Evidence Theory about statistical convergence

Zhuoyan Xu

7/30/2019

This is an extension of **Theorem 1 (Statistical convergence)** in [1].

For the binary tensor $\mathcal{Y} = [y_{i_1,\ldots,i_K}] \in \{0,1\}^{d_1\times\cdots\times d_K}$, we assume its entries are realizations of independent Bernoulli random variables, in that:

$$\mathcal{Y}|\Theta \sim \text{Bernoulli}\{f(\Theta)\}, \quad \text{with } P\left(y_{i_1,\ldots,i_K}=1\right) = f\left(\theta_{i_1,\ldots,i_K}\right)$$

First we define:

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \sum_{i_1,\ldots,i_K}\left[1_{\left\{y_{i_1,\ldots,i_K}=1\right\}}\log f\left(\theta_{i_1,\ldots,i_K}\right) + 1_{\left\{y_{i_1,\ldots,i_K}=0\right\}}\log\left\{1-f\left(\theta_{i_1,\ldots,i_K}\right)\right\}\right]$$

We assumed parameter tensor Θ admits a tucker decomposition as:

$$\Theta = \mathcal{G}\times_1 N_1^{d_1 r_1}\times_2 N_2^{d_2 r_2}\ldots\times_K N_K^{d_K r_K}$$

Where $d_1, d_2, \ldots, d_K$ is dimension of tensor. The $r_1, r_2, \ldots, r_K$ is the dimension of core tensor. The $N_1, \ldots, N_K$ are all orthogonal matrix. We use $\text{rank}_T(\Theta) = (r_1, \ldots, r_K)$ to denotes the tucker rank.

To incorporate the tucker decomposition structure, we consider a constrained optimization:

$$\max_{\Theta\in\mathcal{D}} \mathcal{L}_{\mathcal{Y}}(\Theta), \quad \text{where } \mathcal{D}\subset\mathcal{S} = \left\{\Theta : \text{rank}_T(\Theta) = (r_1,\ldots,r_K)\,,\ \text{and } \|\Theta\|_\infty \le \alpha\right\}$$

Then we define L_α and γ_α the same as section3.2 in [1].

Define:

$$\text{Loss}\left(\hat{\Theta}, \Theta_{\text{true}}\right) = \frac{1}{\sqrt{\prod_k d_k}}\left\|\hat{\Theta} - \Theta_{\text{true}}\right\|_F$$

We have:

Theorem 1 *Suppose $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ is an order-K binary tensor following model (2), with the link function f, and the true coefficient tensor $\Theta_{true} \in \mathcal{D}$, Then there exist an absolute constant $C_1 > 0$, and a constant $C_2 > 0$ that depends only on K, such that, with probability at least $1 - \exp\left(-C_1 \log K \sum_k d_k\right)$.*

$$\operatorname{Loss}\left(\hat{\Theta}_{MLE}, \Theta_{true}\right) \leq \min\left\{2\alpha, C_2 \frac{L_\alpha}{\gamma_\alpha} \sqrt{\frac{\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k}{\Pi_k d_k}}\right\}$$

proof The main proof following Appendix B.1 in [1].

By the second-order Taylor's theorem:

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{ true}}) + \langle S_{\mathcal{Y}}(\Theta_{\text{ true}}), \Theta - \Theta_{\text{ true}}\rangle + \frac{1}{2}\operatorname{vec}(\Theta - \Theta_{\text{ true}})^T \mathcal{H}_{\mathcal{Y}}(\check{\Theta}) \operatorname{vec}(\Theta - \Theta_{\text{ true}})$$

We first bound the linear term, by lemma 4 in [1],

$$|\langle S_{\mathcal{Y}}(\Theta_{\text{ true}}), \Theta - \Theta_{\text{ true}}\rangle| \leq \|S_{\mathcal{Y}}(\Theta_{\text{ true}})\|_\sigma \|\Theta - \Theta_{\text{ true}}\|_*$$

By lemma 2 in [1], with probability at least $1 - \exp\left(-C_1 \log K \sum_k d_k\right)$:

$$\|S_{\mathcal{Y}}(\Theta_{\text{ true}})\|_\sigma \leq C_2 L_\alpha \log K \sqrt{\sum_k d_k} \tag{1}$$

According to Theorem 9 and Corollary 10 in [2], for the orthonormal tucker decomposition:

$$\|\Theta\|_* = \|\mathcal{G}\|_*$$

According to Theorem 6 and Corollary 8 in [2], for the orthonormal tucker decomposition:

$$\|\Theta\|_F = \|\mathcal{G}\|_F$$

According to Theorem 5.2 in [3], For any positive integers $K \geq 3$, $d_1 \leq \cdots \leq d_K$ and an K-order tensor $\mathcal{A} \in R^{d_1 \times \cdots \times d_K}$, we have:

$$\left\|A_{(1)}\right\|_* \leq \|\mathcal{A}\|_* \leq \sqrt{\prod_{k=2}^{K-1} d_k}\left\|A_{(1)}\right\|_*$$

Without loss of generality, we assume $r_1 \leq r_2 \leq \ldots \leq r_K$, then:

$$
\begin{aligned}
\|\Theta\|_* =& \|\mathcal{G}\|_* \leq \sqrt{\prod_{k=2}^{K-1} r_k} \left\|G_{(1)}\right\|_* \leq \sqrt{\prod_{k=2}^{K-1} r_k \sqrt{r_1}} \left\|G_{(1)}\right\|_F \\
=& \sqrt{\prod_{k=1}^{K-1} r_k} \|\mathcal{G}\|_F = \sqrt{\prod_{k=1}^{K-1} r_k} \|\Theta\|_F
\end{aligned}
$$

Then we have:

$$
\|\Theta - \Theta_{true}\|_* \leq \sqrt{\prod_{k=1}^{K-1} 2r_k} \|\Theta - \Theta_{true}\|_F \tag{2}
$$

Combining 1 and 2, we have, with probability at least $1-\exp\left(-C_1 \log K \sum_k d_k\right)$:

$$
\left|\left\langle S_{\mathcal{Y}}\left(\Theta_{\text{ true }}\right), \Theta - \Theta_{\text{ true }}\right\rangle\right| \leq C_2 L_\alpha \sqrt{\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k} \left\|\Theta - \Theta_{\text{ true }}\right\|_F
$$

where the constant C_2 absorbs all factors that depend only on K.
The following steps are the same as proof in section B.1 in [1].
Henceforth,

$$
\frac{1}{\sqrt{\prod_k d_k}} \left\|\hat{\Theta} - \Theta_{\text{ true }}\right\|_F \leq \frac{2C_2 L_\alpha \sqrt{\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k}}{\gamma_\alpha \sqrt{\Pi_k d_k}} = 2C_2 \frac{L_\alpha}{\gamma_\alpha} \sqrt{\frac{\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k}{\Pi_k d_k}}
$$

References

[1] Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*, 2018.

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

[3] Shenglong Hu. Relations of the nuclear norm of a tensor and its matrix flattenings. *Linear Algebra and its Applications*, 478:188–199, 2015.