Error control of seeded matching

Jiaxin Hu

incomplete note

March 23, 2022

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. The output π_1 of seeded matching Algorithm 1 has at most r_0 errors.*

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.

Note that

$$\begin{aligned}\mathbb{P}\left(\sum_{i \in S^c} H_{ii} < t_1\right) &= \mathbb{P}\left(\frac{1}{(n-s)s^{m-1}} \sum_{i \in S^c} H_{ii} < \frac{t_1}{(n-s)s^{m-1}}\right) \\ &\leq 2\exp\left(-\min\left\{\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right\}(n-s)s^{m-1}\left(\rho - \frac{t_1}{(n-s)s^{m-1}}\right)^2\right), \end{aligned} \tag{2}$$

for $\rho - \frac{t_1}{(n-s)s^{m-1}} \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$, where the inequality follows from Lemma 1.

Consider an arbitrary $\pi \in \Pi_r$ and let the $R = \{i \in S^c : \pi(i) \neq i\}$ denote the set of errors in π, where $|R| = r$. Then, by Lemma 1, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{i\in S^c} H_{i\pi(i)} > t_2\right) &\leq \mathbb{P}\left(\sum_{i\in S^c/R} H_{ii} > t_2 - t'\right) + \mathbb{P}\left(\sum_{i\in R} H_{i\pi(i)} > t'\right) \\
&= \mathbb{P}\left(\frac{1}{(n-s-r)s^{m-1}}\sum_{i\in S^c/R} H_{ii} > \frac{t_2 - t'}{(n-s-r)s^{m-1}}\right) + \mathbb{P}\left(\frac{1}{rs^{m-1}}\sum_{i\in R} H_{i\pi(i)} > \frac{t'}{rs^{m-1}}\right) \\
&\leq 2\exp\left(-\min\left\{\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right\}(n-s-r)s^{m-1}\left(\frac{t_2-t'}{(n-s-r)s^{m-1}} - \rho\right)^2\right) \\
&\quad + \exp\left(-\frac{(t')^2}{4rs^{m-1}}\right),
\end{aligned}$$

for $\frac{t_2-t'}{(n-s-r)s^{m-1}} - \rho \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$ and $\frac{t'}{rs^{m-1}} \in [0, \sqrt{2}]$. Note that $\min\left\{\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right\} \geq \frac{1}{32}$, and $|\Pi_r| = \binom{n}{r} \leq \frac{n^r}{r}$. By union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{r\geq r_0}\max_{\pi\in\Pi_r}\sum_{i\in S^c} H_{i\pi(i)} > t_2\right) \\
&\leq \sum_{r\geq r_0}^{n}\frac{n^r}{r}\left\{2\exp\left(-\frac{(n-s-r)s^{m-1}}{32}\left(\frac{t_2-t'}{(n-s-r)s^{m-1}} - \rho\right)^2\right) + \exp\left(-\frac{(t')^2}{4rs^{m-1}}\right)\right\}. \qquad (3)
\end{aligned}$$

Now, we only need to verify there exists proper $t_1 > t_2$ such that the probabilities (2) and (3) tends to 0 as $n \to \infty$. We check the constraint for t_1, t', t_2, respectively.

For t_1, we have

$$\begin{aligned}
&\begin{cases} \rho - \frac{t_1}{(n-s)s^{m-1}} \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}] \\ \rho - \frac{t_1}{(n-s)s^{m-1}} > \left((n-s)s^{m-1}\right)^{-1/2} \end{cases} \\
\Rightarrow \quad & f(\rho)(n-s)s^{m-1} \leq t_1 \leq \left(\rho - \frac{1}{\sqrt{(n-s)s^{m-1}}}\right)(n-s)s^{m-1},
\end{aligned}$$

where $f(\rho) = \rho - \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}$, the upper bound follows from the decay of probability (2) (second constraint), and the lower bound follows from Lemma 1 (first constraint).

For t' and any $r \geq r_0$, we have

$$\begin{cases} \frac{t'}{rs^{m-1}} \in [0, \sqrt{2}] \\ \frac{(t')^2}{4rs^{m-1}} \geq r\log n - \log r \end{cases} \quad \Rightarrow \quad 4r^{1/2}\sqrt{r\log n - \log r}s^{(m-1)/2} \leq t' \leq \sqrt{2}rs^{m-1},$$

where the lower bound follows from the decay of probability (3) (second constraint), and the upper bound follows from Lemma 1 (first constraint).

For t_2 and any $r \geq r_0$, we have

$$\begin{cases} \frac{t_2-t'}{(n-s-r)s^{m-1}} - \rho \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}] \\ \frac{(n-s-r)s^{m-1}}{32}\left(\frac{t_2-t'}{(n-s-r)s^{m-1}} - \rho\right)^2 \geq r\log n - \log r \end{cases}$$
$$\Rightarrow \quad \rho(n-s-r)s^{m-1} + 8\sqrt{(r\log n - \log r)(n-s-r)s^{m-1}} + t' \leq t_2 \leq g(\rho)(n-s-r)s^{m-1} + t',$$

where $g(\rho) = \rho + \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}$, the lower bound follows from the decay of probability (3) (second constraint), and the upper bound follows from Lemma 1 (first constraint).

$\square$

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

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

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