Notes for remaining questions

This note works on remaining questions for IEEE revision. Blue texts are the modifications to the revision.

1. *Give proof/counterexample to the equivalence/ difference among the following three statements?*

   (a) $F(d) > G(d) + c$, for an arbitrary c;

   (b) $F(d) > G(d) + c$, for an arbitrary constant c;

   (c) $F(d) \geq G(d)$.

   Solution: We first consider the case where $c \geq 0$.

   - (a) $\Rightarrow$ (b): *Proof:* Note that $\{$arbitrary constant $c\} \subset \{$arbitrary $c\}$. Therefore, the statement in (a) indicates (b) by taking c as arbitrary constant.
   - (b) $\not\Rightarrow$ (a): *Proof:* Take $F(d) = G(d) + c' + 1$ for an arbitrary constant $c' \geq 0$. The F, G satisfy the statement in (b) but contradict to the statement (a) with function $c = f(d)$ where $f(d) > c' + 1$ for all $d > D$, when $d > D$.
   - (a) (b) $\Rightarrow$ (c): *Proof:* Take $c = 0$. The statements (a), (b) indicate $F(d) > G(d)$, which further indicates $F(d) \geq G(d)$.
   - (c) $\not\Rightarrow$ (a) (b): *Proof:* Let $F(d) = G(d)$. Then, F, G contradict to the statements in (a), (b) with $c = 1$.

   Then, we consider the case where $c < 0$.

   - (a) $\Rightarrow$ (b): *Proof:* Same as the positive case.
   - (b) $\not\Rightarrow$ (a): *Proof:* Same idea as the positive case but take $F(d) = G(d) + c' - 1$ for an arbitrary constant $c' < 0$.
   - (c) $\Rightarrow$ (a) (b): *Proof:* Let $F(d) = G(d)$. Then, F, G satisfy statements in (a), (b).
   - (a) (b) $\not\Rightarrow$ (c): Let $F(d) = G(d) + c/2$ for arbitrary (or arbitrary constant) $c < 0$. Then, F, G satisfy the statement (a) (or (b)) but contradict to the statement (c) since $F(d) < G(d)$.

   In MLE and polynomial-time algorithm achievability, we should write "arbitrary constant ϵ ¿ 0" because we want $p^{\epsilon} \geq \log p$.

2. *Prove that degree assumptions are redundant in all lower bound results. Prove the lower bound Thm in 1st submission implies the impossibility Thm in resubmission, but not vice versa. Which version is stronger?*

We define the estimation space of $(\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}})$ for clarity.

$$E = \{(\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}) : \hat{z} \text{ is a function } [n] \to [r],\ \hat{\theta}(i) > 0, i \in [n]\}.$$

The estimation space E is a more general space than the true parameter spaces $\mathcal{P}(\gamma) \cap \{\text{extra assumptions on } \boldsymbol{\theta}\}$, without any balanced assumption on community size and any assumptions (except positivity) on $\boldsymbol{\theta}$. The estimation space E is unchanged regardless what true parameter space is considered. Even under the true non-degree model with $\boldsymbol{\theta} = \mathbf{1}$, we still consider the estimation space E rather than the subset $E \cap \{\hat{\boldsymbol{\theta}} = \mathbf{1}\}$. Because we consider the dTBM estimator without prior knowledge on true $\boldsymbol{\theta}$. Hence, in dTBM minimax bounds, we take the infimum over $\hat{z}$ such that $(\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}) \in E$.

Next, we consider the impossibility of dTBM under different true parameter spaces.

Lemma 1. Let $A \subset B$ be two true parameter spaces of dTBM. The impossibility of dTBM on A indicates the impossibility of dTBM on B, but not vice versa.

Proof of Lemma 1. For any $\hat{z}$ such that $(\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}) \in E$, by $A \subset B$, we have

$$\sup_{(z,\boldsymbol{\theta},\mathcal{S})\in B} \ell(\hat{z}, z) \geq \sup_{(z,\boldsymbol{\theta},\mathcal{S})\in A} \ell(\hat{z}, z). \tag{1}$$

If we have impossibility on A, we have

$$\inf_{\hat{z},(\hat{z},\hat{\boldsymbol{\theta}},\hat{\mathcal{S}})\in E} \sup_{(z,\boldsymbol{\theta},\mathcal{S})\in B} \ell(\hat{z}, z) \geq \inf_{\hat{z},(\hat{z},\hat{\boldsymbol{\theta}},\hat{\mathcal{S}})\in E} \sup_{(z,\boldsymbol{\theta},\mathcal{S})\in A} \ell(\hat{z}, z) \geq 1/p,$$

which indicates the impossibility on B. On the other hand, impossibility on B does not guarantee the impossibility on A since by inequality (1)

$$\inf_{\hat{z},(\hat{z},\hat{\boldsymbol{\theta}},\hat{\mathcal{S}})\in E} \sup_{(z,\boldsymbol{\theta},\mathcal{S})\in B} \ell(\hat{z}, z) \geq 1/p \quad \not\Rightarrow \quad \inf_{\hat{z},(\hat{z},\hat{\boldsymbol{\theta}},\hat{\mathcal{S}})\in E} \sup_{(z,\boldsymbol{\theta},\mathcal{S})\in A} \ell(\hat{z}, z) \geq 1/p.$$

□

We have following claims related to the revision.

Claim: Degree assumptions are redundant in all impossibility results.

Proof. Consider the true parameter spaces $A = \mathcal{P}(\gamma) \cap \{\text{extra assumptions on } \boldsymbol{\theta}\}$ and $B = \mathcal{P}(\gamma)$. Note that $A \subset B$. Therefore, by Lemma 1, the impossibility on A indicates the impossibility on B, and thus the assumptions on $\boldsymbol{\theta}$ is redundant for impossibility results. □

We can remove all the extra assumptions on $\boldsymbol{\theta}$ in the impossibility results.

Claim: The statistical impossibility in the original submission does not indicate the impossibility in resubmission, and vice versa.

(The computational impossibility is the same in two submissions.)

Proof. Note that $\Delta_{\min}^2 \asymp \Delta_{\boldsymbol{X}}^2$ under the balanced degree $\boldsymbol{\theta}$. We have

$$\{\Delta_{\min}^2 \asymp p^{-(K-1)}, \boldsymbol{\theta} \text{ is balanced}\} \subset \{\Delta_{\boldsymbol{X}}^2 \asymp p^{-(K-1)}\}.$$

Then, we consider two true parameter spaces corresponding to the original and second submission, respectively,

$$\mathcal{P}_1 = \mathcal{P} \cap \{\Delta_{\min}^2 \asymp p^{-(K-1)}\}, \quad \mathcal{P}_2 = \mathcal{P} \cap \{\Delta_{\boldsymbol{X}}^2 \asymp p^{-(K-1)}\}.$$

Here, we omit some common constraints in two submissions including $r \lesssim p^{1/3}$ and $K \geq 1$.

There is no containment relationship between $\mathcal{P}_1$ and $\mathcal{P}_2$. Specifically,

- $\mathcal{P}_1 \not\subset \mathcal{P}_2$: Counter-example: Take $a = p^{-1/2}, M = p^{-1}$ in the Example 3 of revised manuscript. We have $\Delta_{\min}^2 = p^{-1}$ and $\Delta_{\boldsymbol{X}}^2 = 1$. With proper $z, \boldsymbol{\theta}$, we have $(z, \boldsymbol{\theta}, \mathcal{S}) \in \mathcal{P}_1$ but $(z, \boldsymbol{\theta}, \mathcal{S}) \notin \mathcal{P}_2$.
- $\mathcal{P}_2 \not\subset \mathcal{P}_1$: Counter-example: Take $a = p^{-1/4}, M = p^{1/2}$ in the Example 3 of revised manuscript. We have $\Delta_{\min}^2 = p^{-1/2}$ and $\Delta_{\boldsymbol{X}}^2 = p^{-1}$. With proper $z, \boldsymbol{\theta}$, we have $(z, \boldsymbol{\theta}, \mathcal{S}) \in \mathcal{P}_2$ but $(z, \boldsymbol{\theta}, \mathcal{S}) \notin \mathcal{P}_1$.

To compare the impossibility, we need to consider following two minimax rates.

$$\inf_{\hat{z}, (\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}) \in E} \sup_{(z, \boldsymbol{\theta}, \mathcal{S}) \in \mathcal{P}_1} \ell(\hat{z}, z) \quad \text{and} \quad \inf_{\hat{z}, (\hat{z}, \hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}) \in E} \sup_{(z, \boldsymbol{\theta}, \mathcal{S}) \in \mathcal{P}_2} \ell(\hat{z}, z),$$

Since no containment relationship exists between $\mathcal{P}_1$ and $\mathcal{P}_2$, we need to exactly solve the minimizer $\hat{z}$ over E and the maximizers $(z, \boldsymbol{\theta}, \mathcal{S})$ over $\mathcal{P}_1, \mathcal{P}_2$ in the two rates. By Neyman-Pearson Lemma, the profile MLE is the optimal $\hat{z}$ in E. However, we do not know the maximizer of the supremum. Therefore, there is no enough information to infer the relationship between two minimax rates. The impossibility of dTBM on $\mathcal{P}_1$ does not indicates the impossibility of dTBM on $\mathcal{P}_2$, and vice versa.

□

Remark 1. Both impossibilities on $\mathcal{P}_1$ and $\mathcal{P}_2$ are true, though we can not infer the impossibility on one space from the other one. Based on current proof, we can show the impossibility of dTBM on the space $\mathcal{P}_0 = \{\Delta_{\min}^2 \asymp p^{-(K-1)}, \boldsymbol{\theta} \text{ is balanced}\}$, which satisfies $\mathcal{P}_0 \subset \mathcal{P}_1$ and $\mathcal{P}_0 \subset \mathcal{P}_2$. Hence, by Lemma 1, we have the impossibilities on $\mathcal{P}_1$ and $\mathcal{P}_2$.

We can combine the results in original and second submission for the statistical impossibility; i.e., "when $\Delta_{\min}^2 \lesssim p^{-(K-1)}$ or $\Delta_{\boldsymbol{X}}^2 \lesssim p^{-(K-1)}$, the estimator obeys...".

3. *Rewrite the motivating example on mixture Gaussian tensor models.*

Example 1 (Weighted tensor Gaussian mixture model)**.** Gaussian mixture model is widely applied in applications including pattern recognition, image processing, and machine learning. We say an order-K sample tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ follows a tensor Gaussian distribution with mean tensor $\mathcal{S} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, denoted as $\mathcal{X} \sim \mathcal{N}(\mathcal{S}; \boldsymbol{I}_{p_1}, \ldots, \boldsymbol{I}_{p_K})$, if $\text{vec}(\mathcal{Y}) \sim \mathcal{N}(\text{vec}(\mathcal{S}), \boldsymbol{I}_{p_1 \cdots p_K})$, where $\boldsymbol{I}_p$ is the identity matrix of dimension p.

Weighted tensor Gaussian mixture model assumes the tensor data $\mathcal{Y}_i \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ for all $i \in [N]$ follow the mixture tensor Gaussian distribution with density

$$f(\mathcal{Y}_i) = \sum_{a=1}^{r} \pi_a \phi_a(\theta_i \mathcal{S}_a; \boldsymbol{I}_{p_1}, \ldots, \boldsymbol{I}_{p_K}), \quad i \in [N], \tag{2}$$

where θ_i is the weight of the i-th observation, r is the number of clusters, π_a is the mixture weight, $\mathcal{S}_a$ is the mean tensor and ϕ_a denotes the tensor Gaussian density of the a-th cluster, respectively, for all $r \in [a]$.

Let $\mathcal{Y} \in \mathbb{R}^{N \times p_1 \times \cdots \times p_K}$ denote the combined data with slices $\mathcal{Y}(i,:) = \mathcal{Y}_i$ and $\mathcal{S} \in \mathbb{R}^{r \times p_1 \times \cdots \times p_K}$ denote the combined mean tensor with slices $\mathcal{S}(a,:) = \mathcal{S}_a$. Consider the dTBM model

$$\mathbb{E}[\mathcal{Y}] = \mathcal{S} \times_1 \boldsymbol{\Theta}\boldsymbol{M} \times_2 \boldsymbol{I}_{p_1} \times_3 \cdots \times_{K+1} \boldsymbol{I}_{p_K}, \tag{3}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \cdots, \theta_N)$ is the weight matrix and $\boldsymbol{M} \in \{0,1\}^{N \times r}$ is the membership matrix for N samples. Then, the weighted tensor Gaussian mixture model (2) with prior probability $\pi_a = |\{i \in [N] : \boldsymbol{M}_{i,a} = 1\}|/N$ for all $a \in [r]$ is equal to the dTBM model (3).

I feel such tensor mixture may not be a good motivating example. Because Gaussian mixture essentially tackles the 1-dimensional clustering rather than multiway clustering on multiple modes.

4. *What is the between "arbitrarily worse" vs "worse". Rethink your example.*

   - "Arbitrarily worse": the (statistical or computational) critical values without balanced assumption can be *arbitrarily larger* than that with balanced assumption; the "arbitrarily large" means "as large as possible", which indicates $\Delta_{\min}^2 = \mathcal{O}(1)$ in our context.
   - "worse": the critical values without balanced assumption can be *larger* than that with balanced assumption.

   Example 3 in revised manuscript only indicates that the phase transition of dTBM with unbalanced degree is *worse* than that with balanced degree. Because Example 3 only shows the statistical critical value increases to $p^{-1/2}$, rather than arbitrarily large to $\mathcal{O}(1)$.

   Now, let $a = 1$ and $M = p$ in Example 3. We have $\Delta_{\min}^2 = 1$ and $\Delta_{\boldsymbol{X}}^2 < 2p^{-1}$, with which dTBM is statistically impossible by revised Theorem 1. Then, the new example indicates the statistical critical value in dTBM with unbalanced degree increases to $\mathcal{O}(1)$. Therefore, we say the phase transition of dTBM with respect to $\Delta_{\min}^2$ is arbitrarily worse than that with balanced assumption.

   Current explanation that discusses the relative magnitude of m is better in my opinion.

   Remark 2. Example 3 and the statistical impossibility in resubmission indicate that our original Theorem 1 is not sharp; i.e., when $\gamma > -(K-1)$, dTBM still can be statistically impossible. The sharpness of threshold $-(K-1)$ comes from the achievability side with balanced assumption. Similarly, our original Theorem 2 is also not sharp; i.e., when $\gamma > -K/2$, dTBM still can be computationally inefficient. The sharpness of threshold $-K/2$ also comes from the achievability of the polynomial algorithm with extra constraints on $\boldsymbol{\theta}$.

5. *Sharpness of MLE bounds on $\hat{\theta}$, $\hat{\mathcal{S}}$, $\hat{\mathcal{X}}$? Conjecture, intuition, heuristics.*

Conjecture 1 (MLE sharpness)**.** Let $\hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}$ and $\hat{\mathcal{X}}$ denote the MLE of dTBM, and $\boldsymbol{\theta}, \mathcal{S}, \mathcal{X}$ denote the true parameters. With high probability, we have

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \lesssim \frac{p-r}{p^K}\sigma^2, \quad \|\hat{\mathcal{S}} - \mathcal{S}\|_F^2 \lesssim \frac{r^K}{p^K}\sigma^2, \quad \|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \lesssim \frac{pr + r^K}{p^K}\sigma^2.$$

Remark 3. Above conjecture follows the heuristics

$$\|\hat{\beta} - \beta\|_F^2 \lesssim \frac{\#\text{ degree of freedom of } \beta}{\#\text{ sample size}},$$

where β is the parameter of interest and sample size refers to the number of sample involved in the estimator $\hat{\beta}$. Specifically, in our context, p^K contributes to the estimations of $\hat{\boldsymbol{\theta}}, \hat{\mathcal{S}}, \hat{\mathcal{X}}$. The degree of freedom in $\boldsymbol{\theta}$ is $p - r$ where $-r$ comes from the constraint that $\|\boldsymbol{\theta}_{z^{-1}(a)}\| = |z^{-1}(a)|$ for all $a \in [r]$; the degree of freedom of $\mathcal{S}$ is r^K; and the degree of freedom in $\mathcal{X}$ is $pr + r^K$ where pr comes from the factor matrix $\boldsymbol{\Theta}\boldsymbol{M}$ and r^K comes from the core tensor.

6. *Other technical flaws you detected by yourself.*

- By Question 5, the estimation error of $\hat{\boldsymbol{\theta}}$ in Lemma 12 can be sharpened, and consequently the condition $K \geq 3$ in MLE achievability may be removed.
- A related problem with Question 4. In Example 3, I only show the phase transition will be arbitrarily worse when M achieves the extreme values $\mathcal{O}(p)$. This example does not support the necessity of the constant 1 in the balanced assumption; i.e., when $M = c$ for some positive constant $c \neq 1$, Example 3 can not illustrate the arbitrarily worse phase transition w.r.t. $\Delta_{\min}^2$. However, I have no idea on how to show the necessity of constant 1.

7. *Conditions for each theorem/lemma*

Here we summarize the parameter conditions for all the theortical results in the main text and the supplement.

Results	high prob vs deterministic	finite p vs $p \to \infty$	balance degree	lower bounded degree	$r \geq$?	$K \geq$?
Thm 1	deterministic	finite $p \geq 2$	×	×	$1 \leq r \leq p$	$K \geq 2$
Lem 1	deterministic	$p \to \infty$	√	√	$2 \leq r \leq p$	$K \geq 1$
Thm 2 (Imp)	deterministic	$p \to \infty$	×	×	$2 \leq r \leq \mathcal{O}(p^{1/3})$	$K \geq 1$
Thm 2 (Ach)	high prob	$p \to \infty$	√	√	fixed $r \geq 1$	$K \geq 3$
Thm 3 (Imp)	deterministic	$p \to \infty$	×	×	$2 \leq r \leq p$	$K \geq 2$
Lem 2	deterministic	finite $p \geq 1$	×	√	$1 \leq r \leq p$	$K \geq 1$
Thm 4	high prob	$p \to \infty$	×(√)	√	fixed $r \geq 1$	$K \geq 2$
Thm 5	high prob	$p \to \infty$	×	√	fixed $r \geq 1$	$K \geq 2$
Cor 1 (Thm 3 Ach)	high prob	$p \to \infty$	√	√	fixed $r \geq 1$	$K \geq 2$
Prop 1	high prob	$p \to \infty$	√	√	fixed $r \geq 1$	$K \geq 2$

Table 1: Conditions for main Theorems in the main text.

Remark 4 (Reflection)**.** For the results valid under $1 \leq r \leq p$ and $K \geq 1$, we omit the conditions on r, K in the statement.

Remark 5 (When we need $r \geq 2$)**.** By our convention in Assumption 1, $\Delta_{\min}^2 = 1$ when $r = 1$. Therefore, in impossibility results (Thm 2 (Imp), Thm 3 (Imp)) we need $r \geq 2$ to let the condition $\Delta_{\min}^2 \leq p^\gamma$ feasible. For Lem 1, we need $z(i) \neq z(j)$, which also indicates we need $r \geq 2$.

Remark 6 (When we allow $K = 1$)**.** Our dTBM with $K = 1$ does not directly reduce to the vector clustering. The core tensor reduces to a vector $\boldsymbol{s} \in \mathbb{R}^r$ and the angle gap between the elements in $\boldsymbol{s}$ (scalars) is always 0, which leads to the non-identifiable for the dTBM. The statistical impossibility results Thm 2 still holds since $K \geq 1$ includes a larger range than $K \geq 2$. Lem 1 and Lem 2 hold since both sides in the inequalities are 0. The algorithm will produce random results under the case $K = 1$ due to the non-identifiability. Hence, we exclude the $K = 1$ for algorithm results (Thm 4, Thm 5, Cor 1, Prop 1).

Results	high prob vs deterministic	finite p vs $p \to \infty$	balance degree	lower bounded degree	$r \geq$?	$K \geq$?
Lem 3	deterministic	finite $p \geq 2$	$\times$	$\times$	$2 \leq r \leq p$	$K \geq 2$
Lem 4	deterministic	finite dimension	-	-	-	-
Lem 5	deterministic	finite dimension	-	-	-	-
Lem 6	deterministic	finite $p \geq 1$	$\times$	$\surd$	$1 \leq r \leq p$	$K \geq 1$
Lem 7	high prob	$p \to \infty$	$\times$	$\times$	$1 \leq r \leq p$	$K \geq 1$
Lem 8	deterministic	$p \to \infty$	$\surd$	$\times$	$2 \leq r \leq p$	$K \geq 1 (d \geq 1)$
Lem 9	deterministic	finite $p \geq 1$	$\times$ ($\surd$)	$\surd$	$1 \leq r \leq p$	$K \geq 1$
Lem 10	deterministic	$p \to \infty$	$\times$ ($\surd$)	$\surd$	fixed $r \geq 2$	$K \geq 2$
Lem 11	deterministic	$p \to \infty$	$\times$ ($\surd$)	$\surd$	fixed $r \geq 2$	$K \geq 2$
Lem 12	high prob	$p \to \infty$	$\surd$	$\surd$	fixed $r \geq 1$	$K \geq 2$

Table 2: Conditions for Lemmas in the Supplement. The $\times(\surd)$ indicates that only part (MLE related proof) of the proof requires the balance assumption.

Remark 7 (Stability condition)**.** The local stability condition required by Theorem 5 has been updated as the arXiv version.

8. *Explanation for the slide bounds c_3 and c_4.*

   Previous explanations about the bounds c_3 and c_4 are claimed to avoid the purely zero slides and unbounded entries. This is not correct, since the slides has p^{K-1} entries. The bound c_3 avoids the degenerate slides with norm tending to 0; while bound c_4 avoids the inflating slides with norm tending to ∞ as $p \to \infty$.

9. *Explanation for necessity of balance assumption.*

   Previous explanation above Example 3 aims to point out the intrinsic necessity of the nuisance parameter constraints to establish the phase transition. However, Example 3 only illustrates the necessity of balance assumption for the impossibility result. We have no achievability result without the balance assumption. Therefore, we argue the helpfulness of the balance assumption in our work rather than the necessity. It may not be necessary to have constraints for all problems with nuisance parameters.

10. *Similar assumptions in literature.*

    Previous Remark 3 states that our balance assumption can be relaxed to the same condition in Ke et al. (2019) with stricter constraint on core tensor. This is not true. Though we restrict

magnitude of core tensor, we still need balance assumption to guarantee the angular signal in core tensor can be preserved in mean tensor.

11. *Gaussian mixture example.*

    I feel the Gaussian mixture example is not very appropriate. The traditional Gaussian mixture actually tackles the 1-dimensional clustering (though the observations can be matrices or tensors), rather than the multiway clustering on different modes.

12. *Sharpen the MLE result to relax the $K \geq 3$ condition in MLE achievability. Unsolved.*

13. *Sparsity as future direction.*

    The diagonal-deletion technique is also applied for binary data clustering, motivated by the bias-variance trade-off. The sparse tensor decomposition/clustering may be a good future direction.

14. *Statistical impossibility.*

    We need the statistical impossibility result with $\Delta^2_{\boldsymbol{X}}$. We use such result to illustrate the insufficiency of $\Delta^2_{\min}$ without balance assumption in Example 3.

References

Ke, Z. T., Shi, F., and Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.