

Seeded Algorithm

Jiaxin Hu

March 31, 2022

This note combines previous results of error control and the non-iterative clean up. Theorem 1 is the theoretical guarantee for the complete seeded matching Algorithm 1.

Notations.

- $I^m = \{(i_1, \dots, i_m) : i_k \in I, k \in [m]\}$: the order m product space of index set I ;
- $\mathcal{A} \in \mathbb{R}^{n^m}$: an order m tensor with dimension n on each mode and all entries in \mathbb{R} ;
- $\mathcal{S} = \{(i, k) \in [n]^2 : a_i, b_k \geq \xi, d_p(\mu_i, \nu_k) \leq \zeta\}$: the set of seeds and $s = |\mathcal{S}|$;
- $S = \{i \in [n] : (i, k) \in \mathcal{S} \text{ for some } k \in [n]\}$: the set of the first coordinate in the seed \mathcal{S} ;
- $T = \{k \in [n] : (i, k) \in \mathcal{S} \text{ for some } i \in [n]\}$: the set of the second coordinate in the seed \mathcal{S} ;
- $\pi_0 : S \mapsto T$: permutation corresponding to the seed \mathcal{S} satisfying $(i, \pi_0(i)) \in \mathcal{S}$ for all $i \in S$.

The Sub-Algorithm 1 is equivalent to do the one-way matching for the rows between $\text{Mat}_1(\mathcal{A}'), \text{Mat}_1(\mathcal{B}') \in \mathbb{R}^{n-s \times s^{m-1}}$, where

\mathcal{A}' contains $\{\mathcal{A}_{i,\omega} : i \in [n]/S, \omega \in S^{m-1}\}$, and \mathcal{B}' contains $\{\mathcal{B}_{k,\omega} : k \in [n]/T, \omega \in T^{m-1}\}$,

with $S^{m-1} = \{(i_2, \dots, i_m) : i_l \in S, l = 2, \dots, m\}$ and $T^{m-1} = \{(k_2, \dots, k_m) : k_l \in T, l = 2, \dots, m\}$.

Theorem 1 (Guarantee for Algorithm 1). *Let $\rho = \sqrt{1 - \sigma^2}$ and $s_0 = C(\log n^{1/4} + 1)^{1/(m-1)}$. Suppose $\sigma \leq c/s_0^{1/3}$ for sufficiently small constant c . Choose thresholds $\xi \geq c_1\sqrt{s_0}$ with universal positive constant c_1 and $\zeta \leq \sqrt{\sigma/n^{m-1}}$. Algorithm 1 recover the true permutation π^* with probability tends to 1.*

Remark 1 (Compared with previous results.). Compared with the result in note 0306_22-proof, we relax the conditions from $\sigma < c/\log^{1/3(m-1)} n$ to $\sigma \leq c/\log^{1/3(m-1)} n^{1/4}$. The exponent $1/4$ over n comes from the choice of $r_0 = \mathcal{O}(n^{3/4})$ in Theorem 2. The range of r_0 is determined by Theorem 3, however, there may exist some problem in Theorem 3. See the Fixme below.

Proof of Theorem 1. Based on Theorem 2 and Theorem 3, the output $\hat{\pi}$ of Sub-Algorithm 1 and Sub-Algorithm 2 fully recovers the true permutation if the number of seeds s satisfying $s^{m-1} \gtrsim \log n^{1/4} + 1$ and we take $r_0 = \mathcal{O}(n^{3/4})$.

Algorithm 1 Gaussian tensor matching with seed improvement

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^m}$, threshold ξ, ζ .

- 1: Calculate the distance statistics $d_p(\mu_i, \nu_k)$ for each pair of $(i, k) \in [n]^2$.
- 2: Obtain the high-degree set $\mathcal{S} = \{(i, k) \in [n]^2 : a_i, b_k \geq \xi, d_p(\mu_i, \nu_k) \leq \zeta\}$, where $a_i = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega \in [n]^{m-1}} \mathcal{A}_{i, \omega}$, $b_k = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega \in [n]^{m-1}} \mathcal{B}_{k, \omega}$.
- 3: **if** there exists a permutation $\pi_0 : S \mapsto T$ such that $\mathcal{S} = \{(i, \pi_0(i)) : i \in S, \pi(i) \in T\}$ **then**
- 4: Run Sub-Algorithm 1 with seed π_0 and obtain output π_1 .
- 5: Run Sub-Algorithm 2 with π_1 and obtain output $\hat{\pi}$.
- 6: **else**
- 7: Output error.
- 8: **end if**

Output: Estimated permutations $\hat{\pi}$ or error.

Sub-Algorithm 1: Seeded matching

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^m}$, seed $\pi_0 : S \mapsto T$.

- 9: Let S^c denote the complement $[n]/S$, and T^c denote $[n]/T$. Obtain the similarity matrix $H = \llbracket H_{ik} \rrbracket \in \mathbb{R}^{(n-s) \times (n-s)}$ where $H_{ik} = \sum_{\omega \in S^{m-1}} \mathcal{A}_{i, \omega} \mathcal{B}_{k, \pi_0(\omega)}$ for any $i \in S^c$ and $k \in T^c$.
- 10: Find the optimal bipartite permutation $\tilde{\pi}_1$ such that

$$\tilde{\pi}_1 = \arg \max_{\pi : S^c \mapsto T^c} \sum_{i \in S^c} H_{i, \pi(i)}.$$

Let π_1 denote the matching on $[n]$ such that $\pi_1|_S = \pi_0$ and $\pi_1|_{S^c} = \tilde{\pi}_1$.

Output: Estimated permutations $\hat{\pi}_1$.

Sub-Algorithm 2: Non-iterative clean-up

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^m}$ and permutation $\pi_1 : [n] \mapsto [n]$.

- 11: For each pair $(i, k) \in [n]^2$, calculate $W_{ik} = \sum_{\omega \in [n]^{m-1}} \mathcal{A}_{i, \omega} \mathcal{B}_{k, \pi_1(\omega)}$.
- 12: Sort $\{W_{ik} : (i, k) \in [n]^2\}$ and let \hat{S} denote the set of indices of largest n elements.
- 13: **if** there exists a permutation $\hat{\pi}$ such that $\hat{S} = \{(i, \hat{\pi}(i)) : i \in [n]\}$ **then**
- 14: Output $\hat{\pi}$.
- 15: **else**
- 16: Output error.
- 17: **end if**

Output: Estimated permutations $\hat{\pi}$ or error.

[FIXME (Jiaxin): The Theorem 3 indicates we can choose $r_0 = \mathcal{O}(n - \sqrt{n} \log^{1/2(m-1)} n)$. However, if we take $r_0 \asymp n - n^{1/2+\epsilon}$, for some small $\epsilon \in (0, 1/2)$, we will have $s_0^{m-1} \gtrsim \log \left(\frac{1}{n^{1/2-\epsilon+1}} + 1 \right) \asymp \log n^{\epsilon-1/2}$. Then by Theorem 1, when n goes larger, we need fewer seeds and thus has a looser upper bound for σ , which is counter-intuitive and contradicts to the result in Ding et al. (2021). Therefore, I guess there may be some problems I did not recognize in Theorem 3.]

Hence, we only need to show the set

$$\mathcal{S} = \{(i, k) \in [n]^2 : a_i, b_k \geq \xi, d_p(\mu_i, \nu_k) \leq \zeta\},$$

where

$$a_i = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega \in [n]^{m-1}} \mathcal{A}_{i,\omega}, \quad b_k = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega \in [n]^{m-1}} \mathcal{B}_{k,\omega},$$

with proper thresholds ξ and ζ has enough true pairs and no fake pairs with high probability.

Note that for fake pair $(i, k) \in [n]^2$, i.e, $i \neq \pi^*(k)$, we have

$$\mathbb{P}(a_i \geq \xi, b_k \geq \xi) = \mathbb{P}(a_i \geq \xi) \mathbb{P}(b_k \geq \xi) = Q^2(\xi),$$

where Q is the complementary CDF of normal distribution. For true pair $(i, k) \in [n]^2$, i.e, $i = \pi^*(k)$, we have

$$\begin{aligned} \mathbb{P}(a_i \geq \xi, b_k \geq \xi) &= \mathbb{P}(a_i \geq \xi, \sqrt{1 - \sigma^2} a_i + \sigma z_i \geq \xi) \\ &\geq \mathbb{P}(a_i \geq \xi / \sqrt{1 - \sigma^2}, z_i \geq 0) \\ &\geq \frac{1}{2} Q(\xi / \sqrt{1 - \sigma^2}) \\ &\geq Q(\xi) \exp(-C \sigma^2 \xi^2), \end{aligned}$$

where C is a positive constant.

Take $\zeta \leq \sqrt{\sigma/n^{m-1}}$. To let \mathcal{S} satisfy the conditions for in Theorem 2, we need

1. \mathcal{S} has s true pairs with high probability (the expectation of the true pairs in \mathcal{S} is larger than s)

$$nQ(\xi) \exp(-C \sigma^2 \xi^2) \geq s; \tag{1}$$

2. \mathcal{S} has no fake pairs (the expectation of the fake pairs in \mathcal{S} converges to 0 as $n \rightarrow \infty$)

$$n^2 Q^2(\xi) C_2 \exp\left(-\frac{1}{\sigma}\right) = o(1). \tag{2}$$

Take $\xi \geq c_1 \sqrt{s}$. By inequality (1), we have $Q(\xi) \geq \frac{s}{n} \exp(C c_1^2 \sigma^2 s)$. Plugging the inequality for $Q(\xi)$ into the inequality (2), we have

$$C_2 s^2 \exp\left(2C c_1^2 \sigma^2 s - \frac{1}{\sigma}\right) = o(1),$$

which implies $\sigma \leq \frac{c}{s^{1/3}}$ with small constant c such that $2C c_1^2 c^2 - \frac{1}{c^2} < 0$.

□

Useful Theorems and Lemmas for the proof of Theorem 1

Theorem 2 (Guarantee for Sub-Algorithm 1). *Suppose the seed π_0 corresponds to s true pairs and no fake pairs. Assume $s^{m-1} \gtrsim \log n - \log(r_0 + 1) + 1$. The output π_1 of seeded matching Sub-Algorithm 1 has at most r_0 errors for $r_0 \in \mathbb{N} \cap [0, n - s]$.*

Proof of Theorem 2. See note 0323_22_seeded.

□

Theorem 3 (Guarantee for Sub-Algorithm 2). *Suppose the input permutation π_1 has at most r fake pairs such that $(n - r)^{(m-1)/2} \gtrsim n^{(m-1)/4} \log^{1/4} n + \log^{1/2} n$. Then, the output of non-iterative clean up Sub-Algorithm 2 is equal to the true permutation with a high probability; i.e., $\hat{\pi} = \pi^*$ with a high probability as $n \rightarrow \infty$.*

Proof of Theorem 3. See note 0321.22_cleanup. □

Lemma 1 (Tail bounds for the product of normal variables). *Consider the correlated pairs of normal variables (X_i, Y_i) for $i \in [n]$, where $X_i, Y_i \sim N(0, 1)$. Let $M = \frac{1}{n} \sum_{i \in [n]} X_i Y_i$. If $\text{cov}(X_i, Y_i) = \rho > 0$, then we have*

$$\mathbb{P}(|M - \rho| \geq t) \leq 4 \exp \left(- \min \left\{ \frac{1}{32\rho^2}, \frac{1}{16(1 - \rho^2)} \right\} nt^2 \right) \leq 4 \exp \left(- \frac{nt^2}{32} \right),$$

for constant $t \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1 - \rho^2}\}]$. If $\text{cov}(X_i, Y_i) = 0$, then, we have

$$\mathbb{P}(|M| \geq t) \leq 2 \exp \left(- \frac{nt^2}{4} \right),$$

for constant $t \in [0, \sqrt{2}]$.

Proof of Lemma 1. See note 0306.22_proof. □

References

Ding, J., Ma, Z., Wu, Y., and Xu, J. (2021). Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115.