constants independent of the tensor dimension and rank. Therefore,

$$\|\hat{\Theta}-\Theta^{\mathrm{true}}\|^2_{F,\Pi} \to 0, \quad \text{as} \quad \frac{|\Omega|}{\sum_k d_k} \to \infty,$$

provided that $r_{\max} = O(1)$. □

1.3 Convexity of the log-likelihood function

Theorem 1.1. *Define the function*

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}) = \sum_{\omega\in\Omega}\sum_{\ell\in[L]}\left\{\mathbb{1}_{\{y_\omega=\ell\}}\log\left[f(b_\ell-\theta_\omega)-f(b_{\ell-1}-\theta_\omega)\right]\right\}, \tag{16}$$

where $f(\cdot)$ satisfies Assumption 1. Then, $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta,\boldsymbol{b})$ is concave in $(\Theta,\boldsymbol{b})$.

Proof. Define $d_{\mathrm{total}} = \prod_k d_k$. By abuse of notation, we use $(\Theta,\boldsymbol{b})$ to denote the length-$(d_{\mathrm{total}}+L-1)$-vector collecting all parameters together. Let us denote a bivariate function

$$\begin{aligned}\lambda :& \mathbb{R}^2 \mapsto \mathbb{R} \\ (u,v) \mapsto& \lambda(u,v) = \log\left[f(u)-f(v)\right].\end{aligned}$$

It suffices to show that $\lambda(u,v)$ is concave in (u,v) where $u > v$.

Suppose that the claim holds (which we will prove in the next paragraph). Based on (16), u, v are both linear functions of $(\Theta, \boldsymbol{b})$:

$$u = \boldsymbol{a}_1^T(\Theta, \boldsymbol{b}), \quad v = \boldsymbol{a}_2^T(\Theta, \boldsymbol{b}), \quad \text{for some } \boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^{d_{\text{total}}+L-1}.$$

Then, $\lambda(u, v) = \lambda(\boldsymbol{a}_1^T(\Theta, \boldsymbol{b}),\ \boldsymbol{a}_2^T(\Theta, \boldsymbol{b}))$ is concave in $(\Theta, \boldsymbol{b})$ by the definition of concavity. Therefore, we can conclude that $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$ is concave in $(\Theta, \boldsymbol{b})$ because $\mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b})$ is the sum of $\lambda(u, v)$.

Now, we prove the concavity of $\lambda(u, v)$. Note that

$$\lambda(u, v) = \log\big[f(u) - f(v)\big] = \log\Big[\int \mathbb{1}_{[u,v]}(x) f'(x) dx\Big],$$

where $\mathbb{1}_{[u,v]}$ is an indicator function that equals 1 in the interval $[u, v]$, and 0 elsewhere. Furthermore, $\mathbb{1}_{[u,v]}(x)$ is log-concave in (u, v, x), and by Assumption 1, $f'(x)$ is log-concave in x. It follows that $\mathbb{1}_{[u,v]}(x) f'(x)$ is a log-concave in (u, v, x). By Lemma 1, we conclude that $\lambda(u, v)$ is concave in (u, v) where $u > v$. □

Lemma 1 (Corollary 3.5 in Brascamp and Lieb [2002]). *Let $F(x, y) : \mathbb{R}^{m+n} \to \mathbb{R}$ be an integrable function where $x \in \mathbb{R}^m, y \in \mathbb{R}^n$. Let*

$$G(x) = \int_{\mathbb{R}^n} F(x, y) dy.$$

If $F(x, y)$ is log concave in (x, y), then $G(x)$ is log concave in x.

1.4 Auxiliary lemmas

This section collects lemmas that are useful for the proofs of the main theorems.

Definition 1 (Atomic M-norm [Ghadermarzy et al., 2019]). Define $T_{\pm} = \{\mathcal{T} \in \{\pm 1\}^{d_1 \times \cdots \times d_K} : \operatorname{rank}(\mathcal{T}) = 1\}$. The atomic M-norm of a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined as

$$\begin{aligned} \|\Theta\|_M &= \inf\{t > 0 : \Theta \in t\operatorname{conv}(T_{\pm})\} \\ &= \inf\left\{\sum_{\mathcal{X} \in T_{\pm}} c_x : \Theta = \sum_{\mathcal{X} \in T_{\pm}} c_x \mathcal{X},\ c_x > 0\right\}. \end{aligned}$$

Definition 2 (Spectral norm [Lim, 2005]). The spectral norm of a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined as

$$\|\Theta\|_\sigma = \sup\Big\{\langle \Theta, \boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_K\rangle : \|\boldsymbol{x}_k\|_2 = 1,\ \boldsymbol{x}_k \in \mathbb{R}^{d_k},\ \text{for all } k \in [K]\Big\}.$$

Definition 3 (Nuclear norm [Friedland and Lim, 2018]). The nuclear norm of a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined as

$$\|\Theta\|_* = \inf\left\{\sum_{i \in [r]} |\lambda_i| : \Theta = \sum_{i=1}^r \lambda_i \boldsymbol{x}_1^{(i)} \otimes \cdots \otimes \boldsymbol{x}_K^{(i)},\ \|\boldsymbol{x}_k^{(i)}\|_2 = 1,\ \boldsymbol{x}_k^{(i)} \in \mathbb{R}^{d_k},\ \text{for all} k \in [K],\ i \in [r]\right\},$$

where the infimum is taken over all $r \in \mathbb{N}$ and $\|\boldsymbol{x}_k^{(i)}\|_2 = 1$ for all $i \in [r]$ and $k \in [K]$.

Lemma 2 (M-norm and infinity norm [Ghadermarzy et al., 2019]). *Let $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be an order-K, rank-$(r_1, \ldots, r_K)$ tensor. Then*

$$\|\Theta\|_\infty \leq \|\Theta\|_M \leq \left(\frac{\prod_k r_k}{r_{\max}}\right)^{\frac{3}{2}} \|\Theta\|_\infty.$$

Lemma 3 (Nuclear norm and F-norm)**.** *Let $\mathcal{A} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ be an order-K tensor with Tucker rank$(\mathcal{A}) = (r_1, \ldots, r_K)$. Then*

$$\|\mathcal{A}\|_* \le \sqrt{\frac{\prod_k r_k}{\max_k r_k}}\|\mathcal{A}\|_F,$$

where $\|\cdot\|_*$ denotes the nuclear norm of the tensor.

Proof. Without loss of generality, suppose $r_1 = \min_k r_k$. Let $\mathcal{A}_{(k)}$ denote the mode-k matricization of $\mathcal{A}$ for all $k \in [K]$. By Wang et al. [2017, Corollary 4.11], and the invariance relationship between a tensor and its Tucker core [Jiang et al., 2017, Section 6], we have

$$\|\mathcal{A}\|_* \le \sqrt{\frac{\prod_{k\ge 2} r_k}{\max_{k\ge 2} r_k}}\|\mathcal{A}_{(1)}\|_*, \tag{17}$$

where $\mathcal{A}_{(1)}$ is a d_1-by-$\prod_{k\ge 2} d_k$ matrix with matrix rank r_1. Furthermore, the relationship between the matrix norms implies that $\|\mathcal{A}_{(1)}\|_* \le \sqrt{r_1}\|\mathcal{A}_{(1)}\|_F = \sqrt{r_1}\|\mathcal{A}\|_F$. Combining this fact with the inequality (17) yields the final claim. □

Lemma 4. *Let $\mathcal{A}$, $\mathcal{B}$ be two order-K tensors of the same dimension. Then*

$$|\langle \mathcal{A}, \mathcal{B}\rangle| \le \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*.$$

Proof. By Friedland and Lim [2018, Proposition 3.1], there exists a nuclear norm decomposition of $\mathcal{B}$, such that

$$\mathcal{B} = \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}, \quad \boldsymbol{a}_r^{(k)} \in \mathbf{S}^{d_k-1}(\mathbb{R}), \quad \text{for all } k \in [K],$$

and $\|\mathcal{B}\|_* = \sum_r |\lambda_r|$. Henceforth we have

$$\begin{aligned}|\langle \mathcal{A}, \mathcal{B}\rangle| &= |\langle \mathcal{A}, \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}\rangle| \le \sum_r |\lambda_r| |\langle \mathcal{A}, \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}\rangle| \\ &\le \sum_r |\lambda_r| \|\mathcal{A}\|_\sigma = \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*,\end{aligned}$$

which completes the proof. □

The following lemma provides the bound on the spectral norm of random tensors. The result was firstly presented in Nguyen et al. [2015], and we adopt the version from Tomioka and Suzuki [2014].

Lemma 5 (Spectral norm of random tensors [Tomioka and Suzuki, 2014])**.** *Suppose that $\mathcal{S} = [\![s_\omega]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_\omega) = 0, \quad \textit{and} \quad \mathbb{E}(e^{ts_\omega}) \le e^{t^2L^2/2}.$$

Then the spectral norm $\|\mathcal{S}\|_\sigma$ satisfies that,

$$\|\mathcal{S}\|_\sigma \le \sqrt{8L^2 \log(12K) \sum_k d_k + \log(2/\delta)},$$

with probability at least $1 - \delta$.

Lemma 6. *Suppose that $\mathcal{S} = [\![s_\omega]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_\omega) = 0, \quad \text{and} \quad |s_\omega| \le U.$$

Then we have

$$\mathbb{P}\left(\|\mathcal{S}\|_\sigma \ge C_2 U \sqrt{\sum_k d_k} \right) \le \exp\left(-C_1 \log K \sum_k d_k \right)$$

where $C_1 > 0$ is an absolute constant, and $C_2 > 0$ is a constant that depends only on K.

Proof. Note that the random variable $U^{-1}s_\omega$ is zero-mean and supported on $[-1, 1]$. Therefore, $U^{-1}s_\omega$ is sub-Gaussian with parameter $\frac{1-(-1)}{2} = 1$, i.e.

$$\mathbb{E}(U^{-1}s_\omega) = 0, \quad \text{and} \quad \mathbb{E}(e^{tU^{-1}s_\omega}) \le e^{t^2/2}.$$

It follows from Lemma 5 that, with probability at least $1-\delta$,

$$\|U^{-1}\mathcal{S}\|_\sigma \le \sqrt{(c_0 \log K + c_1)\sum_k d_k + \log(2/\delta)},$$

where $c_0, c_1 > 0$ are two absolute constants. Taking $\delta = \exp(-C_1 \log K \sum_k d_k)$ yields the final claim, where $C_2 = c_0 \log K + c_1 + 1 > 0$ is another constant. $\square$

Lemma 7. *Let X, Y be two discrete random variables taking values on L possible categories, with category probabilities $\{p_\ell\}_{\ell\in[L]}$ and $\{q_\ell\}_{\ell\in[L]}$, respectively. Suppose p_ℓ, $q_\ell > 0$ for all $i \in [L]$. Then, the Kullback-Leibler (KL) divergence satisfies that*

$$\mathrm{KL}(X||Y) \stackrel{def}{=} -\sum_{\ell\in[L]} \mathbb{P}_X(\ell) \log\left\{ \frac{\mathbb{P}_Y(\ell)}{\mathbb{P}_X(\ell)} \right\} \le \sum_{\ell\in[L]} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

Proof. Using the fact $\log x \le x - 1$ for $x > 0$, we have that

$$\begin{aligned}
\mathrm{KL}(X||Y) &= \sum_{\ell\in[L]} p_\ell \log \frac{p_\ell}{q_\ell} \\
&\le \sum_{\ell\in[L]} \frac{p_\ell}{q_\ell}(p_\ell - q_\ell) \\
&= \sum_{\ell\in[L]} \left(\frac{p_\ell}{q_\ell} - 1 \right)(p_\ell - q_\ell) + \sum_{\ell\in[L]} (p_\ell - q_\ell).
\end{aligned}$$

Note that $\sum_{\ell\in[L]}(p_\ell - q_\ell) = 0$. Therefore,

$$\mathrm{KL}(X||Y) \le \sum_{\ell\in[L]} \left(\frac{p_\ell}{q_\ell} - 1 \right)(p_\ell - q_\ell) = \sum_{\ell\in[L]} \frac{(p_\ell - q_\ell)^2}{q_\ell}.$$

$\square$

Lemma 8 (KL divergence and F-norm). *Let $\mathcal{Y} \in [L]^{d_1 \times \cdots \times d_K}$ be an ordinal tensor generated from the model (1) with the link function f and parameter tensor Θ. Let $\mathbb{P}_\Theta$ denote the joint categorical distribution of $\mathcal{Y}|\Theta$ induced by the parameter tensor Θ, where $\|\Theta\|_\infty \leq \alpha$. Define*

$$A_\alpha = \min_{\ell \in [L], |\theta| \leq \alpha} \left[f(b_\ell - \theta) - f(b_{\ell-1} - \theta) \right]. \tag{18}$$

Then, for any two tensors Θ, Θ^* in the parameter spaces, we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\Theta^*}) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta - \Theta^*\|_F^2.$$

Proof. Suppose that the distribution over the ordinal tensor $\mathcal{Y} = [\![y_\omega]\!]$ is induced by $\Theta = [\![\theta_\omega]\!]$. Then, based on the generative model (1),

$$\mathbb{P}(y_\omega = \ell | \theta_\omega) = f(b_\ell - \theta_\omega) - f(b_{\ell-1} - \theta_\omega),$$

for all $\ell \in [L]$ and $\omega \in [d_1] \times \cdots \times [d_K]$. For notational convenience, we suppress the subscribe in θ_ω and simply write θ (and respectively, θ^*). Based on Lemma 7 and Taylor expansion,

$$\begin{aligned}
\mathrm{KL}(\theta || \theta^*) &\leq \sum_{\ell \in [L]} \frac{\left[f(b_\ell - \theta) - f(b_{\ell-1} - \theta) - f(b_\ell - \theta^*) + f(b_{\ell-1} - \theta^*) \right]^2}{f(b_\ell - \theta^*) - f(b_{\ell-1} - \theta^*)} \\
&\leq \sum_{\ell=2}^{L-1} \frac{\left[\dot{f}(b_\ell - \eta_\ell) - \dot{f}(b_{\ell-1} - \eta_{\ell-1}) \right]^2}{f(b_\ell - \theta^*) - f(b_{\ell-1} - \theta^*)} (\theta - \theta^*)^2 + \frac{\dot{f}^2(b_1 - \eta_1)}{f(b_1 - \theta^*)} (\theta - \theta^*)^2 \\
&\quad + \frac{\dot{f}^2(b_{L-1} - \eta_{L-1})}{1 - f(b_{L-1} - \theta^*)} (\theta - \theta^*)^2,
\end{aligned}$$

where η_ℓ and $\eta_{\ell-1}$ fall between θ and θ^*. Therefore,

$$\mathrm{KL}(\theta || \theta^*) \leq \left(\frac{4(L-2)}{A_\alpha} + \frac{2}{A_\alpha} \right) \dot{f}^2(0) (\theta - \theta^*)^2 = \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) (\theta - \theta^*)^2, \tag{19}$$

where we have used Taylor expansion, the bound (18), and the fact that $\dot{f}(\cdot)$ peaks at zero for an unimodal and symmetric function. Now summing (19) over the index set $\omega \in [d_1] \times \cdots \times [d_K]$ gives

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\Theta^*}) = \sum_{\omega \in [d_1] \times \cdots \times [d_K]} \mathrm{KL}(\theta_\omega || \theta^*_\omega) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta - \Theta^*\|_F^2.$$

□

Remark 1. In particular, let $\mathbb{P}_{\mathbf{0}}$ denote the distribution of $\mathcal{Y}|\mathbf{0}$ induced by the zero parameter tensor. Then we have

$$\mathrm{KL}(\mathbb{P}_\Theta || \mathbb{P}_{\mathbf{0}}) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta\|_F^2.$$

Lemma 9. *Assume the same setup as in Theorem 4.2. Without loss of generality, suppose $d_1 = \max_k d_k$. Define $R = \max_k r_k$ and $d_{total} = \prod_{k \in [K]} d_k$. For any constant $0 \leq \gamma \leq 1$, there exist a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{P}$ satisfying the following four properties:*

(i) $Card(\mathcal{X}) \geq 2^{Rd_1/8} + 1$, where Card denotes the cardinality;

(ii) $\mathcal{X}$ *contains the zero tensor* $\mathbf{0} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$*;*

(iii) $\|\Theta\|_\infty \leq \gamma \min\left\{\alpha,\ C^{-1/2}\sqrt{\frac{Rd_1}{d_{total}}}\right\}$ *for any element* $\Theta \in \mathcal{X}$*;*

(iv) $\|\Theta_i - \Theta_j\|_F \geq \frac{\gamma}{4}\min\left\{\alpha\sqrt{d_{total}},\ C^{-1/2}\sqrt{Rd_1}\right\}$ *for any two distinct elements* $\Theta_i,\ \Theta_j \in \mathcal{X}$*,*

Here $C = C(\alpha, L, f, \boldsymbol{b}) = \frac{(4L-6)\dot{f}^2(0)}{A_\alpha} > 0$ *is a constant independent of the tensor dimension and rank.*

Proof. Given a constant $0 \leq \gamma \leq 1$, we define a set of matrices:

$$\mathcal{C} = \left\{\boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{d_1 \times R} : a_{ij} \in \left\{0, \gamma \min\left\{\alpha, C^{-1/2}\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}\right\},\ \forall (i,j) \in [d_1] \times [R]\right\}.$$

We then consider the associated set of block tensors:

$$\begin{aligned}\mathcal{B} = \mathcal{B}(\mathcal{C}) = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} \colon \Theta = \boldsymbol{A} \otimes \mathbf{1}_{d_3} \otimes \cdots \otimes \mathbf{1}_{d_K},\\ \text{where } \boldsymbol{A} = (\boldsymbol{M}|\cdots|\boldsymbol{M}|\boldsymbol{O}) \in \mathbb{R}^{d_1 \times d_2},\ \boldsymbol{M} \in \mathcal{C}\},\end{aligned}$$

where $\mathbf{1}_d$ denotes a length-d vector with all entries 1, $\boldsymbol{O}$ denotes the $d_1 \times (d_2 - R\lfloor d_2/R \rfloor)$ zero matrix, and $\lfloor d_2/R \rfloor$ is the integer part of d_2/R. In other words, the subtensor $\Theta(\boldsymbol{I}, \boldsymbol{I}, i_3, \ldots, i_K) \in \mathbb{R}^{d_1 \times d_2}$ are the same for all fixed $(i_3, \ldots, i_K) \in [d_3] \times \cdots \times [d_K]$, and furthermore, each subtensor $\Theta(\boldsymbol{I}, \boldsymbol{I}, i_3, \ldots, i_K)$ itself is filled by copying the matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times R}$ as many times as would fit.

By construction, any element of $\mathcal{B}$, as well as the difference of any two elements of $\mathcal{B}$, has Tucker rank at most $\max_k r_k \leq R$, and the entries of any tensor in $\mathcal{B}$ take values in $[0, \alpha]$. Thus, $\mathcal{B} \subset \mathcal{P}$. By Lemma 10, there exists a subset $\mathcal{X} \subset \mathcal{B}$ with cardinality $\text{Card}(\mathcal{X}) \geq 2^{Rd_1/8} + 1$ containing the zero $d_1 \times \cdots \times d_K$ tensor, such that, for any two distinct elements Θ_i and Θ_j in $\mathcal{X}$,

$$\|\Theta_i - \Theta_j\|_F^2 \geq \frac{Rd_1}{8}\gamma^2 \min\left\{\alpha^2, \frac{C^{-1}Rd_1}{d_{\text{total}}}\right\} \lfloor \frac{d_2}{R} \rfloor \prod_{k \geq 3} d_k \geq \frac{\gamma^2 \min\left\{\alpha^2 d_{\text{total}}, C^{-1}Rd_1\right\}}{16}.$$

In addition, each entry of $\Theta \in \mathcal{X}$ is bounded by $\gamma \min\left\{\alpha, C^{-1/2}\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}$. Therefore the Properties (i) to (iv) are satisfied. ☐

Lemma 10 (Varshamov-Gilbert bound)**.** *Let* $\Omega = \{(w_1, \ldots, w_m) \colon w_i \in \{0,1\}\}$*. Suppose* $m > 8$*. Then there exists a subset* $\{w^{(0)}, \ldots, w^{(M)}\}$ *of* Ω *such that* $w^{(0)} = (0, \ldots, 0)$ *and*

$$\|w^{(j)} - w^{(k)}\|_0 \geq \frac{m}{8}, \quad \text{for } 0 \leq j < k \leq M,$$

where $\|\cdot\|_0$ *denotes the Hamming distance, and* $M \geq 2^{m/8}$*.*

Lemma 11 (Theorem 2.5 in Tsybakov [2008])**.** *Assume that a set* $\mathcal{X}$ *contains element* $\Theta_0, \Theta_1, \ldots, \Theta_M$ *(*$M \geq 2$*) such that*

- $d(\Theta_j,\ \Theta_k) \geq 2s > 0$*,* $\forall 0 \leq j \leq k \leq M$*;*
- $\mathbb{P}_0$ *is absolutely continuous with respect to* $\mathbb{P}_j$*,* $\forall j = 1, \ldots, M$*, and*

$$\frac{1}{M}\sum_{j=1}^{M} \text{KL}(\mathbb{P}_j || \mathbb{P}_0) \leq \alpha \log M$$

where $d \colon \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty]$ *is a semi-distance function,* $0 < \alpha < 1/8$ *and* $\mathbb{P}_j = \mathbb{P}_{\Theta_j}$*,* $j = 0, 1 \ldots, M$*.*

Then

$$\inf_{\hat{\Theta}} \sup_{\Theta \in \mathcal{X}} \mathbb{P}_{\Theta}(d(\hat{\Theta}, \Theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0.$$

Lemma 12 (Lemma 28 in Ghadermarzy et al. [2019]). *Define* $\mathbb{B}_M(\beta) = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \|\Theta\|_M \leq \beta\}$*. Let* $\Omega \subset [d_1] \times \cdots \times [d_K]$ *be a random set with* $m = |\Omega|$*, and assume that each entry in* Ω *is drawn with replacement from* $[d_1] \times \cdots \times [d_K]$ *using probability* Π*. Define*

$$\|\Theta\|_{F,\Pi}^2 = \frac{1}{m} \mathbb{E}_{\Omega \in \Pi} \|\Theta\|_{F,\Omega}^2.$$

Then, there exists a universal constant $c > 0$*, such that, with probability at least* $1 - \exp\left(-\frac{\sum_k d_k}{\sum_k \log d_k}\right)$ *over the sampled set* Ω*,*

$$\frac{1}{m} \|\Theta\|_{F,\Omega}^2 \geq \|\Theta\|_{F,\Pi}^2 - c\beta \sqrt{\frac{\sum_k d_k}{m}}$$

holds uniformly for all $\Theta \in \mathbb{B}_M(\beta)$*.*

2 Extension of Theorem 4.1 to unknown cut-off points

When the cut-off points $\boldsymbol{b}$ is unknown, we estimate $(\hat{\Theta}, \hat{\boldsymbol{b}})$ by

$$(\hat{\Theta}, \hat{\boldsymbol{b}}) = \underset{(\Theta, \boldsymbol{b}) \in \mathcal{P} \times \mathcal{B}}{\arg\max}\ \mathcal{L}_{\mathcal{Y},\Omega}(\Theta, \boldsymbol{b}), \tag{20}$$

where

$$\mathcal{P} = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \text{rank}(\mathcal{P}) \leq \boldsymbol{r},\ \|\Theta\|_\infty \leq \alpha\}, \quad \mathcal{B} = \{\boldsymbol{b} \in \mathbb{R}^{L-1} : \|\boldsymbol{b}\|_\infty \leq \beta,\ \min_\ell (b_\ell - b_{\ell-1}) \geq \Delta\}.$$

The estimation accuracy is assessed using the mean squared error (MSE):

$$\text{MSE}\left(\hat{\Theta}, \Theta^{\text{true}}\right) = \frac{1}{\prod_k d_k} \|\hat{\Theta} - \Theta^{\text{true}}\|_F, \quad \text{MSE}\left(\hat{\boldsymbol{b}}, \boldsymbol{b}^{\text{true}}\right) = \frac{1}{L-1} \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F.$$

By abuse of notation, we use $(\Theta, \boldsymbol{b})$ to denote the length-$(d_{\text{total}} + L - 1)$-vector collecting all parameters together. The total MSE is defined as

$$\text{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}),\ (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\right) = \frac{1}{\prod_k d_k + L - 1} \|(\hat{\Theta}, \hat{\boldsymbol{b}}) - (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\|_F^2.$$

We introduce several quantities that will be used in our theory:

1. We make the convention that $b_0 = -\infty$, $b_L = \infty$, $f(-\infty) = 0$, $f(\infty) = 1$, and $\dot{f}(-\infty) = \ddot{f}(-\infty) = \dot{f}(\infty) = \ddot{f}(\infty) = 0$.

2. The difference function $g_\ell(\theta)$ is defined as $g_\ell(\theta) = f(b_\ell - \theta) - f(b_{\ell-1} - \theta)$ for all $\theta \in \mathbb{R}$ and $\ell \in [L]$.

3. Define $n_\ell = \sum_{\omega \in \Omega} \mathbb{1}_{\{y_\omega = \ell\}}$, i.e., the number of tensor entries taking value on $\ell \in [L]$.

4. With a little abuse of notation, we re-define the constants in (2) as

$$U_{\alpha,\beta,\Delta} = \max_{\substack{|\theta| \leq \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \max_{\ell \in [L-1]} \frac{|\dot{g}_\ell(\theta)|}{g_\ell(\theta)}, \quad \text{and} \quad L_{\alpha,\beta,\Delta} = \min_{\substack{|\theta| \leq \alpha, \\ \boldsymbol{b} \in \mathcal{B}}} \min_{\ell \in [L-1]} \frac{\dot{g}_\ell^2(\theta) - \ddot{g}_\ell(\theta) g_\ell(\theta)}{g_\ell^2(\theta)}. \tag{21}$$

5. We define three additional constants:

$$C_{\alpha,\beta,\Delta} = \max_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \max_{\ell\in[L-1]} \max\left\{ \frac{\dot{f}(b_\ell-\theta)}{g_\ell(\theta)},\ \frac{\dot{f}(b_\ell-\theta)}{g_{\ell+1}(\theta)} \right\},$$

$$\begin{aligned} D_{\alpha,\beta,\Delta} &= \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \min\left\{ \frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_\ell-\theta)}{g_\ell(\theta)}\right),\ -\frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_\ell-\theta)}{g_{\ell+1}(\theta)}\right) \right\} \\ &= \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \min\Bigg\{ -\frac{\ddot{f}(b_\ell-\theta)g_\ell(\theta)-\dot{f}(b_\ell-\theta)\dot{g}_\ell(\theta)}{g_\ell^2(\theta)}, \\ &\qquad\qquad \frac{\ddot{f}(b_\ell-\theta)g_{\ell+1}(\theta)-\dot{f}(b_\ell-\theta)\dot{g}_{\ell+1}(\theta)}{g_{\ell+1}^2(\theta)} \Bigg\}, \end{aligned}$$

$$\tilde{D}_{\alpha,\beta,\Delta} = \max_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \max_{\ell\in[L-1]} \max\left\{ \frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_\ell-\theta)}{g_\ell(\theta)}\right),\ -\frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_\ell-\theta)}{g_{\ell+1}(\theta)}\right) \right\}. \tag{22}$$

We make the following assumptions about the link function.

Assumption 1. *The link function $f\colon \mathbb{R}\mapsto[0,1]$ satisfies the following properties:*

1. *$f(\theta)$ is twice-differentiable and strictly increasing in θ.*
2. *$\dot{f}(\theta)$ is strictly log-concave and symmetric with respect to $\theta=0$.*
3. *The function $\frac{\dot{f}(b_\ell-\theta)}{g_\ell(\theta)}$ is strictly increasing with respect to θ for all $\boldsymbol{b}\in\mathcal{B}$.*
4. *The function $\frac{\dot{f}(b_\ell-\theta)}{g_{\ell+1}(\theta)}$ is strictly decreasing with respect to θ for all $\boldsymbol{b}\in\mathcal{B}$.*

Remark 2. The condition $\Delta=\min_\ell(b_\ell-b_{\ell-1})>0$ on the feasible set $\mathcal{B}$ guarantees the strict positiveness of $g_\ell(\theta)=f(b_\ell-\theta)-f(b_{\ell-1}-\theta)$. Therefore, the denominators in the above quantities $U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta}, C_{\alpha,\beta,\Delta}, D_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}$ are well-defined. Furthermore, by Theorem 1.1, $g_\ell(\theta)$ is strictly log-concave in θ for all $\boldsymbol{b}\in\mathcal{B}$. Based on Assumption 1 and closeness of the feasible set, we have $U_{\alpha,\beta,\Delta}>0$, $L_{\alpha,\beta,\Delta}>0$, $C_{\alpha,\beta,\Delta}>0$, $D_{\alpha,\beta,\Delta}>0$, $\tilde{D}_{\alpha,\beta,\Delta}>0$.

Remark 3. In particular, for logistic link $f(x)=\frac{1}{1+e^{-x}}$, we have

$$C_{\alpha,\beta,\Delta} = \max_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \max_{\ell\in[L-1]} \max\left\{ \frac{1}{e^{b_\ell-b_{\ell-1}}-1}\frac{1+e^{-(b_{\ell-1}-\theta)}}{1+e^{-(b_\ell-\theta)}},\ \frac{1}{1-e^{-(b_{\ell+1}-b_\ell)}}\frac{1+e^{-(b_{\ell+1}-\theta)}}{1+e^{-(b_\ell-\theta)}} \right\} > 0,$$

$$D_{\alpha,\beta,\Delta} = \min_{\substack{|\theta|\le\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \frac{e^{-(b_\ell-\theta)}}{(1+e^{-(b_\ell-\theta)})^2} > 0.$$

Theorem 2.1 (Statistical convergence with unknown $\boldsymbol{b}$)**.** *Consider an ordinal tensor $\mathcal{Y}\in[L]^{d_1\times\cdots\times d_K}$ generated from model (1) with the link function f and parameters $(\Theta^{true},\boldsymbol{b}^{true})\in\mathcal{P}\times\mathcal{B}$. Suppose the link function f satisfies Assumption 1. Define $r_{\max}=\max_k r_k$, and assume $r_{\max}=\mathcal{O}(1)$. Then with very high probability, the estimator in (20) satisfies*

$$\mathrm{MSE}\left(\hat{\Theta},\Theta^{true}\right) \le \min\left(4\alpha^2,\ \frac{c_1 U_{\alpha,\beta,\Delta}^2 r_{\max}^{K-1}}{L_{\alpha,\beta,\Delta}^2}\frac{\sum_k d_k}{\prod_k d_k}\right), \tag{23}$$

and

$$\text{MSE}\left(\hat{\boldsymbol{b}}, \boldsymbol{b}^{true}\right) \le \min\left(4\beta^2,\ \frac{C'_{\alpha,\beta,\Delta} r_{\max}^{K-1}}{D^2_{\alpha,\beta,\Delta}} \frac{\left(\sum_k d_k\right)\left(\prod_k d_k\right)}{\min_\ell (n_\ell + n_{\ell+1})^2}\right), \tag{24}$$

where $c_1, U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta}, C'_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}$ *are constants independent of the tensor dimension and rank.*

In particular, we have

$$\text{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}),\ (\Theta^{true}, \boldsymbol{b}^{true})\right) = \mathcal{O}\left(\frac{\sum_k d_k}{\min\left\{\prod_k d_k,\ \min_\ell (n_\ell + n_{\ell+1})^2\right\}}\right), \quad \text{as} \quad d_{\min} \to \infty.$$

Remark 4. The total MSE has two components, where the first component $\mathcal{O}\left(\frac{\sum_k d_k}{\prod_k d_k}\right)$ is from the error in $\hat{\Theta}$, and the second component $\mathcal{O}\left(\frac{\sum_k d_k}{\min_\ell (n_\ell + n_{\ell+1})^2}\right)$ is from the error in $\hat{\boldsymbol{b}}$. When the L labels are moderately balanced in that $\min_\ell (n_\ell + n_{\ell+1}) \gg \sqrt{\prod_k d_k}$, the error in $\hat{\Theta}$ dominates the total MSE. When the L labels are severely imbalanced in that $\min_\ell (n_\ell + n_{\ell+1}) \ll \sqrt{\prod_k d_k}$, the error in $\hat{\boldsymbol{b}}$ dominates the total MSE.

Remark 5. We discuss the impact of imbalance between labels to the estimation accuracy. Consider the special case $d_1 = \cdots = d_K = d$. The bounds (23) and (24) demonstrate that, both $\hat{\Theta}$ and $\hat{\boldsymbol{b}}$ achieve consistency as long as $\min_\ell (n_\ell + n_{\ell+1})$ is slightly larger than $\mathcal{O}(d^{(K+1)/2})$. Note that in this case $\frac{\min_\ell (n_\ell + n_{\ell+1})}{\max_\ell n_\ell} \asymp \frac{\min_\ell (n_\ell + n_{\ell+1})}{d^K} \to 0$, indicating a long-run imbalance between labels. The obtained consistency highlights the robustness of our proposed estimator (20) to the label imbalance.

Proof of Theorem 2.1. The constant bounds with α and β can be obtained trivially from the definition of the feasible sets. We focus on proving the non-constant bounds.

Similar to the proof of Theorem 4.1, we suppress Ω in the subscript. Based on the definition of $(\hat{\Theta}, \hat{\boldsymbol{b}})$, we have the following inequalities:

$$\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \hat{\boldsymbol{b}}) \ge \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}, \hat{\boldsymbol{b}}) \quad \text{and} \quad \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \hat{\boldsymbol{b}}) \ge \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b}^{\text{true}}). \tag{25}$$

Following the same argument as in Theorem 4.1 and the first inequality in (25), we obtain that

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \le \frac{c_1^2 U^2_{\alpha,\beta,\Delta} r_{\max}^{K-1}}{L^2_{\alpha,\beta,\Delta}} \sum_k d_k, \tag{26}$$

where $U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta} > 0$ are two constants defined in (21).

Next we bound $\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2$ given $\hat{\Theta}$. We consider the profile log-likelihood $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b})$ as a function of $\boldsymbol{b} \in \mathcal{B}$. For notational convenience, we drop $\hat{\Theta}$ from $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}, \boldsymbol{b})$ and simply write $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b})$. It follows from the expression of $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b})$ that

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell} &= \sum_{\omega \in \Omega} \left[\mathbb{1}_{\{y_\omega = \ell\}} \frac{\dot{f}(b_\ell - \theta_\omega)}{g_\ell(\theta_\omega)} - \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\dot{f}(b_\ell - \theta_\omega)}{g_{\ell+1}(\theta_\omega)}\right], \\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2} &= \sum_{\omega \in \Omega} \left[\mathbb{1}_{\{y_\omega = \ell\}} \frac{\ddot{f}(b_\ell - \theta_\omega) g_\ell(\theta_\omega) - \dot{f}^2(b_\ell - \theta_\omega)}{g_\ell^2(\theta_\omega)} - \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\ddot{f}(b_\ell - \theta_\omega) g_{\ell+1}(\theta_\omega) + \dot{f}^2(b_\ell - \theta_\omega)}{g_{\ell+1}^2(\theta_\omega)}\right], \\
&\quad \text{for all } \ell \in [L-1], \\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell+1}} &= \sum_{\omega \in \Omega} \mathbb{1}_{\{y_\omega = \ell+1\}} \frac{\dot{f}(b_\ell - \theta_\omega) \dot{f}(b_{\ell+1} - \theta_\omega)}{g_{\ell+1}^2(\theta_\omega)} \quad \text{and} \quad \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b'_\ell} = 0 \text{ if } |\ell - \ell'| > 1.
\end{aligned}$$

Therefore, all entries in $\nabla_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}$ are upper bounded by $\{C_{\alpha,\beta,\Delta}\max_\ell(n_\ell+n_{\ell+1})\}>0$, and $\nabla^2_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}$ is a tridiagonal matrix.

By the second-order Taylor's expansion of $\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b})$ around $\boldsymbol{b}^{\text{true}}$, we obtain

$$\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) = \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + (\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}})^T\nabla_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + \frac{1}{2}(\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}})^T\nabla^2_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}})(\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}}), \tag{27}$$

where $\check{\boldsymbol{b}} = \gamma\boldsymbol{b}^{\text{true}} + (1-\gamma)\hat{\boldsymbol{b}}$ for some $\gamma\in[0,1]$, and $\nabla^2_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}})$ denotes the $(L-1)$-by-$(L-1)$ Hession matrix evaluated at $\check{\boldsymbol{b}}$.

The linear term in (27) can be bounded by Cauchy-Schwartz inequality,

$$\begin{aligned}(\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}})^T\nabla_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) &\le \|\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}}\|_F\|\nabla_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}})\|_F \\ &\le C'_{\alpha,\beta,\Delta}\|\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}}\|_F\sqrt{L-1}\sqrt{\left(\sum_k d_k\right)\left(\prod_k d_k\right)},\end{aligned} \tag{28}$$

for some constant $C'_{\alpha,\beta,\Delta}>0$, where the last inequality follows from Lemma 14,

$$\left|\frac{\partial\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\Big|_{(\hat{\Theta},\boldsymbol{b}^{\text{true}})}\right| \le C'_{\alpha,\beta,\Delta}\sqrt{\left(\sum_k d_k\right)\left(\prod_k d_k\right)}, \quad \text{for all } \ell\in[L-1] \text{ with very high probability.}$$

We next bound the quadratic term in (27). Note that

$$\begin{aligned}&(\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}})^T\nabla^2_{\boldsymbol{b}}\mathcal{L}_{\mathcal{Y}}(\check{\boldsymbol{b}})(\boldsymbol{b}^{\text{true}}-\hat{\boldsymbol{b}}) \\ &= \sum_{\ell\in[L-1]}\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell-b_\ell^{\text{true}})^2 + 2\sum_{\ell\in[L-1]/\{1\}}\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell\partial b_{\ell-1}}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell-b_\ell^{\text{true}})(\hat{b}_{\ell-1}-b_{\ell-1}^{\text{true}}) \\ &\le \sum_{\ell\in[L-1]}\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)(\hat{b}_\ell-b_\ell^{\text{true}})^2 + \sum_{\ell\in[L-1]/\{1\}}\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell\partial b_{\ell-1}}\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}\right)\left[(\hat{b}_\ell-b_\ell^{\text{true}})^2+(\hat{b}_{\ell-1}-b_{\ell-1}^{\text{true}})^2\right] \\ &= \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_1^2}+\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_1\partial b_2}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_1-b_1^{\text{true}})^2 + \left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{L-1}^2}+\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{L-2}\partial b_{L-1}}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_{L-1}-b_{L-1}^{\text{true}})^2 \\ &\quad + \sum_{\ell\in[L-2]/\{1\}}\left(\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2}+\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell\partial b_{\ell-1}}+\frac{\partial^2\mathcal{L}_{\mathcal{Y}}}{\partial b_{\ell+1}\partial b_\ell}\right)\Big|_{\boldsymbol{b}=\check{\boldsymbol{b}}}(\hat{b}_\ell-b_\ell^{\text{true}})^2 \\ &\le -D'_{\alpha,\beta,\Delta}\sum_{\ell\in[L-1]}(\hat{b}_\ell-b_{\text{true},\ell})^2 \\ &= -D'_{\alpha,\beta,\Delta}\|\hat{\boldsymbol{b}}-\boldsymbol{b}^{\text{true}}\|_F^2,\end{aligned} \tag{29}$$

where

$$
\begin{aligned}
D'_{\alpha,\beta,\Delta} &= \min_{\substack{|\theta|\leq\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} -\left(\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell^2} + \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell \partial b_{\ell-1}} + \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial b_{\ell+1} \partial b_\ell}\right) \\
&= \min_{\substack{|\theta|\leq\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \left\{ \sum_{\omega\in\Omega} -\mathbb{1}_{\{y_\omega=\ell\}} \left(\frac{\ddot{f}(b_\ell-\theta_\omega)g_\ell(\theta_\omega) - \dot{f}(b_\ell-\theta_\omega)\dot{g}_\ell(\theta_\omega)}{g_\ell^2(\theta_\omega)} \right) \right. \\
&\qquad \left. + \sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell+1\}} \left(\frac{\ddot{f}(b_\ell-\theta_\omega)g_{\ell+1}(\theta_\omega) - \dot{f}(b_\ell-\theta_\omega)\dot{g}_{\ell+1}(\theta_\omega)}{g_{\ell+1}^2(\theta_\omega)} \right) \right\} \\
&= \min_{\substack{|\theta|\leq\alpha,\\ \boldsymbol{b}\in\mathcal{B}}} \min_{\ell\in[L-1]} \left\{ \underbrace{-\sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell\}} \frac{\partial}{\partial\theta_\omega} \left(\frac{\dot{f}(b_\ell-\theta_\omega)}{g_\ell(\theta_\omega)} \right)}_{>0} + \underbrace{\sum_{\omega\in\Omega} \mathbb{1}_{\{y_\omega=\ell+1\}} \frac{\partial}{\partial\theta_\omega} \left(\frac{\dot{f}(b_\ell-\theta_\omega)}{g_{\ell+1}(\theta_\omega)} \right)}_{>0} \right\} \\
&\geq D_{\alpha,\beta,\Delta} \min_{\ell\in[L-1]} (n_\ell + n_{\ell+1}).
\end{aligned}
$$

Combining inequalities (27), (28) and (29) yields

$$
\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) \leq \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) + C'_{\alpha,\beta,\Delta} \sqrt{(L-1)\left(\sum_k d_k\right)\left(\prod_k d_k\right)} \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F - \frac{D_{\alpha,\beta,\Delta}}{2} \min_\ell (n_\ell + n_{\ell+1}) \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2.
$$

Since $\hat{\boldsymbol{b}}$ satisfies $\mathcal{L}_{\mathcal{Y}}(\hat{\boldsymbol{b}}) - \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) \geq 0$, we have that

$$
C'_{\alpha,\beta,\Delta} \sqrt{(L-1)\left(\sum_k d_k\right)\left(\prod_k d_k\right)} \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F - \frac{D_{\alpha,\beta,\Delta}}{2} \min_\ell (n_\ell + n_{\ell+1}) \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2 \geq 0.
$$

Finally,

$$
\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2 \leq \frac{C'_{\alpha,\beta,\Delta}(L-1)}{D^2_{\alpha,\beta,\Delta}} \frac{\left(\sum_k d_k\right)\left(\prod_k d_k\right)}{\min_\ell (n_\ell + n_{\ell+1})^2}. \tag{30}
$$

Combing (26) and (30) with the following inequality completes the proof.

$$
\begin{aligned}
\text{MSE}\left((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\right) &\leq \frac{1}{\prod_k d_k} \|(\hat{\Theta}, \hat{\boldsymbol{b}}) - (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\|_F^2 \\
&= \frac{1}{\prod_k d_k} \|(\hat{\Theta}, \hat{\boldsymbol{b}}) - (\hat{\Theta}, \boldsymbol{b}^{\text{true}}) + (\hat{\Theta}, \boldsymbol{b}^{\text{true}}) - (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})\|_F^2 \\
&\leq \frac{1}{\prod_k d_k} \|(\hat{\Theta} - \Theta^{\text{true}})\|_F^2 + \frac{1}{\prod_k d_k} \|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F^2.
\end{aligned}
$$

□

Lemma 13 (CLT for independent Bernoulli r.v.'s)**.** *Let $\{X_n\}$ be a series of independent Bernoulli random variables with possibly different success probabilities $\{p_n\}$. Suppose the success probabilities $\{p_n\}$ are bounded, in that, $c_1 \leq p_n \leq 1 - c_2$ for some positive constants $c_1, c_2 > 0$ and all $n \in [m]$. Define $s_m^2 = \sum_{n=1}^m p_n(1-p_n)$ and $Y_n = X_n - p_n$. Then,*

$$
\frac{1}{s_m} \sum_{n=1}^m Y_n \xrightarrow{\mathcal{D}} N(0,1), \qquad \text{as } m \to \infty.
$$

Proof. We apply Lyapunov central limit theorem (CLT) to $\{Y_n\}$. Let us verify Lyapunov's condition for $\delta = 1$:

$$\mathbb{E}[|Y_n|^3] = p_n(1-p_n)^3 + (1-p_n)p_n^3 \le p_n(1-p_n)[(1-p_n)^2 + p_n^2] \le p_n(1-p_n).$$

Summation of the above inequality shows

$$\sum_{n=1}^{m} \mathbb{E}[|Y_n|^3] \le \sum_{n=1}^{m} p_n(1-p_n) = s_m^2.$$

Thus, the Lyapunov condition is satisfied whenever

$$\frac{1}{s_m^3}\sum_{n=1}^{m} \mathbb{E}[|Y_n|^3] \le \frac{s_m^2}{s_m^3} \to 0,$$

or simply $s_m = \sum_{n=1}^{m} p_n(1-p_n) \ge m \min\{c_1(1-c_1), c_2(1-c_2)\} \to \infty$. Applying Lyapunov CLT to $\{Y_n\}$ completes the proof. □

Lemma 14 (Bound on score function)**.** *Consider the same set-up as in Theorem 2.1. Let $\mathcal{L}_{\mathcal{Y}}(\Theta, \boldsymbol{b})$ denote the log-likelihood function of $(\Theta, \boldsymbol{b})$ given data tensor $\mathcal{Y}$. Then, with very high probability,*

$$\left| \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\bigg|_{(\Theta, \boldsymbol{b}^{true})} \right| \le \left(C_{\alpha,\beta,\Delta} + \tilde{D}_{\alpha,\beta,\Delta}\|\Theta^{true} - \Theta\|_F^2 \right) \sqrt{\prod_k d_k}, \quad \text{for all } \ell \in [L-1],$$

In particular, with very high probability,

$$\left| \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\bigg|_{(\hat{\Theta}, \boldsymbol{b}^{true})} \right| \le C'_{\alpha,\beta,\Delta} \sqrt{\left(\sum_k d_k\right)\left(\prod_k d_k\right)}, \quad \text{for all } \ell \in [L-1],$$

where $C'_{\alpha,\beta,\Delta} > 0$ is a constant that depends on $U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta}, C_{\alpha,\beta,\Delta}, \tilde{D}_{\alpha,\beta,\Delta}$ defined in (21) and (22).

Proof. We only prove the case for $\ell = 1$. Other cases can be proved similarly.

$$\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\bigg|_{(\Theta, \boldsymbol{b}^{\text{true}})} = \underbrace{\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\bigg|_{(\Theta, \boldsymbol{b}^{\text{true}})} - \mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\bigg|_{(\Theta, \boldsymbol{b}^{\text{true}})}\right]}_{:=A} + \underbrace{\mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\bigg|_{(\Theta, \boldsymbol{b}^{\text{true}})}\right] - \mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\bigg|_{(\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})}\right]}_{:=B}. \tag{31}$$

We have used the fact that the score function has mean zero, $\mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\big|_{(\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})}\right] = 0$. Here all expectations are taken with respect to $\mathcal{Y} \sim \mathbb{P}(\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})$.

We now bound the two deviation terms in (31) separately. The term A in (31) is the stochastic deviation of log-likelihood to its expectation:

$$\begin{aligned} A &= \sum_{\omega \in \Omega} \left\{ [\mathbb{1}_{y_\omega = 1} - g_1(\theta_\omega^{\text{true}})] \frac{\dot{f}(b_1 - \theta_\omega)}{g_1(\theta_\omega)} - [\mathbb{1}_{y_\omega = 2} - g_2(\theta_\omega^{\text{true}})] \frac{\dot{f}(b_1 - \theta_\omega)}{g_2(\theta_\omega)} \right\} \\ &\le C_{\alpha,\beta,\Delta} \sum_{\omega \in \Omega} \underbrace{[\mathbb{1}_{y_\omega = 1} + \mathbb{1}_{y_\omega = 2} - g_1(\theta_\omega^{\text{true}}) - g_2(\theta_\omega^{\text{true}})]}_{:=W_\omega}. \end{aligned}$$

Note that $\{W_\omega\}$ are independent, centered Bernoulli random variables with bounded success probabilities $g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})$. By Lemma 13, we have

$$\sum_{\omega\in\Omega} W_\omega \xrightarrow{\mathcal{D}} N\left(0,\ \sum_{\omega\in\Omega}\left(g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right)\left(1 - g_1(\theta_\omega^{\text{true}}) - g_2(\theta_\omega^{\text{true}})\right)\right).$$

Hence, with the fact that $\sum_{\omega\in\Omega}\left(g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})\right)\left(1 - g_1(\theta_\omega^{\text{true}}) - g_2(\theta_\omega^{\text{true}})\right) \leq \frac{1}{4}\prod_k d_k$,

$$|A| \leq C_{\alpha,\beta,\Delta}\left|\sum_{\omega\in\Omega} W_\omega\right| \leq C_{\alpha,\beta,\Delta}\sqrt{\prod_k d_k}, \quad \text{with very high probability.} \tag{32}$$

The second term B in (31) is the bias induced by Θ:

$$\begin{aligned}
|B| &= \left|\sum_{\omega\in\Omega} g_1(\theta_\omega^{\text{true}})\left(\frac{\dot{f}(b_1-\theta_\omega)}{g_1(\theta_\omega)} - \frac{\dot{f}(b_1-\theta_\omega^{\text{true}})}{g_1(\theta_\omega^{\text{true}})}\right) - \sum_{\omega\in\Omega} g_2(\theta_\omega^{\text{true}})\left(\frac{\dot{f}(b_1-\theta_\omega)}{g_2(\theta_\omega)} - \frac{\dot{f}(b_1-\theta_\omega^{\text{true}})}{g_2(\theta_\omega^{\text{true}})}\right)\right| \\
&= \left|\sum_{\omega\in\Omega} g_1(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}} - \theta_\omega)\left\{\frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_1-\theta)}{g_1(\theta)}\right)\Big|_{\rho_\omega\theta_\omega + (1-\rho_\omega)\theta_\omega^{\text{true}}}\right\}\right. \\
&\qquad \left. - \sum_{\omega\in\Omega} g_2(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}} - \theta_\omega)\left\{\frac{\partial}{\partial\theta}\left(\frac{\dot{f}(b_1-\theta)}{g_2(\theta)}\right)\Big|_{\rho'_\omega\theta_\omega + (1-\rho'_\omega)\theta_\omega^{\text{true}}}\right\}\right| \\
&\leq \tilde{D}_{\alpha,\beta,\Delta}\sum_{\omega\in\Omega}|g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}})||\theta_\omega^{\text{true}} - \theta_\omega|.
\end{aligned}$$

By Cauchy-Schwartz inequality with the fact that $g_1(\theta_\omega^{\text{true}}) + g_2(\theta_\omega^{\text{true}}) \leq 1$,

$$|B| \leq \tilde{D}_{\alpha,\beta,\Delta}\sqrt{\prod_k d_k}\|\Theta^{\text{true}} - \Theta\|_F. \tag{33}$$

Plugging (32) and (33) back to (31) yields that

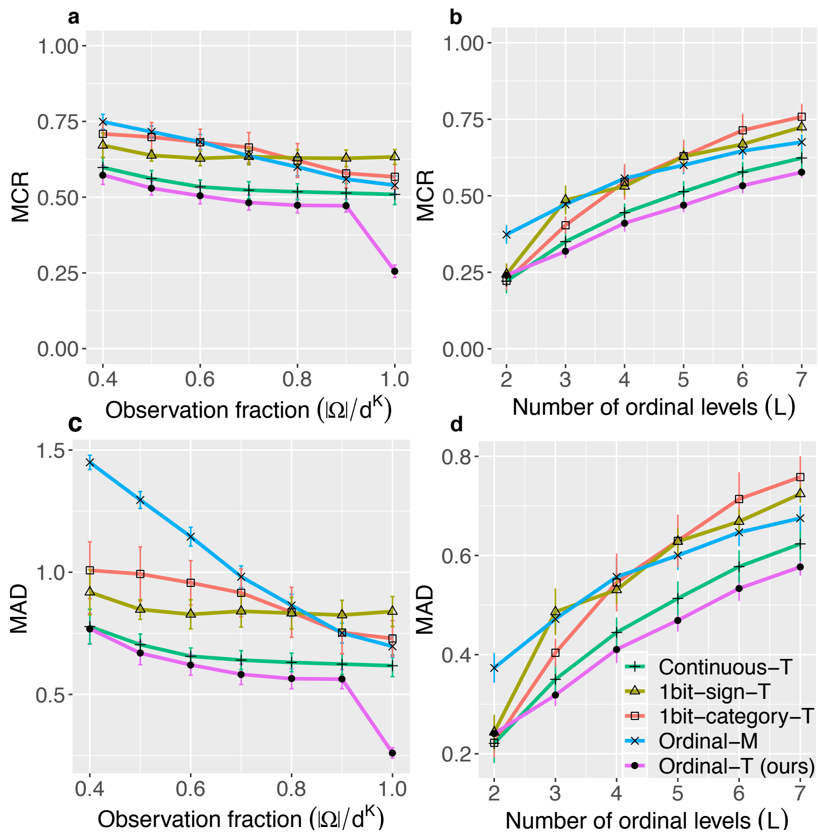
$$\frac{\partial\mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\Theta,\boldsymbol{b}^{\text{true}})} \leq \left(C_{\alpha,\beta,\Delta} + \tilde{D}_{\alpha,\beta,\Delta}\|\Theta^{\text{true}} - \Theta\|_F\right)\sqrt{\prod_k d_k}.$$

holds with very high probability. The second inequality in the lemma comes from (26) that $\|\Theta^{\text{true}} - \hat{\Theta}\|_F \leq \mathcal{O}(\sqrt{\sum_k d_k})$. □

3 Additional results on simulations

We also compare the methods by their performance in predicting the median labels, $y_\omega^{\text{median}} = \min\{\ell\colon \mathbb{P}(y_\omega = \ell) \geq 0.5\}$. Under the latent variable model (4) and Assumption 1, the median label is the quantized θ_ω without noise; i.e. $y_\omega^{\text{median}} = \sum_\ell \mathbb{1}_{\theta_\omega\in(b_{\ell-1},b_\ell]}$. We utilize the same simulation setting as in the earlier experiment. Figure S1 shows that our method outperforms the others in both MCR and MAD. The improved accuracy comes from the incorporation of multilinear low-rank structure, multi-level observations, and the ordinal structure. Interestingly, for the three multilevel methods (**1bit-sign-T**, **Ordinal-M**, and **Ordinal-T**), the median estimator tends to yield smaller MAD than the mode estimator, $\text{MAD}(\mathcal{Y}^{\text{median}}, \hat{\mathcal{Y}}^{\text{median}}) \leq \text{MAD}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}})$ (Figures 3a-b vs.

Figures S1a-b). On the other hand, the mode estimator tends to yield smaller MCR than the median estimator, $\text{MCR}(\mathcal{Y}^{\text{mode}}, \hat{\mathcal{Y}}^{\text{mode}}) \leq \text{MCR}(\mathcal{Y}^{\text{median}}, \hat{\mathcal{Y}}^{\text{median}})$ (Figures 3c-d vs. Figures S1c-d). This tendency is from the property that the median estimator $\hat{y}_\omega^{(\text{median})}$ minimizes $R_1(z) = \mathbb{E}_{y_\omega}|y_\omega - z|$, whereas the mode estimator $\hat{y}_\omega^{(\text{mode})}$ minimizes $R_2(z) = \mathbb{E}_{y_\omega}\mathbb{1}_{\{y_\omega=z\}}$. Here the expectation is over the categorical distribution of y_ω given parameters $\hat{\Theta}$ and $\hat{\boldsymbol{b}}$.



Supplementary Figure S1: Performance comparison for predicting median labels. (a, c) Prediction error versus sample complexity $\rho = |\Omega|/d^K$ when $L = 5$. (b, d) Prediction error versus the number of ordinal levels L, when $\rho = 0.8$.

4 Additional explanations of HCP analysis

We perform clustering analyses based on the Tucker representation of the estimated tensor parameter $\hat{\Theta}$. The procedure is motivated from the higher-order extension of Principal Component Analysis (PCA) or Singular Value Decomposition (SVD). Recall that, in the matrix case, we usually perform clustering on an $m \times n$ (normalized) matrix X based on the following procedure. First, we factorize X into

$$X = U\Sigma V^T,$$

where Σ is a diagonal matrix and U, V are factor matrices with orthogonal columns. Second, we take each column of V as a principal axis and each row in $U\Sigma$ as principal component. A subsequent multivariate clustering method (such as K-means) is then applied to the m rows of $U\Sigma$.

We apply a similar clustering procedure to the estimated parameter tensor $\hat{\Theta}$. Based on Tucker representation of $\hat{\Theta}$, we have

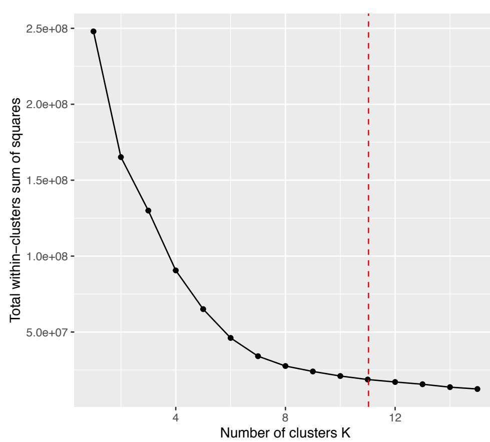
$$\hat{\Theta} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}_1 \times_2 \cdots \times_K \hat{\boldsymbol{M}}_K, \tag{34}$$

where $\hat{\mathcal{C}} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the estimated core tensor, $\hat{\boldsymbol{M}}_k \in \mathbb{R}^{d_k \times r_k}$ are estimated factor matrices with orthogonal columns, and $\times_k$ denotes the tensor-by-matrix multiplication [Kolda and Bader, 2009]. The mode-k matricization of (34) gives

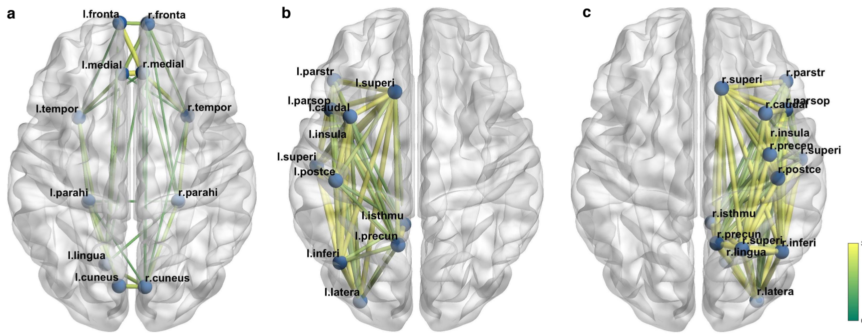
$$\hat{\Theta}_{(k)} = \hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)} \left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1 \right),$$

where $\hat{\Theta}_{(k)}, \hat{\mathcal{C}}_{(k)}$ denote the mode-k unfolding of $\hat{\Theta}$ and $\hat{\mathcal{C}}$, respectively. Then, the mode-k clustering can be performed as follows. First, we take columns in $\left(\hat{\boldsymbol{M}}_K \otimes \cdots \otimes \hat{\boldsymbol{M}}_1 \right)$ as principal axes and rows in $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$ as principal components. Then, we perform K-means clustering method to the d_k rows of the matrix $\hat{\boldsymbol{M}}_k \hat{\mathcal{C}}_{(k)}$.

We perform a clustering analysis on the 68 brain nodes using the produce described above. Our ordinal tensor method outputs the estimated parameter tensor $\hat{\Theta} \in \mathbb{R}^{68 \times 68 \times 136}$ with rank $(23, 23, 8)$. We apply K-means to the mode-1 principal component matrix of size 68×184 ($184 = 23 \times 8$). The elbow method suggests 11 clusters among the 68 nodes (see Figure S2). We find that the clustering captures the spatial separation between brain regions very well (Table S1). In particular, cluster I represents the connection between the left and right hemispheres, whereas clusters II-III represent the connection within each of the half brains (Figure S3). Other smaller clusters represent local regions driving by similar nodes (Table S1). For example, the cluster IV/VII consists of nodes in the supramarginal gyrus region in the left/right hemisphere. This region is known to be involved in visual word recognition and reading [Stoeckel et al., 2009]. The identified similarities among nodes without external annotations illustrate the applicability of our method to clustering analysis.



Supplementary Figure S2: Elbow plot for determining the number of clusters in K-means.



Supplementary Figure S3: Top three clusters in the HCP analysis. (a) Cluster I reflects the connections between two brain hemispheres. (b)-(c) Cluster II/III consists of nodes within left/right hemisphere only. Node name are shown in abbreviation. Edges are colored based on predicted connection level averaged across individuals.

Cluster	I		
Brain nodes	l.frontalpole, l.temporalpole, l.medialorbitofrontal, l.cuneus, l.parahippocampal, l.lingual, r.frontalpole, r.temporalpole, r.medialorbitofrontal, r.cuneus, r.parahippocampal		
Cluster	II		
Brain nodes	l.caudalmiddlefrontal, l.inferiorparietal, l.insula, l.isthmuscingulate, l.lateraloccipital(2), l.parsopercularis, l.parstriangularis, l.postcentral, l.precuneus, l.superiorfrontal, l.superiortemporal(3)		
Cluster	III		
Brain nodes	r.caudalmiddlefrontal, r.inferiorparietal, r.insula, r.isthmuscingulate, r.lateraloccipital(2), r.lingual, r.parsopercularis, r.parstriangularis, r.postcentral, r.precentral, r.precuneus, r.superiorfrontal(3), r.superiorparietal, r.superiortemporal(3)		
Cluster	IV	V	VI
Brain nodes	l.supramarginal(4)	l.inferiortemporal(3)	l.middletemporal(3)
Cluster	VII	VIII	VIIII
Brain nodes	r.supramarginal(4)	r.inferiortemporal(3)	r.middletemporal(3)
Cluster	X	XI	
Brain nodes	l.superiorfrontal(2)	l.precentral, l.superiorparietal	

Supplementary Table S1: Node clusters in the HCP analysis. The first alphabet in the node name indicates the left (L) or right (R) hemisphere. The number in the parentheses indicates the node count in each cluster.

References

Herm Jan Brascamp and Elliott H Lieb. On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. In *Inequalities*, pages 441–464. Springer, 2002.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.

Navid Ghadermarzy, Yaniv Plan, and Özgür Yilmaz. Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619, 2019.

Bo Jiang, Fan Yang, and Shuzhong Zhang. Tensor and its Tucker core: the invariance relationships. *Numerical Linear Algebra with Applications*, 24(3):e2086, 2017.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51 (3):455–500, 2009.

Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.*, pages 129–132. IEEE, 2005.

Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229, 2015.

Cornelia Stoeckel, Patricia M Gough, Kate E Watkins, and Joseph T Devlin. Supramarginal gyrus involvement in visual word recognition. *Cortex*, 45(9):1091–1096, 2009.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520: 44–66, 2017.