

# Algorithmic guarantees

## 1 General setting

We first introduce the regularity condition on the loss function  $\mathcal{L}$  and set  $\mathcal{S}$ .

**Definition 1.** Let  $f$  be a real-valued function. We say  $f$  satisfies  $\text{RCG}(\alpha, \beta, \mathcal{S})$  condition for  $\alpha, \beta > 0$  and the set  $\mathcal{S}$  if,

$$\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq \alpha \|x - x'\|_2^2 + \beta \|\nabla f(x) - \nabla f(x')\|_2^2,$$

for any  $x, x' \in \mathcal{S}$ .

Define

$$\begin{aligned} \bar{\lambda} &:= \max \{ \sigma_{\max}(\mathcal{M}_1(\mathcal{B})), \sigma_{\max}(\mathcal{M}_2(\mathcal{B})), \sigma_{\max}(\mathcal{M}_3(\mathcal{B})) \}, \\ \underline{\lambda} &:= \min \{ \sigma_{\min}(\mathcal{M}_1(\mathcal{B})), \sigma_{\min}(\mathcal{M}_2(\mathcal{B})), \sigma_{\min}(\mathcal{M}_3(\mathcal{B})) \}, \end{aligned}$$

and  $\kappa = \bar{\lambda}/\underline{\lambda}$  can be regarded as a tensor condition number. Here  $\mathcal{M}_i$  is the matricization operator with respect to  $i$ -th mode.

We define some constants related to side information  $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$  as

$$\gamma := \prod_{k=1}^3 \sigma_{\max}(\mathbf{X}_k)^2, \quad \gamma_1 := \prod_{k=1}^3 \sigma_{\min}(\mathbf{X}_k)^2, \quad \text{and} \quad \gamma_2 := \prod_{k=1}^3 \|\mathbf{X}_k\|_F^2.$$

Without loss of generality, we scale the the side information matrices  $\mathbf{X}_k$  so that  $\|\mathbf{X}_k\|_{\infty} \leq 1$  for all  $k = 1, 2, 3$ .

**Lemma 1.1.** Suppose  $f: \mathbb{R}^{d_1 \times d_2 \times d_3} \rightarrow \mathbb{R}$  satisfies  $\text{RCG}(\alpha, \beta, \mathcal{S})$  where  $\mathcal{S} = \{\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3)\}$ . Define  $g: \mathbb{R}^{p_1 \times p_2 \times p_3} \rightarrow \mathbb{R}$  as  $g(\mathcal{B}) = f(\mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})$  for all  $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$  and  $\mathcal{S}' = \{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} : \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3)\}$ . Then,  $g$  satisfies  $\text{RCG}(\alpha\gamma_1, \beta/\gamma_2, \mathcal{S}')$ .

*Proof.* Notice that for any  $\mathcal{B}_1, \mathcal{B}_2 \in \mathcal{S}'$ ,

$$\begin{aligned} &\langle \nabla g(\mathcal{B}_1) - \nabla g(\mathcal{B}_2), \mathcal{B}_1 - \mathcal{B}_2 \rangle \\ &= \langle (\nabla f(\mathcal{B}_1 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}) - \nabla f(\mathcal{B}_2 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})) \times \{\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\}, \mathcal{B}_1 - \mathcal{B}_2 \rangle \\ &= \langle \nabla f(\mathcal{B}_1 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}) - \nabla f(\mathcal{B}_2 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}), (\mathcal{B}_1 - \mathcal{B}_2) \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\} \rangle \\ &\geq \alpha \|(\mathcal{B}_1 - \mathcal{B}_2) \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}\|_F^2 + \beta \|\nabla f(\mathcal{B}_1 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}) - \nabla f(\mathcal{B}_2 \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})\|_F^2 \\ &\geq \alpha \gamma_1 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2 + \frac{\beta}{\gamma_2} \|\nabla g(\mathcal{B}_1) - \nabla g(\mathcal{B}_2)\|_F^2, \end{aligned}$$

where the first inequality uses the fact that  $f$  satisfies  $\text{RCG}(\alpha, \beta, \mathcal{S})$  and the last inequality uses Cauch Schwartz inequality.  $\square$

Since negative log-likelihoods of poisson and binomial distribution are not strongly convex and smooth in the unbounded domain. We thus introduce the following assumption on  $\mathcal{B}_{\text{true}}$  to ensure that  $\mathcal{B}_{\text{true}}$  is in a bounded set.

**Assumption 1.** Suppose  $\mathcal{B}_{\text{true}} = \mathcal{C}^* \times \{\mathbf{M}_1^*, \mathbf{M}_2^*, \mathbf{M}_3^*\}$ , where  $\mathbf{M}_k^* \in \mathbb{R}^{p_k \times r_k}$  is a orthogonal matrix for  $k = 1, 2, 3$ . There exists some constants  $\{\mu_k\}_{k=1}^3$ ,  $B$  such that  $\|\mathbf{M}_k^*\|_{2,\infty}^2 \leq \frac{\mu_k r_k}{p_k}$  for  $k = 1, 2, 3$  and  $\bar{\lambda} \leq$

$B \sqrt{\frac{\prod_{k=1}^3 p_k}{\prod_{k=1}^3 \mu_k r_k}}$ . Here  $\|\mathbf{M}_k^*\|_{2,\infty}$  is the largest row-wise  $\ell_2$  norm of  $\mathbf{M}_k^*$ .

**Remark 1.** This condition guarantees that  $\mathcal{B}_{\text{true}}$  is entry-wise upperbounded by  $B$ , which guarantees the local strong convexity and smoothness of the negative log-likelihood function.

We define searching space  $\mathcal{S}$  as follows:

$$\begin{aligned}\mathcal{S} &= \mathcal{S}_c \times \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_3, \text{ where} \\ \mathcal{S}_k &= \left\{ (\mathbf{M}_k \in \mathbb{R}^{p_k \times r_k} : \|\mathbf{M}_k\|_{2,\infty} \leq b \sqrt{\frac{\mu_k r_k}{p_k}} \right\} \text{ for } k = 1, 2, 3, \\ \mathcal{S}_c &= \left\{ \mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \max_k \|\mathcal{M}_k(\mathcal{C})\|_2 \leq b^{-3} B \sqrt{\frac{\prod_{k=1}^3 p_k}{\prod_{k=1}^3 \mu_k r_k}} \right\}.\end{aligned}$$

## 2 General tensor case from exponential family

Suppose we observe  $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$  from exponential family with canonical parameter  $\Theta = \mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$  such that

$$\mathbb{P}(\mathcal{Y}_{ijk} | \Theta_{ijk}) = c(\mathcal{Y}_{ijk}, \phi) \exp \left( \frac{\mathcal{Y}_{ijk} \Theta_{ijk} - b(\Theta_{ijk})}{\phi} \right),$$

where  $b(\cdot)$  is a known function,  $\phi > 0$  is the dispersion parameter, and  $c(\cdot)$  is a known normalization function. Then we consider the following negative log-likelihood to estimate  $\mathcal{B}_{\text{true}}$ ,

$$\mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = -\langle \mathcal{Y}, \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\} \rangle + \sum_{ijk} b(\mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}).$$

**Example 1 (Gaussian).** Suppose we observe  $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$  that satisfies

$$\mathcal{Y}_{ijk} \sim \text{Gaussian}(\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}, \sigma) \text{ independently.}$$

Then the corresponding negative log-likelihood is,

$$\mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = \frac{1}{2} \|\mathcal{Y} - \mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}\|_F^2$$

**Example 2 (Poisson).** Suppose we observe  $\mathcal{Y} \in \mathbb{N}^{d_1 \times d_2 \times d_3}$  that satisfies

$$\mathcal{Y}_{ijk} \sim \text{Poisson}(\exp(\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})) \text{ independently.}$$

Then the corresponding negative log-likelihood is,

$$\mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = \sum_{ijk} \left( -\mathcal{Y}_{ijk} [\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}]_{ijk} + \exp([\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}]_{ijk}) \right).$$

**Example 3 (Bernoulli).** Suppose we observe  $\mathcal{Y} \in \{0, 1\}^{d_1 \times d_2 \times d_3}$  that satisfies

$$\mathcal{Y}_{ijk} \sim \text{Bernoulli}(\text{logistic}(\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})) \text{ independently,}$$

where  $\text{logistic}(x) = (1 + e^{-x})^{-1}$ . Then the corresponding negative log-likelihood is,

$$\mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = - \sum_{ijk} \left( \mathcal{Y}_{ijk} [\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}]_{ijk} + \log \left( 1 + \exp([\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}]_{ijk}) \right) \right).$$

**Theorem 2.1.** Suppose Assumption 1 holds for Poisson and Bernoulli cases and

1. Initialization:  $\|\mathcal{B}_{\text{true}} - \mathcal{B}^{(0)}\|_F^2 \leq c_1 \alpha \beta \kappa^{-2} \underline{\lambda}^2$
2. Signal to noise ratio:  $\underline{\lambda}^2 \geq c_2 \frac{\kappa^4}{\alpha^3 \beta} \left( \frac{\prod_k r_k}{\max_k r_k} \gamma \sum_k p_k \right).$

where  $c_1, c_2 > 0$  are universal constants. Then, with probability at least  $1 - \exp(-c_3 \sum_k p_k)$ , we have

$$\|\mathcal{B}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2 \lesssim \underbrace{\left( \frac{\prod_k r_k}{\max_k r_k} \sum_k p_k \right)}_{\text{Statistical error}} + \underbrace{\rho^T \|\mathcal{B}^{(0)} - \mathcal{B}_{\text{true}}\|_F^2}_{\text{Algorithmic error}},$$

for all  $t \geq 1$  where  $\rho \in (0, 1)$  is a contraction parameter, and  $c_1, c_2, c_3 > 0$  are some constants.

**Remark 2.** Combining Lemma 1.1 and proofs of the theorems in Han et al. [2020], we have

1. (Gaussian case) We have  $\alpha = \frac{\gamma_1}{2}$  and  $\beta = \frac{1}{2\gamma_2}$ .
2. (Poisson case) We have  $\alpha = \frac{\gamma_1}{e^B + e^{-B}}$  and  $\beta = \frac{1}{\gamma_2(e^B + e^{-B})}$
3. (Bernoulli case) We have  $\alpha = \frac{\gamma_1}{2(e^B + 3)}$  and  $\beta = \frac{1}{2\gamma_2}$

*Proof.* We bound the statistical error and apply the result to Theorem 3.1. in Han et al. [2020]. We show that with probability at least  $1 - \exp(-C_2 \sum_k p_k)$ ,

$$\xi = \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \nabla \mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3), \mathcal{T} \rangle \leq C_1 \sqrt{\frac{\prod_k r_k}{\max_k r_k} \gamma \sum_k p_k},$$

for some constants  $C_1, C_2 > 0$ . By definition of  $\mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ , we have

$$\begin{aligned} \xi &= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \nabla \mathcal{L}(\mathcal{B} | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3), \mathcal{T} \rangle \\ &= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle (\mathcal{Y} - b'(\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})) \times \{\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\}, \mathcal{T} \rangle \\ &= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \mathcal{E} \times \{\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\}, \mathcal{T} \rangle, \end{aligned} \tag{1}$$

where  $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{Y} - b'(\mathcal{B}_{\text{true}} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\})$ . Based on Proposition 3, the boundedness of  $b''(\cdot)$  implies that  $\mathcal{E}$  is a sub-Gaussian- $(\phi U)$  tensor. Let  $\check{\mathcal{E}} \stackrel{\text{def}}{=} \mathcal{E} \times \{\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\}$ . By proposition 2,  $\check{\mathcal{E}}$  is a  $(p_1, p_2, p_3)$ -dimensional sub-Gaussian tensor with parameter bounded by  $\phi U \sqrt{\gamma}$ . Applying Cauchy-Schwartz inequality to (1) yields

$$\xi \leq \|\check{\mathcal{E}}\|_2 \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \|\mathcal{T}\|_*, \tag{2}$$

where  $\|\cdot\|_2$  denotes the tensor spectral norm and  $\|\cdot\|_*$  denotes the tensor nuclear norm. The nuclear norm  $\|\mathcal{T}\|_*$  is bounded by  $\|\mathcal{T}\|_* \leq \sqrt{\frac{\prod_{k=1}^3 r_k}{\max_k r_k}} \|\mathcal{T}\|_F$  [Wang et al., 2017, Wang and Li, 2018]. The spectral norm  $\|\check{\mathcal{E}}\|_2$  is bounded by  $\|\check{\mathcal{E}}\|_2 \leq C_1 \sqrt{\gamma \sum_k p_k}$  with probability at least  $1 - \exp(-C_2 \log K \sum_k p_k)$  [Tomioka and Suzuki, 2014]. Combining these two bounds with (2), we have, with probability at least  $1 - \exp(-C_2 \sum_k p_k)$ ,

$$\xi \leq C_1 \sqrt{\frac{\prod_k r_k}{\max_k r_k} \gamma \sum_k p_k}.$$

Applying the upper bound of  $\xi$  to Theorem 3.1 in Han et al. [2020] completes the proof.  $\square$

## References

- Rungang Han, R. Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *ArXiv*, abs/2002.11255, 2020.
- Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.
- Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*, 2018.
- 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.