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 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

$$\bar{\lambda} := \max \left\{ \sigma_{\max}\left(\mathcal{M}_1(\mathcal{B})\right), \sigma_{\max}\left(\mathcal{M}_2(\mathcal{B})\right), \sigma_{\max}\left(\mathcal{M}_3(\mathcal{B})\right) \right\},$$
$$\underline{\lambda} := \min \left\{ \sigma_{\min}\left(\mathcal{M}_1(\mathcal{B})\right), \sigma_{\min}\left(\mathcal{M}_2(\mathcal{B})\right), \sigma_{\min}\left(\mathcal{M}_3(\mathcal{B})\right) \right\},$$

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 $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ as

no need for this quantity

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

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

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

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

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

can be improved to beta / gamma_1

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

definition of singular value.

|B|_F * min singular value (A) \leq |AB|_F \leq |B|_F *max singular value (A)

Convince yourself this property by geometric interpretation of SVD.

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.

I have pointed out this property to you back in 2019 (check your note on random sketching).

Assumption 1. Suppose $\mathcal{B}_{\text{true}} = \mathcal{C}^* \times \{\boldsymbol{M}_1^*, \boldsymbol{M}_2^*, \boldsymbol{M}_3^*\}$, where $\boldsymbol{M}_k^* \in \mathbb{R}^{p_k \times r_k}$ is a orthogonal matrix for $k = 1, 2, 3$. There exists some constants $\{\mu_k\}_{k=1}^3$, B such that $\|\boldsymbol{M}_k^*\|_{2,\infty}^2 \leq \frac{\mu_k r_k}{p_k}$ 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 $\|\boldsymbol{M}_k^*\|_{2,\infty}$ is the largest row-wise ℓ_2 norm of $\boldsymbol{M}_k^*$.

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 spaace $\mathcal{S}$ as follows:

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

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 $\Theta = \mathcal{B}_{\text{true}} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ 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 $\mathcal{B}_{\text{true}}$,

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

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

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

Then the corresponding negative log-likelihood is,

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

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

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

Then the corresponding negative log-likelihood is,

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

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

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

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

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

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. Let $\mathcal{B}^{(t)}$ be t-th iteration output of the alternating gradient descent algorithm with a suitable step size. Then, with probability at least $1 - \exp(-c_3 \sum_k p_k r_k)$, we have

$$\|\mathcal{B}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2 \lesssim \underbrace{\left(r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k\right)}_{\text{Statistical error}} + \underbrace{\rho^T \|\mathcal{B}^{(0)} - \mathcal{B}_{\text{true}}\|_F^2}_{\text{Algorithmic error}},$$

for all $t \geq 1$ where $\rho = \rho(\alpha, \beta, \kappa) \in (0, 1)$ 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 $1 - \exp(-C_2 \sum_k p_k r_k)$,

$$\xi = \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \nabla \mathcal{L}(\mathcal{B} | \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3), \mathcal{T} \rangle \leq C_1 \sqrt{\gamma_1} \phi U \left(r_1 r_2 r_3 + \sum_{k=1}^{3} p_k r_k\right)^{1/2},$$

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

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

where $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{Y} - b'(\mathcal{B}_{\text{true}} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2 . \boldsymbol{X}_3\})$. Based on Proposition 3, Assumption A2 implies that $\mathcal{E}$ is a sub-Gaussian-(ϕU) tensor. Decompose the side information matrices $\boldsymbol{X}_k$ into $\boldsymbol{X}_k = \boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k^T$, where $\boldsymbol{U}_k \in \mathbb{R}^{d_k \times d_k}$, $\boldsymbol{V}_k \in \mathbb{R}^{p_k \times p_k}$ are singular vectors and $\Sigma_k \in \mathbb{R}^{d_k \times r_k}$ is a diagonal matrix whose entries are singular values for $k = 1, 2, 3$. Notice that

$$\begin{aligned}
\sup_{\substack{\mathcal{T} \in \mathbb{R}^{d_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \mathcal{E} \times_1 \boldsymbol{X}_1, \mathcal{T} \rangle &= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{d_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \boldsymbol{X}_1^T \mathcal{M}_1(\mathcal{E}), \mathcal{M}_1(\mathcal{T}) \rangle \\
&= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{d_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \langle \Sigma_1^T \boldsymbol{U}_1^T \mathcal{M}_1(\mathcal{E}), \boldsymbol{V}_1^T \boldsymbol{M}_1(\mathcal{T}) \rangle \\
&\leq \sup_{\substack{\mathcal{T} \in \mathbb{R}^{d_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F^2 \leq 1}} \sigma_{\max}(\boldsymbol{X}_1) \langle \mathcal{E} \times_1 \boldsymbol{U}_1, \mathcal{T} \rangle,
\end{aligned}$$

where the last inequality uses the fact that the multiplication of orthonormal matrix does change the space

Gaussian width=sup_{parameter in family}<noise, parameter>

of $\mathcal{T}$. Applying this inequality with $k = 1, 2, 3$ to (1) gives us

$$\begin{aligned}
\xi &= \sqrt{\gamma} \sup_{\substack{\mathcal{T}\in\mathbb{R}^{p_1\times p_2\times p_3} \\ \mathrm{rank}(\mathcal{T})\le(r_1,r_2,r_3) \\ \|\mathcal{T}\|_F^2\le 1}} \langle \mathcal{E}\times\{\boldsymbol{U}_1,\boldsymbol{U}_2,\boldsymbol{U}_3\},\mathcal{T}\rangle \\
&\le \sqrt{\gamma} \sup_{\substack{\mathcal{T}\in\mathbb{R}^{p_1\times p_2\times p_3} \\ \mathrm{rank}(\mathcal{T})\le(r_1,r_2,r_3) \\ \|\mathcal{T}\|_F^2\le 1}} \langle \mathcal{E}',\mathcal{T}\rangle,
\end{aligned}$$

where $\mathcal{E}' \stackrel{\text{def}}{=} \mathcal{E}\times\{\boldsymbol{U}_1,\boldsymbol{U}_2,\boldsymbol{U}_3\} \in \mathbb{R}^{p_1\times p_2\times p_3}$ and $\gamma = \prod_{k=1}^3 \sigma_{\max}(\boldsymbol{X}_k)^2$. By orthonormality of $\boldsymbol{U}_k$ for k = 1,2,3, $\mathcal{E}'$ is again a sub-Gaussian-(ϕU) tensor whose entries are independent. Therefore, combination of Lemma and Lemma E.5 in Han et al. [2020] yields

$$\sup_{\substack{\mathcal{T}\in\mathbb{R}^{d_1\times p_2\times p_3} \\ \mathrm{rank}(\mathcal{T})\le(r_1,r_2,r_3) \\ \|\mathcal{T}\|_F^2\le 1}} \langle \mathcal{E}',\mathcal{T}\rangle \le C_1\phi U\left(r_1r_2r_3+\sum_{k=1}^3 p_kr_k\right)^{1/2},$$

with probability at least $1-\exp\left(-C_2\sum_k p_kr_k\right)$ for some constants $C_1, C_2 > 0$.

Finally, we have

$$\xi \le C_1\sqrt{\gamma_1}\phi U\left(r_1r_2r_3+\sum_{k=1}^3 p_kr_k\right)^{1/2}.$$

Applying the upper bound of ξ to Theorem 3.1 in Han et al. [2020] completes the proof with explicit $\rho = 1-\frac{\alpha\beta\eta_0}{1000\kappa^2}$, where $\eta_0 \le \frac{1}{28}$ is a constant. Notice the algorithm requires to have a step size $\eta = \frac{\eta_0\beta}{b^6}$ for the result. $\square$

Lemma 2.1 (Proposition 3.2 in Rivasplata [2012])**.** If X is σ-subgaussian, then for any $q > 0$ one has

$$\mathbb{E}|X|^q = q2^{q/2}\sigma^q\Gamma\left(\frac{q}{2}\right).$$

Consequently, for any $q \ge 1$,

$$\left(\mathbb{E}|X|^q\right)^{1/q} \le C\sigma\sqrt{q}.$$

References

Rungang Han, R. Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *ArXiv*, abs/2002.11255, 2020.

Omar Rivasplata. Subgaussian random variables: An expository note. *Internet publication, PDF*, 2012.