

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

1 Proofs

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 $\mathcal{L}_Y(\Theta)$ that

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

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

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

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

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

$$\mathcal{L}_Y(\Theta) = \mathcal{L}_Y(\Theta^{\text{true}}) + \langle \nabla_\Theta \mathcal{L}_Y, \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle + \frac{1}{2} \text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla_\Theta^2 \mathcal{L}_Y(\check{\Theta}) \text{Vec}(\Theta - \Theta^{\text{true}}), \quad (2)$$

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

We first bound the linear term in (2). Note that, by Lemma 4,

$$|\langle \nabla_\Theta \mathcal{L}_Y(\Theta^{\text{true}}), \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle| \leq \|\nabla_\Theta \mathcal{L}_Y(\Theta^{\text{true}})\|_\sigma \|\Theta - \Theta^{\text{true}}\|_*, \quad (3)$$

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

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

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

$$\mathbb{E}(s_\omega) = 0, \quad |s_\omega| \leq U_\alpha, \quad \text{for all } \omega \in [d_1] \times \cdots \times [d_K]. \quad (4)$$

By lemma 2, with probability at least $1 - \exp(-C_1 \sum_k d_k)$, we have

$$\|\nabla_\Theta \mathcal{L}_Y(\Theta^{\text{true}})\|_\sigma \leq C_2 U_\alpha \sqrt{\sum_k d_k}, \quad (5)$$

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

$$|\langle \nabla_{\Theta} \mathcal{L}_{\mathcal{Y}}(\Theta^{\text{true}}), \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle| \leq C_2 U_{\alpha} \sqrt{r_{\max}^{K-1} \sum_k d_k} \|\Theta - \Theta^{\text{true}}\|_F. \quad (6)$$

We next bound the quadratic term in (2). Note that

$$\begin{aligned} \text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla_{\Theta}^2 \mathcal{L}_{\mathcal{Y}}(\check{\Theta}) \text{Vec}(\Theta - \Theta^{\text{true}}) &= \sum_{\omega} \left(\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_{\omega}^2} \Big|_{\Theta=\check{\Theta}} \right) (\theta_{\omega} - \theta_{\text{true},\omega})^2 \\ &\leq -L_{\alpha} \sum_{\omega} (\Theta_{\omega} - \Theta_{\text{true},\omega})^2 \\ &= -L_{\alpha} \|\Theta - \Theta^{\text{true}}\|_F^2, \end{aligned} \quad (7)$$

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

Combining (2), (6) and (7), we have that, for all $\Theta \in \mathcal{P}$, with probability at least $1 - \exp(-C_1 \sum_k d_k)$,

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

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

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

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

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

Henceforth,

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

This completes the proof. \square

Proof of Corollary 1. The result follows immediately from Theorem 4.1 and Lemma 6. \square

1.2 Sample complexity for tensor completion

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 $\mathcal{X} = \{\Theta_i : i = 1, \dots\} \subset \mathcal{P}$ satisfying the properties of (i)-(iv) in Lemma 7. By Lemma 7, such a subset of tensors exist. For any tensor $\Theta \in \mathcal{X}$, let \mathbb{P}_{Θ} 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 6, we have

$$KL(\mathbb{P}_\Theta || \mathbb{P}_0) \leq C \|\Theta\|_F^2, \quad (8)$$

where $C = \frac{(4L-6)f^2(0)}{A_\alpha} > 0$ is a constant independent of the tensor dimension and rank. Combining the inequality (8) with property (iii) of \mathcal{X} , we have

$$KL(\mathbb{P}_\Theta || \mathbb{P}_0) \leq \gamma^2 R_{\max} d_{\max}. \quad (9)$$

From (9) and the property (i), we deduce that the condition

$$\frac{1}{\text{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} KL(\mathbb{P}_\Theta, \mathbb{P}_0) \leq \varepsilon \log \{\text{Card}(\mathcal{X}) - 1\} \quad (10)$$

holds for any $\varepsilon \geq 0$ when $\gamma \in [0, 1]$ is chosen to be sufficiently small depending on ε , e.g., $\gamma \leq \sqrt{3\varepsilon}$. By applying Lemma ?? to (10), and in view of the property (iv), we obtain that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\text{true}} \in \mathcal{X}} \mathbb{P} \left(\|\hat{\Theta} - \Theta^{\text{true}}\|_F \geq \frac{\gamma}{8} \min \left\{ \alpha \sqrt{d_{\text{total}}}, C^{-1/2} \sqrt{R_{\max} d_{\max}} \right\} \right) \geq \frac{1}{2} \left(1 - 2\varepsilon - \sqrt{\frac{16\varepsilon}{R_{\max} d_{\max}}} \right). \quad (11)$$

Note that $\text{Loss}(\hat{\Theta}, \Theta^{\text{true}}) = \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 / d_{\text{total}}$ and $\mathcal{X} \subset \mathcal{P}$. By taking $\varepsilon = 1/8$ and $\gamma = 1/2$, we conclude from (11) that

$$\inf_{\hat{\Theta}} \sup_{\Theta^{\text{true}} \in \mathcal{P}} \mathbb{P} \left(\text{Loss}(\hat{\Theta}, \Theta^{\text{true}}) \geq \frac{1}{256} \min \left\{ \alpha^2, \frac{C^{-1} R_{\max} d_{\max}}{d_{\text{total}}} \right\} \right) \geq \frac{1}{2} \left(\frac{3}{4} - \frac{2}{R_{\max} d_{\max}} \right) \geq \frac{1}{8}.$$

This completes the proof. \square

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 $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$.

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

$$\mathcal{L}_{\mathcal{Y},\Omega}(\Theta) = \mathcal{L}_{\mathcal{Y},\Omega}(\Theta^{\text{true}}) + \langle \nabla_{\Theta} \mathcal{L}_{\mathcal{Y},\Omega}, \text{Vec}(\Theta - \Theta^{\text{true}}) \rangle + \frac{1}{2} \text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla_{\Theta}^2 \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta}) \text{Vec}(\Theta - \Theta^{\text{true}}), \quad (12)$$

where

1. $\nabla_{\Theta} \mathcal{L}_{\mathcal{Y},\Omega}$ is a length- d_{total} vector with $|\Omega|$ nonzero entries, and each entry is upper bounded by $L_\alpha > 0$.
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 $-U_\alpha < 0$.

Similar to (3) and (7), we have

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

and

$$\text{Vec}(\Theta - \Theta^{\text{true}})^T \nabla_{\Theta}^2 \mathcal{L}_{\mathcal{Y},\Omega}(\check{\Theta}) \text{Vec}(\Theta - \Theta^{\text{true}}) \leq -L_\alpha \|\Theta - \Theta^{\text{true}}\|_{F,\Omega}^2. \quad (13)$$

Combining (12)-(13) with the fact that $\nabla_{\Theta} \mathcal{L}_{\mathcal{Y}, \Omega}(\hat{\Theta}) \geq \nabla_{\Theta} \mathcal{L}_{\mathcal{Y}, \Omega}(\Theta^{\text{true}})$, we have

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_{F, \Omega} \leq \frac{2C_2 U_{\alpha} r_{\max}^{(K-1)/2}}{L_{\alpha}} \sqrt{\sum_k d_k}. \quad (14)$$

Lastly, we invoke the result regarding the closeness of Θ to its sampled version Θ_{Ω} , under the entrywise bound condition. Note that $\|\hat{\Theta} - \Theta^{\text{true}}\|_{\infty} \leq 2\alpha$ and $\text{rank}(\hat{\Theta} - \Theta^{\text{true}}) \leq 2r$. By Lemma 9, $\|\hat{\Theta} - \Theta^{\text{true}}\|_M \leq 2^{3(K-1)/2} \left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Therefore, the condition in Lemma 10 holds with $\beta = 2^{3(K-1)/2} \left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. Applying Lemma 10 to (14) gives

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

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,

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

provided that $r_{\max} = O(1)$. □

1.3 Auxiliary lemmas

We begin with various notion of tensor norms that are useful for the proofs of the main theorems.

Definition 1 (Atomic M-norm [?]). Define $T_{\pm} = \{\mathcal{T} \in \{\pm 1\}^{d_1 \times \dots \times d_K} : \text{rank}(\mathcal{T}) = 1\}$. The atomic M-norm of a tensor $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$ is defined as

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

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

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

Definition 3 (Nuclear norm [?]). The nuclear norm of a tensor $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$ is defined as

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

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

Lemma 1 (M-norm and infinity norm [?]). Let $\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be an order- K , rank- (r_1, \dots, 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 2 (Nuclear norm and F-norm). Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ be an order- K tensor with Tucker rank (r_1, \dots, r_K) . Then

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\sum_k r_k}{\max_k r_k}} \|\mathcal{A}\|_F,$$

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 [?], Corollary 4.11, and the invariance relationship between a tensor and its Tucker core [?, Section 6], we have

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_{k \geq 2} r_k}{\max_{k \geq 2} r_k}} \|\mathcal{A}_{(1)}\|_*, \quad (15)$$

where $\mathcal{A}_{(1)}$ is a d_1 -by- $\prod_{k \geq 2} d_k$ matrix with matrix rank r_1 . Furthermore, the relationship between the matrix norms implies that $\|\mathcal{A}_{(1)}\|_* \leq \sqrt{r_1} \|\mathcal{A}_{(1)}\|_F = \sqrt{r_1} \|\mathcal{A}\|_F$. Combining this fact with the inequality (15) yields the final claim. \square

Lemma 3. Let \mathcal{A}, \mathcal{B} be two order- K tensors of the same dimension. Then

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

Proof. By [?], Proposition 3.1, there exists a nuclear norm decomposition of \mathcal{B} , such that

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

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

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

which completes the proof. \square

The following lemma provides the bound on the spectral norm of random tensors. The result was firstly presented in [?], and we adopt the version from [?].

Lemma 4 (?). Suppose that $\mathcal{S} = \llbracket s_\omega \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$ is an order- K tensor whose entries are independent random variables that satisfy

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

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

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

with probability at least $1 - \delta$.

Lemma 5. Suppose that $\mathcal{S} = \llbracket s_\omega \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$ is an order- K tensor whose entries are independent random variables that satisfy

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

Then we have

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

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.

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

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

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

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 6. 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 $p_\ell, q_\ell > 0$ for all $i \in [L]$. Then, the Kullback-Leibler (KL) divergence satisfies that

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

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

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

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

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

\square

Lemma 7 (KL divergence and F-norm). *Let $\mathcal{Y} \in [L]^{d_1 \times \dots \times d_K}$ be an ordinal tensor generated from the model (1) with the link function f and parameter tensor Θ . Let \mathbb{P}_Θ denote the joint categorical distribution of $\mathcal{Y}|\Theta$ induced by the parameter tensor Θ , where $\|\Theta\|_\infty \leq \alpha$. Define*

$$A_\alpha = \min_{\ell \in [L], |\theta| \leq \alpha} [f(\theta + b_\ell) - f(\theta + b_{\ell-1})]. \quad (16)$$

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

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

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

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

for all $\ell \in [L]$ and $\omega \in [d_1] \times \dots \times [d_K]$. For notational convenience, we suppress the subscribe in θ_ω and simply write θ (and respectively, θ^*). Based on Lemma 5 and Taylor expansion,

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

where $\eta_\ell, \eta_{\ell-1}$ fall between θ and θ^* , and \tilde{b}_ℓ falls between $b_{\ell-1}$ and b_ℓ . Therefore,

$$KL(\theta || \theta^*) \leq \left(\frac{4(L-2)}{A_\alpha} + \frac{2}{A_\alpha} \right) \dot{f}^2(0) (\theta - \theta^*)^2 = \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) (\theta - \theta^*)^2, \quad (17)$$

where we have used Taylor expansion, the bound (16), and the fact that $\dot{f}(\cdot)$ peaks at zero for an unimodal and symmetric function. Now summing (17) over the index set $\omega \in [d_1] \times \dots \times [d_K]$ gives

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

□

Remark 1. In particular, let \mathbb{P}_0 denote the distribution of $\mathcal{Y}|0$ induced by the zero parameter tensor. Then we have

$$KL(\mathbb{P}_\Theta || \mathbb{P}_0) \leq \frac{2(2L-3)}{A_\alpha} \dot{f}^2(0) \|\Theta\|_F^2.$$

Lemma 8. *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 $\mathcal{X} = \{\Theta_i : i = 1, \dots\} \subset \mathcal{P}$ 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 $\mathbf{0} \in \mathbb{R}^{d_1 \times \dots \times d_K}$;

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

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

Here $C = C(\alpha, L, f, \mathbf{b}) = \frac{(4L-6)f^2(0)}{A_\alpha} > 0$ 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:

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

We then consider the associated set of block tensors:

$$\mathcal{B} = \mathcal{B}(\mathcal{C}) = \{ \Theta \in \mathbb{R}^{d_1 \times \dots \times d_K} : \Theta = \mathbf{A} \otimes \mathbf{1}_{d_3} \otimes \dots \otimes \mathbf{1}_{d_K},$$

$$\text{where } \mathbf{A} = (\mathbf{M} | \dots | \mathbf{M} | \mathbf{O}) \in \mathbb{R}^{d_1 \times d_2}, \mathbf{M} \in \mathcal{C},$$

where $\mathbf{1}_d$ denotes a length- d vector with all entries 1, \mathbf{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 $\Theta(\mathbf{I}, \mathbf{I}, i_3, \dots, i_K) \in \mathbb{R}^{d_1 \times d_2}$ are the same for all fixed $(i_3, \dots, i_K) \in [d_3] \times \dots \times [d_K]$, and furthermore, each subtensor $\Theta(\mathbf{I}, \mathbf{I}, i_3, \dots, i_K)$ itself is filled by copying the matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times R}$ 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 8, there exists a subset $\mathcal{X} \subset \mathcal{B}$ with cardinality $\text{Card}(\mathcal{X}) \geq 2^{Rd_1/8} + 1$ containing the zero $d_1 \times \dots \times d_K$ tensor, such that, for any two distinct elements Θ_i and Θ_j in \mathcal{X} ,

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

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. \square

Lemma 9 (Varshamov-Gilbert bound). *Let $\Omega = \{(w_1, \dots, w_m) : w_i \in \{0, 1\}\}$. Suppose $m > 8$. Then there exists a subset $\{w^{(0)}, \dots, w^{(M)}\}$ of Ω such that $w^{(0)} = (0, \dots, 0)$ 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 10 (Theorem 2.5 in ?). *Assume that a set \mathcal{X} contains element $\Theta_0, \Theta_1, \dots, \Theta_M$ ($M \geq 2$) such that*

- $d(\Theta_j, \Theta_k) \geq 2s > 0, \forall 0 \leq j < k \leq M$;
- $\mathbb{P}_j \ll \mathbb{P}_0, \forall j = 1, \dots, M$, and

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

where $d: \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty]$ is a semi-distance function, $0 < \alpha < 1/8$ and $P_j = P_{\Theta_j}, j = 0, 1, \dots, M$.

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 11 (Lemma 28 in ?). Define $\mathbb{B}_M(\beta) = \{\Theta \in \mathbb{R}^{d_1 \times \dots \times d_K} : \|\Theta\|_M \leq \beta\}$. Let $\Omega \subset [d_1] \times \dots \times [d_K]$ be a random set with $m = |\Omega|$, and assume that each entry in Ω is drawn with replacement from $[d_1] \times \dots \times [d_K]$ 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 $\Theta \in \mathbb{B}_M(\beta)$.