Error control of seeded matching

Jiaxin Hu

March 25, 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.

To do list:

- Figure out the proof details for the extreme cases and constants. (done)
- 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. Assume $s^{m-1} \gtrsim \log n - \max\{\log r_0, 0\} + 1$. The output π_1 of seeded matching Algorithm 1 has at most r_0 errors for $r_0 \in \mathbb{N} \cap [0, n]$.*

Remark 1. Note that the constant 1 in the condition for s^{m-1} can be replaced by any small positive constant $\epsilon \in [0,1]$ as long as the $r_0 s^{m-1} \to \infty$ always holds for any r_0 when $n \to \infty$.

Remark 2 (Extreme cases)**.** Note that when $r_0 = 0$, we have $s^{m-1} \gtrsim \log n$. This result coincides with our previous result in note 0306_proof, which investigates the seed condition for π_1 to achieve full recovery. When $r_0 = n$, we need $s^{m-1} = \mathcal{O}(1)$ which indicates we do not have any meaningful constraint for s in this case.

Remark 3 (Compare with Ding et al. (2021))**.** Our result also applies to the matrix case by taking $m = 2$. Compared with Lemma 19 in Ding et al. (2021), we relax the seed condition from $s \gtrsim \log n$ to $s \gtrsim \log n - \log r_0$ when π_1 has errors $r_0 \asymp \log n$.

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

It suffices to show any permutation $\pi : S^c \mapsto T^c$ with more that r_0 errors is not picked by criterion (1) with high probability, where $r_0 \in \mathbb{N} \cap [0, n-s]$; i.e.,

$$\begin{aligned}\mathbb{P}\left(\sum_{i\in S^c} H_{ii} > \max_{r>r_0\in\mathbb{N}\cap[0,n-s]} \max_{\pi\in\Pi_r} \sum_{i\in S^c} H_{i\pi(i)}\right) \\ \geq \mathbb{P}\left(\min_{r>r_0\in\mathbb{N}\cap[0,n-s]} \min_{\pi\in\Pi_r} \left(\sum_{i\in S^c} H_{ii} - \sum_{i\in S^c} H_{i\pi(i)}\right) \geq t\right) \to 1,\end{aligned}$$

as $n \to \infty$ for some positive constant t, 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 $r > r_0 \geq 0$. Let the $R = \{i \in S^c : \pi(i) \neq i\}$ denote the set of errors in π, and we have $|R| = r \geq 1$. Then, consider 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 $t' = \frac{\rho}{4} rs^{m-1}$. Note that by assumption

$$rs^{m-1} \gtrsim r \log n - r \max\{\log r_0, 0\} + r. \tag{2}$$

When $r_0 = o(n)$, the lower bound (2) is dominated by $r \log n$; when $r_0 \asymp n$, the lower bound (2) is dominated by $r \geq r_0$. Hence, we always have $rs^{m-1} \to \infty$ as $n \to \infty$ for $r > r_0$. Then, $t^2/rs^{m-1} \leq \rho/4$ when n is large enough. We have

$$\mathbb{P}\left(\sum_{i \in S^c} H_{ii} - \sum_{i \in S^c} H_{i\pi(i)} < t\right) \leq 3 \exp\left(-\frac{rs^{m-1}}{128}\rho^2\right).$$

Note that $|\Pi_r| = \binom{n}{r} \leq \frac{n^r}{r!}$ for $r \geq 1$. Hence,

$$\begin{aligned}
\mathbb{P}\left(\min_{r > r_0} \min_{\pi \in \Pi_r} \left(\sum_{i \in S^c} H_{ii} - \sum_{i \in S^c} H_{i\pi(i)}\right) < t\right) &\leq \sum_{r \geq r_0} \frac{n^r}{r!} \mathbb{P}\left(\sum_{i \in S^c} H_{ii} - \sum_{i \in S^c} H_{i\pi(i)} < t\right) \\
&\leq 3 \sum_{r > r_0} \frac{n^r}{r!} \exp\left(-\frac{1}{128} rs^{m-1}\rho^2\right) \\
&\leq 3 \sum_{r > r_0} \exp\left(-\frac{1}{256} rs^{m-1}\rho^2\right) \\
&\leq 3 \exp\left(-\frac{1}{256}(r_0+1)s^{m-1}\rho^2\right) \to 0.
\end{aligned}$$

In the above inequalities, the first inequality follows by the union bound; the third inequality follows by the assumption that

$$rs^{m-1} \gtrsim r \log n - r \log r_0 \gtrsim r \log n - r \log r \gtrsim r \log n - \log(r!),$$

where the last inequality follows by the Stirling's approximation that $\log(x!) \asymp x \log x$ and thus $\frac{n^r}{r!} \exp\left(-rs^{m-1}\right) \lesssim 1$; the last inequality follows by the sum of proportional sequence that $\sum_{r > r_0} q_0 q^r \leq \frac{q_0 q}{1-q} \leq q_0 q$ for $q < 1$; and the probability decays to 0 due to the implication of assumption (2).

Therefore, we have finished the proof of Theorem 0.1.

$\square$

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

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.