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 $\mathcal{L}_{\mathcal{Y}}(\Theta)$ that

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

for all $\omega \in [d_1] \times \cdots \times [d_K]$. Define $d_{\text{total}} = \prod_k d_k$. Let $\nabla_\Theta \mathcal{L}_{\mathcal{Y}} \in \mathbb{R}^{d_{\text{total}}}$ denote the vector of gradient with respect to $\text{Vec}(\Theta) \in \mathbb{R}^{d_{\text{total}}}$, 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 (1.1), $\nabla^2_\Theta \mathcal{L}_{\mathcal{Y}}$ 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 $\nabla_\Theta \mathcal{L}_{\mathcal{Y}}$ are upper bounded $U_\alpha > 0$, 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

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

$\check{\Theta} = \gamma \Theta^{\text{true}} + (1-\gamma)\Theta$ 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 (1.1). Note that, by Lemma 4,

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

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 (1.1) 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

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

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

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

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 (1.1), (1.1) and (1.1), we have that, with probability at least $1 - \exp(-C_1 \sum_k d_k)$,

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

We next bound the quadratic term in (1.1). Note that

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

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

Combining (1.1), (1.1) and (1.1), 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 6. □

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 $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{P}$ 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 $\mathbb{P}_\Theta$ 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 6, we have

$$KL(\mathbb{P}_{\Theta}||\mathbb{P}_{\mathbf{0}}) \leq C\|\Theta\|_F^2, \tag{8}$$

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 (1.2) with property (iii) of $\mathcal{X}$, we have

$$\mathrm{KL}(\mathbb{P}_{\Theta}||\mathbb{P}_{\mathbf{0}}) \leq \gamma^2 R_{\max} d_{\max}. \tag{9}$$

From (1.2) and the property (i), we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{X})-1}\sum_{\Theta\in\mathcal{X}} \mathrm{KL}(\mathbb{P}_{\Theta},\mathbb{P}_{\mathbf{0}}) \leq \varepsilon \log\left\{\mathrm{Card}(\mathcal{X})-1\right\} \tag{10}$$

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 Theorem 1.1 to (1.2), and in view of the property (iv), we obtain that

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

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 $\varepsilon = 1/8$ and $\gamma = 1/2$, we conclude from (1.2) that

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

This completes the proof. □

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

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

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. $\nabla^2_\Theta\mathcal{L}_{\mathcal{Y},\Omega}$ 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 (1.1) and (1.1), we have

$$|\langle\nabla_\Theta\mathcal{L}_{\mathcal{Y},\Omega},\ \Theta-\Theta^{\text{true}}\rangle| \leq C_2U_\alpha\sqrt{r_{\max}^{K-1}\sum_k d_k}\|\Theta-\Theta^{\text{true}}\|_{F,\Omega}$$

and

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

Combining (1.2)-(1.2) 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

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

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 $\text{rank}(\hat{\Theta} - \Theta^{\text{true}}) \leq 2\boldsymbol{r}$. By Lemma 9, $\|\hat{\Theta} - \Theta^{\text{true}}\|_M \leq 2^{3(K-1)/2}\alpha \left(\frac{\prod r_k}{r_{\max}}\right)^{3/2}$. 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 (1.2) gives

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

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,

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

provided that $r_{\max} = O(1)$. □

1.3 Auxiliary lemmas

We begin with a set of technical lemmas that are useful for the proofs of the main theorems.

Lemma 1 (?). *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*

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

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 $1 - \delta$.

Remark 1. The above lemma provides the bound on the spectral norm of random tensors. The result was firstly presented in **?**, and we adopt the version from **?**.

Lemma 2. *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*

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

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 3. *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*

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

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

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

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 (1.3) yields the final claim. $\square$

Lemma 4. *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

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

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

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

which completes the proof. $\square$

Lemma 5. *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*

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

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

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

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

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

□

Lemma 6. *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*

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

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 $\mathcal{Y} = [\![y_\omega]\!]$ is induced by $\Theta = [\![\theta_\omega]\!]$. 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 $\omega \in [d_1]\times\cdots\times[d_K]$. For notational convenience, we suppress the subscribe in θ_ω and simply write θ (and respectively, θ^*). Based on Lemma 5 and Taylor expansion,

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

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

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

where we have used Taylor expansion, the bound (6), and the fact that $\dot{f}(\cdot)$ peaks at zero for an unimodal and symmetric function. Now summing (1.3) 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 2. 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 7. *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:

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

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 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 $d_1 \times \cdots \times d_K$ tensor, such that, for any two distinct elements Θ_i and Θ_j in $\mathcal{X}$,

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

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. $\square$

Lemma 8 (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 $M \geq 2^{m/8}$.

Theorem 1.1 (?)**.** *Assume that a set $\mathcal{X}$ contains element $\Theta_0, \Theta_1, \ldots, \Theta_M$ ($M \geq 2$) such that*

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

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

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

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

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

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

Lemma 9 (?)**.** *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 10 (?)**.** *Define $\mathbb{B}_M(\beta) = \{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} \colon \|\Theta\|_M \leq \beta\}$. 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 $\Theta \in \mathbb{B}_M(\beta)$*.*