Error control of seeded matching

Jiaxin Hu

March 24, 2022

Previous note 0306_proof investigates the seed condition for the π_1 to fully recover the true permutation π^*. Note that 0321_clean_up indicates we can achieve fully recovery via a non-iterative clean up of π_1 with controlled error. Therefore, this note aims to investigate the seed condition for π_1 with controlled error. The theorem indicates that the seed condition can be more relaxed when we allow more error in π_1. More details about the constant and extreme cases should be considered in the proof, though I believe the general proof idea makes sense.

To do list:

- Figure out the proof details for the extreme cases and constants.
- Combine this error control result with the clean up result.
- Proof of Conjecture 1.

For self-consistency, we write the seeded algorithm without the non-iterative clean up procedure as the separate Algorithm 1 below.

Algorithm 1 Seeded matching

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

1: For $i \in S^c$ and $k \in T^c$, obtain the similarity matrix $H = [\![H_{ik}]\!]$ as

$$H_{ik} = \sum_{\omega \in S^{m-1}} \mathcal{A}_{i,\omega} \mathcal{B}_{k,\pi_0(\omega)}.$$

2: Find the optimal bipartite permutation $\tilde{\pi}_1$ such that

$$\tilde{\pi}_1 = \underset{\pi: S^c \mapsto T^c}{\arg\max} \sum_{i \in S^c} H_{i,\pi(i)}. \tag{1}$$

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 $\hat{\pi}_1$.

Theorem 0.1 (Error control of seeded matching)**.** *Suppose the seed π_0 corresponds to s true pairs and no fake pairs, where $s^{m-1} \gtrsim \log n - \log r_0/n$ and r_0 satisfies $r_0 \log n - r_0 \log r_0/n \gtrsim 1$. The output π_1 of seeded matching Algorithm 1 has at most r_0 errors.*

Remark 1. Note that the condition for the number of seeds s can be relaxed from $\log^{1/(m-1)} n$ to $(\log n - \log r_0/n)^{1/(m-1)}$ when we allow there are are r_0 errors in π_1. Previous theorem in 0306_proof investigates the seed condition for π_1 to fully recover π^*. So, the relaxation of s is intuitive when we ask π_1 has a controlled error. More details in the proof should be improved. For example, when $r_0 = 0$, the Theorem 0.1 now is meaningless. I will figure out this issue in next step.

Proof of Theorem 0.1. Without loss of generality, we assume the true permutation π^* is the identity mapping.

To show the π_1 has at most r_0 errors, it suffices to the permutation on S^c with errors more than r_0 can not be picked by (1) with probability tends to 1 as $n \to \infty$; i.e., with high probability

$$\sum_{i \in S^c} H_{ii} > \max_{r \geq r_0} \max_{\pi \in \Pi_r} \sum_{i \in S^c} H_{i\pi(i)},$$

where Π_r is the collection of all the permutations on $S^c \mapsto T^c$ has r errors.

Consider an arbitrary $\pi \in \Pi_r$ where $r \geq r_0$. Let the $R = \{i \in S^c : \pi(i) \neq i\}$ denote the set of errors in π with $|R| = r$. Then, the probability

$$\begin{aligned}
\mathbb{P}\left(\sum_{i \in S^c} H_{ii} - \sum_{i \in S^c} H_{i\pi(i)} < t\right) &= \mathbb{P}\left(\sum_{i \in R} H_{ii} - \sum_{i \in R} H_{i\pi(i)} < t\right) \\
&= \mathbb{P}\left(\frac{1}{rs^{m-1}} \sum_{i \in R} H_{ii} - \frac{1}{rs^{m-1}} H_{i\pi(i)} < \frac{t}{rs^{m-1}}\right) \\
&\leq \mathbb{P}\left(\frac{1}{rs^{m-1}} \sum_{i \in R} H_{ii} \leq \frac{t+t'}{rs^{m-1}}\right) + \mathbb{P}\left(\frac{1}{rs^{m-1}} H_{i\pi(i)} > \frac{t'}{rs^{m-1}}\right).
\end{aligned}$$

By Lemma 1, we have

$$\mathbb{P}\left(\frac{1}{rs^{m-1}} \sum_{i \in R} H_{ii} \leq \frac{t+t'}{rs^{m-1}}\right) \leq 2\exp\left(-\frac{rs^{m-1}}{32}\left(\rho - \frac{t+t'}{rs^{m-1}}\right)^2\right),$$

for $\rho - \frac{t+t'}{rs^{m-1}} \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$ and

$$\mathbb{P}\left(\frac{1}{rs^{m-1}} H_{i\pi(i)} > \frac{t'}{rs^{m-1}}\right) \leq \exp\left(-\frac{(t')^2}{4rs^{m-1}}\right),$$

for $\frac{t'}{rs^{m-1}} \in [0, \sqrt{2}]$. Take $t = t' = \frac{\rho}{4} rs^{m-1}$. We have

$$\mathbb{P}\left(\sum_{i \in S^c} H_{ii} - \sum_{i \in S^c} H_{i\pi(i)} < \frac{\rho}{4} rs^{m-1}\right) \leq 4\exp\left(-\frac{1}{128} rs^{m-1} \rho^2\right).$$

Hence, noted that $|\Pi_r| = \binom{n}{r} \leq \frac{n^r}{r}$, we have

$$\begin{aligned}
\mathbb{P}\left(\min_{r\geq r_0}\min_{\pi\in\Pi_r}\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)} < \frac{\rho}{4} r_0 s^{m-1}\right) &\leq \sum_{r\geq r_0}\frac{n^r}{r}\mathbb{P}\left(\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)} < \frac{\rho}{4} r_0 s^{m-1}\right) \\
&\leq \sum_{r\geq r_0}\frac{n^r}{r}\mathbb{P}\left(\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)} < \frac{\rho}{4} r s^{m-1}\right) \\
&\leq 4\sum_{r\geq r_0}\frac{n^r}{r}\exp\left(-\frac{1}{128} r s^{m-1}\rho^2\right).
\end{aligned}$$

Based on the assumption that $s^{m-1} \geq 256(\log n - \log r_0/n)$, we know that for all $r \geq r_0$

$$\frac{n^r}{r}\exp\left(-\frac{1}{128} r s^{m-1}\rho^2\right) \leq \exp\left(-\frac{1}{256} r s^{m-1}\rho^2\right).$$

Thus, by the sum of proportional sequence, we have

$$\begin{aligned}
\mathbb{P}\left(\min_{r\geq r_0}\min_{\pi\in\Pi_r}\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)} < \frac{\rho}{4} r_0 s^{m-1}\right) &\leq 4\frac{\exp\left(-\frac{1}{256} r_0 s^{m-1}\rho^2\right)}{1-\exp\left(-\frac{1}{256} s^{m-1}\rho^2\right)} \\
&\leq 4\exp\left(-\frac{1}{256} r_0 s^{m-1}\rho^2\right),
\end{aligned}$$

which tends to 0 when $r_0 \geq s^{-(m-1)}$, which indicates r_0 should satisfy $r_0 \log n - r_0 \log r_0/n \geq 256$.

Therefore, when r_0 satisfies $r_0 \log n - r_0 \log r_0/n \gtrsim 1$ and $s^{m-1} \geq 256(\log n - \log r_0/n)$, we have

$$\mathbb{P}\left(\min_{r\geq r_0}\min_{\pi\in\Pi_r}\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)} \geq \frac{\rho}{4} r_0 s^{m-1}\right) \to 1,$$

which implies the event $\sum_{i\in S^c} H_{ii} > \max_{r\geq r_0}\max_{\pi\in\Pi_r}\sum_{i\in S^c} H_{i\pi(i)}$ holds with probability tends to 1.

□

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 $H = \frac{1}{n}\sum_{i\in[n]} X_i Y_i$. If cov$(X_i, Y_i) = \rho > 0$, then we have*

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

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

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

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

References