Joint estimation of $\hat{\Theta}$ and $\hat{\boldsymbol{b}}$ in the tensor ordinal model

Miaoyan Wang, Feb 14, 2020.

Remark 1 (Notation)**.** For notational convenience, we consider the special case $d_1 = \cdots = d_K = d$. We focus on the asymptotic region $d \to \infty$ and treat all other quantities as constants, i.e., $\boldsymbol{r}, L, U_{\alpha,\beta,\Delta}, L_{\alpha,\beta,\Delta}, ... = \mathcal{O}(1)$. In particular,

1. $\prod_k d_k \asymp d^K$, $\sum_k d_k \asymp d$.
2. Denote $n_{\max} = \max_\ell n_\ell$ and $n_{\min} = \min_\ell(n_\ell + n_{\ell+1})$. Then, $n_{\max} \asymp d^K$ (important!) but $\mathcal{O}(1) \leq n_{\min} \leq \mathcal{O}(d^K)$.

Note: we use $C >$ to denote a positive constant whose value may change from line to line.

1 Results

1.1 Current bound

Total MSE:

$$\begin{aligned}
\mathrm{MSE}((\hat{\Theta}, \hat{\boldsymbol{b}}), (\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})) &\stackrel{\text{def}}{=} \frac{\prod_k d_k \mathrm{MSE}(\hat{\Theta}, \Theta^{\text{true}}) + (L-1)\mathrm{MSE}(\hat{\Theta}, \Theta^{\text{true}})}{\prod_k d_k + L - 1} \\
&\leq \frac{c_1 \sum_k d_k + c_2 \frac{n_{\max}^2}{n_{\min}^2}}{\prod_k d_k + L - 1} \\
&= \mathcal{O}\left(\frac{\sum_k d_k}{\prod_k d_k}\right) + \mathcal{O}\left(\frac{n_{\max}^2}{n_{\min}^2} \frac{1}{\prod_k d_k}\right) \qquad (1)
\end{aligned}$$

Remark 2. Unfortunately, this total MSE bound does not converge to zero. In the worse case, the bound can be $\asymp d^K$; for example, the bound diverges when $n_{\min} \asymp 1$ (and $n_{\max} \asymp d^K$).

Remark 3. The total MSE bound converges to zero only when

$$\frac{n_{\max}}{n_{\min}} \ll d^{K/2}, \quad \text{or equivalently,} \quad n_{\min} \gg d^{K/2}. \qquad (2)$$

In other words, the current bound (1) tolerates only certain imbalanced classes for which $\mathcal{O}(d^{K/2}) \leq n_{\min} \leq \mathcal{O}(d^K)$.

1.2 Sharper bound

We will prove a sharper bound on the linear term

$$\begin{aligned}
(\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}})^T \nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}}) &\leq \|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F \|\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathcal{Y}}(\boldsymbol{b}^{\text{true}})\|_F \\
&\leq C\|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F \sqrt{d^{K+2}},
\end{aligned}$$

where the last inequality is followed from a sharper bound on the gradient:

$$\left| \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})} \right| \leq C\sqrt{d^{K+2}}, \qquad \text{for all } \ell \in [L-1]. \qquad (3)$$

Suppose (3) holds. Following the same line in the current proof (i.e., Taylor expansion, quadratic bound, etc.), we have

$$\|\boldsymbol{b}^{\text{true}} - \hat{\boldsymbol{b}}\|_F \leq C\frac{d^{(K+2)/2}}{n_{\min}}.$$

(Final results.) Therefore, the total MSE:

$$\text{Total MSE} \leq \mathcal{O}\left(\frac{1}{d^{K-1}}\right) + \mathcal{O}\left(\frac{d^{K+2}}{n_{\min}^2 d^K}\right) = \mathcal{O}\left(\frac{d^2}{\min\{d^{K+1}, n_{\min}^2\}}\right) \tag{4}$$

which convergences to zero whenever

$$n_{\min} \gg d. \tag{5}$$

In other words, the new bound (4) tolerates highly imbalanced classes for which $\mathcal{O}(\sqrt{d}) \leq n_{\min} \leq \mathcal{O}(d^K)$.

Remark 4. The consistency condition (5) is more relaxed than (2) for all $K \geq 3$. Both bounds agree in the matrix case ($K = 2$).

Remark 5. When $n_{\min} \gg d^{(K+1)/2}$, the error in $\hat{\Theta}$ dominates the total MSE. Then $n_{\min} \ll d^{(K+1)/2}$, the error in $\hat{\boldsymbol{b}}$ dominates the total MSE

2 Proofs

Now we prove the inequality (3).

Lemma 1 (Sharper Bound on Gradients)**.** *Consider the same set-up as in Theorem A.1. Then, with very high probability,*

$$\left|\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{true})}\right| \leq C_1 d^{K/2} + C_2 d^{K/2}\|\Theta^{true} - \hat{\Theta}\|_F, \quad \textit{for all } \ell \in [L-1].$$

where $C_1, C_2 > 0$ are two constants. In particular, there exists a constant $d_0 \in \mathbb{N}_+$, such that for all $d \geq d_0$,

$$\left|\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{true})}\right| \leq C d^{(K+2)/2}, \quad \textit{for all } \ell \in [L-1].$$

Corollary 1 (MSE for $\hat{\boldsymbol{b}}$)**.** *Under the same set-up as in Theorem A.1, we have*

$$\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{true}\|_F \leq \frac{C_1 d^{K/2} + C_2 d^{K/2}\|\hat{\Theta} - \Theta^{true}\|_F}{n_{\min}} \leq \frac{C d^{(K+2)/2}}{n_{\min}}.$$

Remark 6. The bound (3) is sharper than the trivial bound $|b_\ell^{\text{true}} - \hat{b}_\ell| \leq 2\beta$. In particular, $\|\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}\|_F \to 0$ as $n_{\min} \asymp (d^{(K+2)/2}) \to \infty$.

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

Note that

$$\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})} = \underbrace{\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})} - \mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})}\right]}_{:=\text{A}} + \underbrace{\mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\hat{\Theta}, \boldsymbol{b}^{\text{true}})}\right] - \mathbb{E}_{\mathcal{Y}}\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial b_1}\Big|_{(\Theta^{\text{true}}, \boldsymbol{b}^{\text{true}})}\right]}_{:=B} \tag{6}$$

where 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 (6) separately. The term A in (6) is the stochastic deviation of log-likelihood to its expectation:

$$A=\sum_{\omega\in\Omega}\left\{\underbrace{\left[\mathbb{1}_{\{y_\omega=1\}}-g_1(\theta_\omega^{\text{true}})\right]\frac{\dot f(b_1-\hat\theta_\omega)}{g_1(\hat\theta_\omega)}-\left[\mathbb{1}_{\{y_\omega=2\}}-g_2(\theta_\omega^{\text{true}})\right]\frac{\dot f(b_1-\hat\theta_\omega)}{g_2(\hat\theta_\omega)}}_{:=W_\omega}\right\}.$$

Note that $\{W_\omega\}$ are i.i.d. random variables, and each W_ω has zero mean and bounded variance:

$$\mathrm{Var}(W_\omega)\le C\left(g_1(\theta_\omega^{\text{true}})(1-g_1(\theta_\omega^{\text{true}}))+g_2(\theta_\omega^{\text{true}})(1-g_2(\theta_\omega^{\text{true}}))\right)\le C/2.$$

By central limit theorem (verify..)

$$\sum_{\omega\in\Omega}W_\omega\xrightarrow{\mathcal{D}}N(0,Cd^K),\quad\text{as}\quad d^K\to\infty.$$

Hence, with very high probability,

$$|A|=\left|\sum_{\omega\in\Omega}W_\omega\right|\le Cd^{K/2}. \tag{7}$$

The second term B in (6) is the bias induced by the inaccuracy of $\hat\Theta$:

$$\begin{aligned}B&=\sum_{\omega\in\Omega}g_1(\theta_\omega^{\text{true}})\left(\frac{\dot f(b_1-\hat\theta_\omega)}{g_1(\hat\theta_\omega)}-\frac{\dot f(b_1-\theta_\omega^{\text{true}})}{g_1(\theta_\omega^{\text{true}})}\right)-\sum_{\omega\in\Omega}g_2(\theta_\omega^{\text{true}})\left(\frac{\dot f(b_2-\hat\theta_\omega)}{g_2(\hat\theta_\omega)}-\frac{\dot f(b_2-\theta_\omega^{\text{true}})}{g_2(\theta_\omega^{\text{true}})}\right)\\
&\le\sum_{\omega\in\Omega}g_1(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}}-\hat\theta_\omega)\left\{\frac{\partial}{\partial\theta}\left(\frac{\dot f(b_1-\theta)}{g_1(\theta)}\right)\Big|_{\rho\hat\theta_\omega+(1-\rho)\theta_\omega^{\text{true}}}\right\}\\
&\quad-\sum_{\omega\in\Omega}g_2(\theta_\omega^{\text{true}})(\theta_\omega^{\text{true}}-\hat\theta_\omega)\left\{\frac{\partial}{\partial\theta}\left(\frac{\dot f(b_2-\theta)}{g_2(\theta)}\right)\Big|_{\rho'\hat\theta_\omega+(1-\rho')\theta_\omega^{\text{true}}}\right\}\\
&\le C\sum_{\omega\in\Omega}\left[g_1(\theta_\omega^{\text{true}})-g_2(\theta_\omega^{\text{true}})\right]\left(\theta_\omega^{\text{true}}-\hat\theta_\omega\right).\end{aligned}$$

By Cauchy-Schwartz inequality,

$$|B|\le Cd^{K/2}\|\Theta^{\text{true}}-\hat\Theta\|_F. \tag{8}$$

Plugging (7) and (8) back to (6) yields that

$$\left|\frac{\partial\mathcal{L}_{\mathcal{Y}}}{\partial b_\ell}\Big|_{(\hat\Theta,\boldsymbol{b}^{\text{true}})}\right|\le C_1d^{K/2}+C_2d^{K/2}\|\Theta^{\text{true}}-\hat\Theta\|_F$$

holds with very high probability. The second inequality in the conclusion comes from the fact that $\|\Theta^{\text{true}}-\hat\Theta\|_F\le\mathcal{O}(d)$ as $d\to\infty$. $\square$